COMMUNICATIONS SOLUTIONS

For those who value information

datapro

Datapro Research Corporation, 1805 Underwood Blvd., Delran, NJ 08075; 609-764-0100/A McGraw-Hill Company
Chicago 312-236-8206/Dallas 214-980-1525/Palo Alto 415-967-6007/Phoenix 602-263-7831/
Washington 301-589-6040

Contents, Volume 1

N.B. For a complete, detailed listing of the contents of this information service, see the Index, Section CS01.

Report	Report Number	No. of Pages	Pub'n. Date
INDEX	CS01-100-101	7	5/80
USER'S GUIDE	CS03-100-101	3	6/79
INQUIRY SERVICE	CS05-100-101	14	6/79
BASIC CONCEPTS			
Topic Index	CS10-000-101	1	2/80
Technology Overviews			
A Brief Historical Perspective of Computer Communications	CS10-120-101	10	6/79
Teleprocessing—The Modern Marriage of Computers and Communications	CS10-150-101	8	6/79
The Data Transmission Channel			
The Special Requirements of Data Communications Traffic	CS10-210-101	10	6/79
The Basic Parameters of a Data Channel	CS10-220-101	9	6/79
General Solutions to Common Communications Interface Problems	CS10-230-101	11	6/79
Multiplexing and Concentration			
The Strengths and Applications of Digital Data Multiplexing	CS10-310-101	10	6/79
The Power of Concentration	CS10-320-101	4	6/79
Digital Transmission Techniques			
The Language and Techniques of Pulse Code Modulation (PCM)	CS10-410-101	16	6/79
The Nature and Organization of Digital Transmission Channels	CS10-411-101	10	6/79
Protocols			
Why Protocols? Why Standards?	CS10-510-101	4	6/79
A Tutorial on Protocols	CS10-520-101	32	7/79
Network Technologies			
A Concise Summary of Public, Leased, and Private Transmission Facilities	CS10-610-101	8	6/79
The Uses of Message Switching in Communications Networks	CS10-620-101	20	6/79
The Techniques and Applications of Packet Switching	CS10-650-101	9	6/79
Operating Systems for Computer Networks	CS10-660-101	11	2/80
Wideband Transmission Facilities			
Understanding Satellite-Based Telecommunications Systems	CS10-710-101	9	6/79
Communications Control Techniques			
A Summary Overview of Communications Network Control Architectures and Protocols	CS10-810-101	23	6/79
SYSTEM COMPONENTS			
Topic Index	CS15-000-101	1	5/80
System Component Overview			
A Brief Classification Guide to Data Communications Hardware	CS15-050-101	12	6/79
The Integrated Uses of Hardware in Network Configurations	CS15-060-101	13	6/79
Terminals			
The Categories and Functions of Data Terminals	CS15-100-101	8	6/79
Typing and Printing at a Distance	CS15-120-101	11	9/79
Display Terminals—The Soft Copy Alternative to Teleprinters	CS15-150-101	12	11/79
User Programmable Terminals	CS15-160-101	8	1/80
Remote Batch Terminals	CS15-170-101	8	2/80
Integrated Point of Sale (POS) Systems	CS15-180-201	8	6/79
Voice Response Data Terminals	CS15-190-101	11	6/79
Interface Equipment			
Modems: Terminal—Line Linking Devices	CS15-220-101	8	4/80
Transmission Line Equipment			
Line Use Optimization with Multiplexing Equipment	CS15-320-101	7	12/79
Concentration and Line Management			
Communications Processors—Special Kinds of Computer-Helping Computers	CS15-420-101	12	7/79
Telephone Equipment			
Private Telephone Systems—PBX and PABX	CS15-610-101	9	6/79
The Equipment and Techniques of Digital Telephony	CS15-612-101	14	6/79
Transmission Facilities			
Current Transmission Facility Characteristics	CS15-710-101	10	6/79
Users' Views of Transmission Facilities	CS15-711-101	5	10/79
PLANNING			
Topic Index	CS20-000-101	1	5/80
Need Assessment Planning Overview			
How to Determine Your Data Communications Needs	CS20-110-101	9	6/79
A Digest of the Latest Communications System Planning Techniques	CS20-115-101	9	6/79
Gaining a Better Perspective of Communications System Applications	CS20-120-101	13	6/79
Planning Ahead for Your Data Communications Applications	CS20-121-101	21	6/79
A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs	CS20-150-101	8	6/79
Terminals			
A Short Checklist of Terminal Planning Factors	CS20-205-101	7	6/79
How to Calculate the Number of Terminals You Will Need in Your Communications Network	CS20-210-101	23	6/79
How to Measure a Terminal's I.Q.	CS20-215-101	6	6/79
Providing User-Oriented System Interfaces	CS20-220-101	9	5/80
Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path	CS20-235-101	5	6/79

Contents, Volume 1

N.B. For a complete, detailed listing of the contents of this information service, see the Index, Section CS01.

Report	Report Number	No. of Pages	Pub'n. Date
Transmission Facilities			
Planning Considerations for Dial-Up Versus Leased-Line Facilities	CS20-325-101	15	6/79
Applications			
Office Automation			
An Office Automation Study	CS20-400-101	16	12/79
EFT			
High Velocity Money—Preparing for EFT	CS20-410-101	10	6/79
Planning Guidelines for an EFT Audit and Control System	CS20-412-101	6	6/79
The Economics and Other Issues of EFT Systems	CS20-415-201	8	6/79
Distributed Processing			
How to Conduct a Distributed Processing Feasibility Study	CS20-510-101	18	6/79
Organizational Design for Distributed Processing	CS20-515-101	10	1/80
System Architecture for Distributed Data Management	CS20-520-101	6	9/79
A Language for Distributed Processing	CS20-530-101	12	10/79
Maintaining Data Order and Consistency in a Multi-Access Environment	CS20-540-101	7	3/80
Electronic Mail			
Planning for Electronic Mail	CS20-610-101	6	6/79
Facsimile—How to Analyze Its Role in Your Organization	CS20-650-101	8	8/79
Teleconferencing			
Planning for All-Electronic Multimedia Corporate Communications	CS20-710-101	7	6/79
MIS			
The Communications Framework of Management Information Systems	CS20-940-101	8	6/79

COMMUNICATIONS SOLUTIONS

Filing Instructions— May 1980

This supplement contains a new report on providing user-oriented system interfaces and another interesting report about using an IBM Series/1 minicomputer as a communications system management tool. You will also find a drastically revised report on network diagnostic tools which includes the results of a recent Datapro user survey.

It takes only a few minutes to file the new reports in your looseleaf binders. Just follow the simple filing instructions on this page.

SECTION	REMOVE		INSERT		DESCRIPTION
	Report Number	No. of Pages	Report Number	No. of Pages	
Volume 1	Contents	2	Contents	2	Contents of Volume 1 (revised)
Index	CS01-050-101	2	—	—	Supplementary Index (merged with master index)
	CS01-100-101	6	CS01-100-101	7	Index (revised)
Planning	CS20-000-101	1	CS20-000-101	1	Topic Index (revised)
	—	—	CS20-220-101	9	Providing User-Oriented System Interfaces (new)
Volume 2	Contents	2	Contents	2	Contents Volume 2 (revised)
Installation and Maintenance	CS35-610-201	11	CS35-610-201	36	Network Diagnostic Tools: An Equipment/ Vendor Survey (revised)
Systems Management	CS50-000-101	1	CS50-000-101	1	Topic Index (revised)
	—	—	CS50-330-101	10	Using a Minicomputer as a Communications System Management Tool (new)
Newsbriefs	—	—	CS99-001-051	4	Newsbriefs (new)

Contents, Volume 2

N.B. For a complete, detailed listing of the contents of this information service, see the Index, Section CS01.

Report	Report Number	No. of Pages	Pub'n Date
SYSTEM DESIGN			
Topic Index	CS25-000-101	1	4/80
Design Motives and Goals			
Resource Sharing—The Motivating Goal of Communications System Design	CS25-110-101	19	6/79
A Concise Overview of Communications System Design Factors	CS25-120-101	16	6/79
Nontechnical Issues in Network Design	CS25-121-101	7	6/79
How to Design a Communications System with Elegance and Simplicity	CS25-190-101	9	6/79
Terminals			
The Impact of Human Biology on Man-Machine Interface Design	CS25-220-101	16	6/79
The Effects of Terminal Dialogue Structure on Network Costs	CS25-221-101	5	6/79
Making a Special Purpose Terminal Work in a General Purpose Systems Environment	CS25-222-101	6	4/80
Line Control			
The Elements of Line Control in Communications System Design	CS25-310-101	12	6/79
A Guide to IBM's System Network Architecture (SNA)	CS25-320-101	11	7/79
Networks			
A 10-Year Summary of Solutions to Packet-Switched Network Problems	CS25-411-101	14	6/79
ARPANET Versus SNA—A Side-by-Side Comparison of Two Packet-Switched Network Design Approaches	CS25-415-101	10	6/79
Security			
Some Insights into Cryptography and Its Uses in Data Communications	CS25-510-101	15	6/79
Prevention of Computer Crime	CS25-520-101	15	10/79
Applications for Multilevel Secure Operating Systems	CS25-530-101	11	12/79
Design Aids			
Categories of Data Transmission Systems	CS25-605-101	6	8/79
How to Develop Computer Programs for Network Design Assistance	CS25-610-101	10	6/79
How to Organize and Use a Database for Network Design and Management	CS25-611-101	8	6/79
A Brief Introduction to the Language and Techniques of Queueing Theory	CS25-615-101	15	6/79
Economic Analyses			
How to Evaluate the Economics of System Design Alternatives	CS25-810-101	26	6/79
SELECTION AND ACQUISITION			
Topic Index	CS30-000-101	1	3/80
Data Communications Hardware			
Basic Guidelines to the Selection and Acquisition of Communications System Hardware	CS30-110-101	13	6/79
Computer Services			
Comparing Interactive Computer Services	CS30-210-101	8	2/80
Hardware Performance Analysis			
Selecting Data Communications Facilities	CS30-300-101	29	3/80
Performance Analysis of Complex Communications Systems	CS30-310-101	10	11/79
Telephone Equipment			
The Evaluation and Selection of a Telephone System	CS30-410-101	22	6/79
Software			
How to Evaluate and Select a Data Communications Monitor	CS30-610-101	6	6/79
Contract Negotiations			
How to Negotiate Hardware Contracts	CS30-710-101	9	6/79
Contracts for Acquiring Software Packages	CS30-711-101	10	6/79
Contracts for Outside Processing Services	CS30-712-101	9	8/79
INSTALLATION AND MAINTENANCE			
Topic Index	CS35-000-101	1	10/79
Maintenance Tools and Techniques			
The Equipment and Techniques of Communications Systems Maintenance	CS35-610-101	7	6/79
Network Diagnostic Tools: An Equipment/Vendor Survey	CS35-610-201	36	5/80
How to Design a Communications System Diagnostic Test Center	CS35-611-101	5	6/79
Maintenance Services			
Understanding Independent Third Party Maintenance Services	CS35-710-101	12	6/79
OPERATIONS MANAGEMENT			
Topic Index	CS40-000-101	1	4/80
Operational Checks			
A Protocol Validation Technique for Operational Confidence	CS40-320-101	13	6/79
Workflow—A Technique for Analyzing JES Systems	CS40-330-101	6	10/79
Operating Improvements			
How to Improve Line Utilization in an Existing System	CS40-410-101	9	6/79
Sophisticated Modem and Multiplexer Utilization	CS40-420-101	5	11/79
Operating Cost Reductions			
Putting the Squeeze on Your Telecommunications Costs	CS40-510-101	10	6/79
Control of Variable Telephone Costs	CS40-520-101	7	9/79
Security			
Data Security and Protection Structures	CS40-750-101	9	6/79
Communications and Data Base System Safeguards	CS40-760-101	28	4/80

Contents, Volume 2

N.B. For a complete, detailed listing of the contents of this information service, see the Index, Section CS01.

Report	Report Number	No. of Pages	Pub'n Date
SYSTEMS MANAGEMENT			
Topic Index	CS50-000-101	1	5/80
System Performance Evaluation			
How to Measure and Evaluate Computer Response and Turnaround Times in an Interactive Network	CS50-120-101	13	6/79
A Communications System Performance Checklist	CS50-130-101	6	6/79
Line Optimization			
The Differences in Line Optimization Approaches to Short- and Long-Line Systems	CS50-220-101	9	6/79
Special Systems Management			
Communications Network Management in a Research and Development Environment	CS50-310-101	10	6/79
Institutional Control of Computing Funds and Resources in a Networking Environment	CS50-320-101	8	1/80
Using a Minicomputer as a Communications System Management Tool	CS50-330-101	10	5/80
Applications Management			
Distributed Processing—Before	CS50-510-101	14	6/79
Distributed Processing—After	CS50-510-201	12	6/79
International Operations			
How to Cut Costs and Improve Service of Your International Telecommunications	CS50-810-101	17	8/79
Nationalism and Data Communications	CS50-815-101	8	6/79
FUTURE SYSTEMS			
Topic Index	CS60-000-101	1	3/80
Applications and Software			
A Realistic Look at the Future of EFT, POS, and Electronic Mail	CS60-160-101	8	6/79
Some Yet-to-be-Solved Problems of Distributed Processing Systems	CS60-170-101	6	6/79
The Future for Communications Software	CS60-180-101	7	6/79
Terminals			
Terminals in the 1980's	CS60-220-101	9	6/79
A Prototype Advanced Terminal System	CS60-230-101	10	2/80
New Transmission Technologies			
Fiber Optic Technology—Now and Tomorrow	CS60-310-101	23	6/79
Network Control Architectures			
AT&T Answers 15 Questions About ACS	CS60-510-101	14	6/79
A Fail-Safe Distributed Local Network for Data Communication	CS60-520-101	9	1/80
Communications and Society			
The Technological Path Toward a Distributed Environment	CS60-820-101	7	6/79
Two Possible Communications Futures: 1984 or the Goldfish Bowl	CS60-850-101	10	6/79
Toward the Information Society—A Positive View	CS60-860-101	9	8/79
Potential Impacts of Educational Telecommunications Systems	CS60-865-101	12	3/80
VENDORS AND SUPPLIERS			
Vendors and Suppliers Index	CS90-000-101	1	3/80
Teleprinter Terminals	CS90-120-101	2	10/79
Display Terminals	CS90-150-101	4	10/79
User Programmable Terminals	CS90-160-101	2	3/80
Remote Batch Terminals	CS90-170-101	2	3/80
Integrated Point of Sale (POS) Systems	CS90-180-101	1	12/79
Voice Response Equipment	CS90-190-101	1	6/79
Modems	CS90-220-101	2	6/79
Multiplexers	CS90-320-101	1	6/79
Communications Processors	CS90-420-101	2	6/79
PBX, PABX, and Other Telephone Equipment	CS90-610-101	5	10/79
STANDARDS AND PROTOCOLS			
American Standard Code for Information Interchange (ASCII)	CS93-110-101	7	6/79
Advanced Data Communications Control Procedures (ADCCP)	CS93-112-101	9	12/79
EIA Standard RS-232C	CS93-115-101	7	6/79
IBM Binary Synchronous Communications (BSC)	CS93-301-101	8	6/79
IBM Synchronous Data Link Control (SDLC)	CS93-302-101	7	6/79
Sperry Univac Universal Data Link Control (UDLC)	CS93-310-101	4	6/79
Burroughs Data Link Control (BDLC)	CS93-311-101	3	6/79
CCITT X.25 Packet Switching Interface	CS93-501-101	9	6/79
GLOSSARY			
Glossary	CS98-100-101	18	6/79
NEWSBRIEFS			
Newsbriefs	CS99-001-051	4	5/80

Supplementary Index

This supplementary Index includes only those items which have been added in the January, February, March and April updates to Communications Solutions. For index references to previously published reports, please turn to the Master Index which begins on the next leaf.

A

ARPANET: CS10-660-101

C

centralized data processing—
cost-effectiveness of: CS20-515-101
Comparing Interactive Computer Services: CS30-210-101
communications—
facilities: CS30-300-101
Communications and Data Base System Safeguards: CS40-760-101
computing resources—
availability to institutional users: CS50-320-101
costs—
of computer services: CS50-320-101

D

data communications—
data protection: CS40-760-101
fail-safe network concept: CS60-520-101
security: CS40-760-101
selection of facilities: CS30-300-101
data processing—
centralizing vs decentralizing: CS20-515-101
data control function: CS20-515-101
systems function: CS20-515-101
distributed processing—
fail-safe network concept: CS60-520-101
organizing for: CS20-515-101
terminals: CS15-160-101

E

encryption: CS40-760-101

F

A Fail-Safe Distributed Local Network for Data Communication: CS60-520-101
facilities, selection of: CS30-300-101
future systems—
fail-safe distributed network concept: CS60-520-101

I

Institutional Control of Computing Funds and Resources in a Networking Environment: CS50-320-101

M

Maintaining Data order and Consistency in a Multi-Access Environment: CS20-540-101
Making a Special Purpose Terminal Work in a General Purpose Systems Environment: CS25-222-101

N

networks—
for research facilities and educational institutions: CS50-320-101
operating system: CS10-660-101

O

Operating Systems for Computer Networks: CS10-660-101
Organizational Design for Distributed Processing: CS20-515-101

P

Potential Impacts of Educational Telecommunications Systems: CS60-865-101
A Prototype Advanced Terminal System: CS60-230-101

R

resource sharing: CS50-320-101

Supplementary Index

S

security—
in a data communications environment: CS40-760-101
maintaining data order and consistency: CS20-540-101
Selecting Data Communications Facilities: CS30-300-101
selection—
interactive computer services: CS30-210-101
services—
comparing interactive computer services: CS30-210-101
software—
network operating system: CS10-660-101
system—
network operating: CS10-660-101

T

telecommunications—
impact on education: CS60-865-101
terminal—
intelligent: CS15-160-101
programmable: CS15-160-101
special purpose: CS25-222-101
Terminals in the 1980's—
a prototype advanced terminal system: CS60-230-101

U

User Programmable Terminals: CS15-160-101□

Index

A

access control: CS25-310-101, CS25-520-101
accounting, network: CS25-110-101
ACS: see advanced communication service
ADCCP: (see advanced data communications control protocol)
Advanced Communication Service: CS15-710-101, CS60-510-101
Advanced Data Communications Control Protocol: CS25-810-101
American Standard Code for Information Interchange (ASCII): CS93-112-101
An Office Automation Study: CS20-400-101
applications—
communications network: CS10-150-101, CS15-420-101, CS20-120-101
distributed processing: CS20-510-101, CS20-530-101, CS50-510-101, CS50-510-201
effect on system design: CS25-120-101
for fiber optics: CS60-310-101
gaining a better perspective of: CS20-120-101
management: CS25-190-101
of PCM: CS10-410-101
-oriented software: CS60-180-101
planning: CS20-121-101, CS25-110-101
unilevel and multilevel, for secure systems: CS25-530-101
Applications for Multilevel Secure Operating Systems: CS25-530-101
architecture—
fiber optics: CS60-310-101
network control: CS25-810-101
systems for distributed data management: CS20-520-101
systems network: CS25-320-101
ARPANET: CS10-660-101; CS25-411-101; CS25-415-101
ARPANET Versus SNA—A Side-by-Side Comparison of Two Packet-Switched Network Control Design Approaches: CS25-415-101
AT&T Answers 15 Questions About ACS: CS60-510-101
automated clearing houses: CS20-410-101
automated office: CS20-400-101

B

bandwidth, definitions of: CS10-220-101, CS10-230-101
bank cards: CS20-410-101
Basic Guidelines to the Selection and Acquisition of Communications System Hardware: CS30-110-101
The Basic Parameters of a Data Channel: CS10-220-101
batch concept: CS15-170-101
BDLC: see Burroughs Data Link Control
benefit-cost analysis: CS25-810-101
Brief Classification Guide to Data Communications Hardware: CS15-050-101
Brief Historical Perspective of Computer Communications: CS10-120-101
Brief Introduction to the Language and Techniques of Queuing Theory: CS25-615-101
BSC: see IBM Binary Synchronous Communications
Burroughs Data Link Control: CS20-940-101, CS93-311-101

C

carrier—
communications: CS20-150-101
network links: CS25-120-101
The Categories and Functions of Data Terminals: CS15-100-101
Categories of Data Transmission Systems: CS25-605-101
CCITT X.25 Packet Switching Interface: CS10-650-101, CS20-940-101, CS93-501-101
centralized data processing—
cost-effectiveness of: CS20-515-101
centralized/distributed networks: CS25-120-101
channel—
analog wideband: CS15-710-101
capacity: CS25-220-101
data communications characteristics of: CS10-220-101
digital transmission: CS10-411-101
ciphers: CS25-510-101
coding techniques: CS10-410-101
communications—
and the automated office: CS20-400-101
departments, need for: CS40-510-101
facilities: CS30-300-101; CS40-410-101
from the environmentalist's viewpoint: CS60-820-101
future of: CS60-850-101; CS60-860-101
hardware: CS15-050-101
links: CS15-060-101
man/computer interface: CS25-220-101
monitor: CS30-610-101
networks—
design: CS10-620-101, CS25-611-101
in R&D environment: CS50-310-101
management: CS25-611-101
packet: CS25-411-101
processors: CS15-060-101, CS15-420-101
Communications and Data Base System Safeguards: CS40-760-101
Communications Applications, Planning Ahead for Your Data: CS20-121-101
communications system—
design: CS10-310-101, CS25-110-101, CS50-130-101
digital: CS10-411-101
maintenance: CS35-610-101
test center design: CS35-611-101
using a minicomputer as a management tool: CS50-330-101
The Communications Framework of Management Information Systems: CS20-940-101
Communications Interface Problems, General Solutions to Common: CS10-230-101
Communications Monitor, How to Evaluate and Select a Data: CS30-610-101
Communications, Nationalism and Data: CS50-815-101
Communications Needs, How to Determine Your Data: CS20-110-101
Communications Network Control Architectures, A Summary Overview of: CS25-810-101
Communications Network Management in a Research and Development Environment: CS50-310-101
Communications Processors: CS90-420-101

Index

Communications Processors—Special Kinds of Computer-Helping Computers: CS15-420-101
Communications Regulations and Tariffs, A Short Guide to the Arcana of National and International Data: CS20-150-101
Communications System Applications, Gaining a Better Perspective of: CS20-120-101
Communications System Design Factors, A Concise Overview of: CS25-120-101
Communications System Design, Resource Sharing—The Motivating Goal of: CS25-110-101
Communications System Diagnostic Test Center, How to Design a: CS35-611-101
Communications System, How to Design with Elegance and Simplicity: CS25-190-101
Communications System Performance Checklist: CS50-130-101
Communications Traffic, The Special Requirements of Data: CS10-210-101
Comparing Interactive Computer Services: CS30-210-101
compiling new telephone system specifications: CS30-410-101
components of optical systems: CS60-310-101
computer aided educational network: CS50-110-101
computer crime, prevention of: CS25-520-101
computer network system security: CS25-530-101, CS40-750-101
Computer Response and Turnaround Times in an Interactive Network, How to Evaluate: CS50-120-101
computers, types of: CS30-110-101
computing resources—
 availability to institutional users: CS50-320-101
concentration: CS10-320-101, CS40-410-101
concentrators: CS10-320-101, CS15-060-101
A Concise Overview of Communications System Design Factors: CS25-120-101
A Concise Summary of Public, Leased, and Private Transmission Facilities: CS10-610-101
conditions of measurement: CS50-120-101
contract negotiations: CS30-710-101
Contracts for Acquiring Software Packages: CS30-711-101
Contracts for Outside Processing Services: CS30-712-101
Control of Variable Telephone Costs: CS40-520-101
controllers, communications: CS15-050-101, CS20-510-101
cost considerations, telecommunications: CS25-810-101
cost saving considerations, communications networks: CS15-320-101
costs—
 basic evaluation of: CS25-810-101
 control of, telecommunications systems: CS40-520-101
 electronic fund transfer: CS20-410-201
 international telecommunications: CS50-810-101
 minimizing network: CS20-710-101
 network: CS25-221-101
 of computer services: CS50-320-101
 telecommunications: CS40-510-101
 telephone communications network: CS20-325-101
criteria, distributed processing systems: CS20-510-101
cryptography: CS25-510-101, CS25-520-101, CS40-750-101
Cryptography and Its Uses in Data Communications, Some Insights Into: CS25-510-101

D

Database for Network Design and Management, How to Organize and Use a: CS25-611-101
data base management—
 involved with system architecture: CS20-520-101
 systems security: CS25-530-101
database use: CS25-611-101
Data Channel, The Basic Parameters of a: CS10-220-101
data communications—
 data protection: CS40-760-101
 dial-up versus leased line facilities: CS20-325-101
 fail-safe network concept: CS60-520-101
 hardware: CS15-050-101
 hierarchies and interfaces: CS20-940-101
 needs, determination of: CS20-110-101
 network: CS15-060-101, CS20-115-101
 processors: CS15-420-101
 protocols: CS93-112-101, CS93-115-101, CS93-301-101, CS93-302-101, CS93-310-101, CS93-311-101
 regulations and tariffs: CS20-150-101
 security: CS40-760-101
 security and privacy: CS25-510-101
 selection of facilities: CS30-300-101
 transmission: CS25-810-101
Data Encryption Standard (DES): CS25-510-101, CS25-520-101
data entry: CS15-100-101
data networks: CS20-710-101
data processing—
 centralizing vs decentralizing: CS20-515-101
 data control function: CS20-515-101
 systems function: CS20-515-101
data processing systems: CS50-510-101, CS50-510-201
data security: CS25-520-101, CS40-750-101
Data Security and Protection Structures: CS40-750-101
data transmission systems, categories of: CS25-605-101
dataphone digital service: CS15-710-101
DDD: see direct distance dialing
DDS: see dataphone digital service
decentralization vs. centralization: CS20-510-101
DES: see data encryption standard
design—
 communications system: CS25-110-101, CS25-310-101
 communications system diagnostic test center: CS35-611-101
 data transmission facilities: CS25-190-101
 distributed processing systems: CS20-510-101, CS20-530-101
 network: CS10-620-101
 network control: CS25-415-101
 packet switching network: CS25-411-101
 telephone network: CS20-710-101
diagnostic equipment: CS35-610-101, CS35-610-201
diagnostic test center: CS35-611-101
diagnostics, need for: CS15-220-101
dialogue—
 man/computer: CS25-220-101
 structure of terminal: CS25-221-101
dial-up lines: CS20-325-101
Differences in Line Optimization Approaches to Short- and Long-Line Systems: CS50-220-101
A Digest of the Latest Communications System Planning Techniques: CS20-115-101
Digital Data Multiplexing, The Strengths and Applications of: CS10-310-101
digital transmission: CS10-710-101, CS15-612-101, CS40-410-101
digital transmission and switching systems: CS15-612-101
direct distance dialing: CS15-710-101
display terminals: CS20-210-101
Display Terminals: CS15-150-101, CS90-150-101
Display Terminals—The Soft Copy Alternative to Teleprinters: CS15-150-101
Distributed Environment, The Technological Path Toward a: CS60-820-101
distributed processing—
 advantages: CS50-510-101, CS50-510-201
 approaches to: CS50-510-101
 continuing problems of: CS60-170-101
 data management, system architecture: CS20-520-101
 fail-safe network concept: CS60-520-101
 feasibility study: CS20-510-101
 in remote job entry: CS15-170-101
 organizing for: CS20-515-101
 programming language for: CS20-530-101
 systems: CS50-510-201, CS60-170-101
 terminals: CS15-160-101
Distributed Processing, A Language For: CS20-530-101
distributed processing systems: CS10-520-201; CS20-530-101
Distributed Processing Systems, Some-Yet-To-Be-Solved Problems of: CS60-170-101
distributed services, SNA: CS25-320-101
Distributed Processing—After: CS50-510-201
Distributed Processing—Before: CS50-510-101

Index

E

The Economics and Other Issues of EFT Systems: CS20-410-201
The Effects of Terminal Dialogue Structures on Network Costs: CS25-221-101
EFT: see electronic fund transfer
EFT Audit and Control System, Planning Guidelines for: CS20-412-101
electronic fund transfer: CS20-120-101, CS20-610-101, CS20-409-101, CS20-410-201, CS20-412-101, CS60-160-101
EIA Standard RS-232C: CS10-230-101, CS10-810-101, CS93-115-101
electronic mail: CS20-610-101, CS20-710-101, CS60-160-101
electronic office system: CS15-100-101
The Elements of Line Control in Communications System Design: CS25-310-101
encryption: CS20-235-101, CS25-520-101, CS40-760-101
The Equipment and Techniques of Communications Systems Maintenance: CS35-610-101
The Equipment and Techniques of Digital Telephony: CS15-612-101
equipment—
 distributed processing: CS20-510-101
 network testing: CS35-610-101
 selection—
  communications processors: CS15-420-101
  multiplexers: CS15-320-101
The Evaluation and Selection of a Telephone System: CS30-410-101
evaluation of manufacturer's proposals: CS20-510-101

F

facilities, selection of: CS30-300-101
Facsimile—How to Analayze Its Role in Your Organization: CS20-650-101
A Fail-Safe Distributed Local Network for Data Communication: CS60-520-101
feasibility study, distributed processing: CS20-510-101
features of various computers: CS30-110-101
Fiber Optic Technology—Now and Tomorrow: CS60-310-101
flexibility of various computers: CS30-110-101
front-end processors: CS15-060-101, CS15-420-101
funds transfer—
 authorization: CS20-409-101
 international: CS20-409-101
The Future for Communications Software: CS60-180-101
future systems—
 fail-safe distributed network concept: CS60-520-101

G

Gaining a Better Perspective of Communications System Applications: CS20-120-101
General Solutions to Common Communications Interface Problems: CS10-230-101
Glossary: CS98-100-101

H

hardware, communications: CS15-050-101
Hardware in Network Configurations, The Integrated Uses of: CS15-060-101
Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path: CS20-235-101
high level data link control: CS25-810-101
High-Velocity Money—Preparing for EFT: CS20-409-101
HDLC: see high level data link control
How to Calculate the Number of Terminals You Will Need in Your Communications Network: CS20-210-101
How to Conduct a Distributed Processing Feasibility Study: CS20-510-101
How to Cut Costs and Improve Service of Your International Telecommunications: CS50-810-101
How to Design a Communications System Diagnostic Test Center: CS35-611-101
How to Design a Communications System with Elegance and Simplicity: CS25-190-101
How to Determine Your Data Communications Needs: CS20-110-101
How to Develop Computer Programs for Network Design Assistance: CS25-610-101
How to Evaluate and Select a Data Communications Monitor: CS30-610-101
How to Evaluate the Economics of System Design Alternatives: CS25-810-101
How to Improve Line Utilization in an Existing System: CS40-410-101
How to Measure and Evaluate Computer Response and Turnaround Times in an Interactive Network: CS50-120-101
How to Measure a Terminal's IQ: CS20-215-101
How to Negotiate Hardware Contracts: CS30-710-101
How to Organize and Use a Database for Network Design and Management: CS25-611-101

I

IBM Binary Synchronous Communications (BSC): CS93-301-101
IBM Synchronous Data Link Control (SDLC): CS93-302-101
IBM's Systems Network Architecture (SNA): CS25-320-101
The Impact of Human Biology on Man-Machine Interface Design: CS25-220-101
implementation—
 of a distributed processing system: CS50-510-201
 planning: CS50-510-201
An In-Depth Guide to IBM's Systems Network Architecture (SNA): CS25-320-101
information—
 handling: CS15-190-101
 networks: CS20-121-101
 storage and retrieval systems: CS20-940-101
 systems: CS50-510-101
Information Society, Toward A Positive View: CS60-860-101
Institutional Control of Computing Funds and Resources in a Networking Environment: CS50-320-101
integrated office communications system: CS20-400-101
Integrated Point of Sale (POS) Systems: CS15-180-201, CS90-180-101
The Integrated Uses of Hardware in Network Configurations: CS15-060-101
integration of communications in society: CS60-860-101
interactive networks: CS50-120-101
interconnect: CS15-610-101
interface—
 communications: CS10-230-101
 devices: CS15-100-101
 man-machine: CS25-220-101
international—
 communications: CS20-150-101
 data communications: CS50-815-101

J

JES: see job entry systems: CS40-330-101
job entry, remote: CS40-330-101
job entry systems: CS40-330-101

L

The Language and Techniques of Pulse Code Modulation (PCM): CS10-410-101
Language for Distributed Processing, A: CS20-530-101
languages—
 Concurrent PASCAL: CS20-530-101
 high-level: CS60-180-101
 network command: CS25-110-101
large scale integrated circuits: CS15-612-101
layered architecture: CS25-320-101
leased lines: CS20-325-101

Index

leasing, teleprinter companies: CS15-120-101
line—
categories: CS10-610-101
communications system control: CS25-310-101
utilization: CS40-410-101, CS50-220-101
Line Use Optimization with Multiplexing Equipment: CS15-320-101
Line Utilization, How to Improve in an Existing System: CS40-410-101
links, communications: CS25-310-101
LSI: see large scale integrated circuits

M

machine time scale: CS30-712-101
Maintaining Data Order and Consistency in a Multi-Access Environment: CS20-540-101
maintenance—
communications systems: CS35-610-101
third-party: CS35-710-101
Maintenance Services, Understanding Independent Third-Party: CS35-710-101
Making a Special Purpose Terminal Work in a General Purpose Systems Environment: CS25-222-101
management—
by exception: CS50-510-201
line: CS40-510-101
of computer security: CS25-520-101
of information needs: CS50-510-101
of information systems: CS20-940-101, CS50-510-201
Management Information Systems, The Communications Framework of: CS20-940-101
man/machine interface: CS20-205-101, CS25-220-101, CS25-221-101
measurement criteria for communications systems: CS50-130-101
message—
handling: CS20-710-101
packet-switching: CS40-410-101
switching: CS10-620-101
telecommunications service (MTS): CS50-810-101
methods of measurement and analysis: CS50-120-101
microcomputers—
in distributed processing systems: CS20-510-101
in telephone switching systems: CS15-612-101
microprocessors: CS20-510-101
minicomputer—
used as a communications system management tool: CS50-330-101
MIS: see management information systems
modems: CS10-230-101, CS15-050-101, CS15-060-101, CS15-220-101, CS20-235-101, CS90-220-101
Modems: CS90-220-101
Modems: Terminal-Line Linking Devices: CS15-220-101
modes, interactive and batch: CS50-510-201
monitors—
evaluation: CS30-610-101
types of: CS35-610-201
multiple processor systems: CS20-530-101
multiplex systems, PCM: CS10-410-101
multiplexers—
and concentrators: CS10-320-101
and configuration planning: CS15-060-101
and system design factors: CS25-120-101
categories: CS15-320-101
network: CS15-050-101
vendors: CS90-320-101
Multiplexers: CS90-320-101
multiplexing: CS10-310-101, CS40-410-101

N

National Bureau of Standards: CS25-520-101
National Security Agency: CS25-520-101
Nationalism and Data Communications: CS50-815-101
The Nature and Organization of Digital Transmission Channels: CS10-411-101
networks—
addressable units: CS25-320-101, CS25-810-101
in remote batch operations: CS15-170-101
architectures: CS20-115-101
characteristics of: CS10-620-101
communications: CS20-325-101
communications problems: CS20-235-101
configurations: CS15-060-101, CS25-810-101
and hardware alternatives: CS25-610-101
considerations for: CS15-320-101
control architectures: CS25-310-101
current transmission characteristics of: CS15-710-101
design: CS10-620-101, CS25-121-101, CS25-120-101
tools: CS25-610-101
diagnostic equipment: CS35-610-201
distributed systems: CS50-510-201
for research facilities and educational institutions: CS50-320-101
interactive: CS50-120-101
management—
in communications: CS25-611-101
in an R&D environment: CS50-310-101
number of terminals needed in: CS20-210-101
operating system: CS10-660-101
organization approaches: CS50-220-101
packet: CS25-411-101
queueing problems in: CS25-615-101
satellite based: CS10-710-101
security: CS25-110-101
special purpose: CS10-120-101
systems planning: CS15-060-101
terminal-oriented: CS10-120-101
testing techniques, digital and analog: CS35-610-101
transparency: CS10-650-101
Network Diagnostic Tools: An Equipment/Vendor Survey: CS35-610-201
Nontechnical Issues in Network Design: CS25-121-101

O

Operating Systems for Computer Networks: CS10-660-101
organization—
system security: CS25-530-101
Organizational Design for Distributed Processing: CS20-515-101
Outside Processing Services, Contracts for: CS30-712-101

P

PABX: CS15-610-101, CS15-612-101
packet switching: CS10-650-101, CS25-415-101
Packet-Switching Network Problems, A 10-Year Summary of Solutions to: CS25-411-101
Packet-Switching, The Techniques and Applications: CS10-650-101
PBX: CS15-610-101
PBX, PABX, and Other Telephone Equipment: CS90-610-101
PCM: see pulse code modulation
performance—
criteria for network: CS20-115-101
of computers: CS30-110-101
measurements for communications networks: CS25-120-101
rating: CS50-130-101
planning—
an integrated office communications system: CS20-400-101
for communications systems: CS20-115-101
for data communications: CS20-110-101
Planning Ahead for Your Data Communications Applications: CS20-121-101
Planning Considerations for Dial-Up Versus Leased-Line Facilities: CS20-325-101
Planning for All-Electronic Multimedia Corporate Communications: CS20-710-101
Planning for Electronic Mail: CS20-610-101
Planning Guidelines for an EFT Audit and Control System: CS20-412-101

Index

point of sale: CS15-180-201, CS60-160-101
Point of Sale (POS) Systems, Integrated: CS15-180-201
POS: see point of sale
Potential Impacts of Educational Telecommunications Systems: CS60-865-101
The Power of Concentration: CS10-320-101
Prevention of Computer Crime: CS25-520-101
price, terms of software and delivery: CS30-711-101
printers—
terminals: CS15-060-101
types of: CS15-120-101
printing, impact/non-impact: CS15-120-101
privacy—
data: CS25-520-101
data base: CS30-712-101
privacy, international data; CS50-815-101
private line service: CS15-710-101
Private Telephone Systems—PBX and PABX: CS15-610-101
processors, communications: CS15-060-101, CS15-420-101, CS90-420-101
program architecture: CS25-610-101
programmable communications processors: CS15-420-101
protection—
data: CS25-520-101
A Protocol Validation Technique for Operational Confidence: CS40-320-101
protocols—
and engineering economy: CS25-810-101
and remote batch networks: CS15-170-101
and special requirements of data communications traffic: CS10-210-101
and system network architecture: CS25-320-101
ASCII: CS93-112-101
EIA RS232C: CS93-115-101
functions: CS10-520-201
in message switching networks: CS10-620-101
needs for: CS10-510-101
standards: CS10-520-201
tutorial on: CS10-520-201
validation: CS40-320-101
A Prototype Advanced Terminal System: CS60-230-101
Providing User-Oriented System Interfaces: CS20-220-101
pulse code modulation: CS10-410-101, CS10-411-101
Putting the Squeeze on Your Telecommunications Costs: CS40-510-101

Q

queuing theory: CS20-210-101, CS25-120-101, CS25-615-101

R

rate of return cost analysis method: CS25-810-101
RBT: see Remote Batch Terminals
A Realistic Look at the Future of EFT, POS, and Electronic Mail: CS60-160-101
real-time MIS: CS50-510-101
regulation—
of communications: CS20-150-101
of international data communications: CS50-815-101
reliability analysis: CS25-610-101
RJE: see Remote Job Entry
Remote Batch Terminals: CS15-170-101
remote batch terminals: CS15-060-101, CS15-170-101, CS20-510-101, CS40-330-101
Remote Batch Terminals: CS90-170-101
Remote Batch Terminals or How to Tap Into Idle Computer Time from Anywhere in the World: CS15-170-101
remote job entry: CS15-170-101, CS40-330-101
resource sharing: CS20-121-101, CS25-110-101, CS25-120-101, CS25-411-101, CS40-410-101, CS50-320-101
Resource Sharing—The Motivating Goal of Communications System Design: CS25-110-101
response time—
analysis: CS25-610-101
and consideration of communication needs: CS20-110-101
and planning for leased-line facilities: CS20-325-101
and terminal planning factors: CS20-210-101
communications network: CS25-120-101
man/computer interface: CS25-220-101
measurement of: CS50-120-101

S

satellite—
channels, unique properties of: CS10-710-101
costs: CS10-710-101
communications: CS10-710-101
services: CS15-710-101
SDLC: see IBM Synchronous Data Link Control
security—
data: CS40-750-101
in a data communications environment: CS40-760-101
maintaining data order and consistency: CS20-540-101
operating systems: CS25-530-101
/privacy, EFTS: CS20-409-101
remote terminal: CS40-750-101
system: CS25-510-101
Selecting Data Communications Facilities: CS30-300-101
selection—
distributed processing system: CS20-510-101
equipment: CS20-510-101
interactive computer services: CS30-210-101
telephone system: CS30-410-101
service bureaus: CS30-712-101
service support systems, resource-sharing: CS25-110-101
services—
comparing interactive computer services: CS30-210-101
Services, Contracts for Outside Processing: CS30-712-101
session structure: CS25-320-101
A Short Checklist of Terminal Planning Factors: CS20-205-101
A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs: CS20-150-101
SNA: see systems network architecture
SNA supported IBM terminals: CS25-320-101
society, information-discussion of: CS60-860-101
software—
communications: CS60-180-101
network operating system: CS10-660-101
source availability and access of: CS30-711-101
specifications of: CS30-711-101
Software, The Future for Communications: CS60-180-101
Some Insights Into Cryptography and Its Uses in Data Communications: CS25-510-101
Some Yet-To-Be-Solved Problems of Distributed Processing Systems: CS60-170-101
The Special Requirements of Data Communications Traffic: CS10-210-101
Sperry Univac Universal Data Link Control (UDLC): CS93-310-101
standard data encyrption algorithm: CS25-510-101
standards: CS10-210-101, CS10-510-101
The Strengths and Applications of Digital Data Multiplexing: CS10-310-101
subsystems, job entry: CS40-330-101
A Summary Overview of Communications Network Control Architectures: CS25-810-101
switched networks: CS15-710-101
switching equipment: CS15-610-101
synchronous data link control: CS25-810-101
system—
data transmission: CS25-605-101
design of facsimile: CS20-650-101
foreign exchange and tie-line: CS40-520-101
job entry: CS40-330-101
network architecture: CS25-230-101, CS25-415-101, CS25-810-101
network operating: CS10-660-101
planning: CS20-115-101
transmission time: CS25-605-101

Index

System Architecture for Distributed Data Management: CS20-520-101
system design: CS20-940-101, CS25-110-101, CS25-120-101, CS25-220-101, CS25-310-101
system management—
using a minicomputer as a management tool: CS50-330-101
Systems Network Architecture (SNA), A Long Leisurely Look at IBM's: CS25-320-101

T

tariffs: CS10-610-101, CS20-150-101, CS40-510-101
tech control center: CS35-611-101
The Techniques and Applications of Packet-Switching: CS10-650-101
The Technological Path Toward a Distributed Environment: CS60-820-101
technology—
communications: CS10-150-101, CS60-820-101
EFT: CS20-410-201
teleprinter: CS15-120-101
telecommunications—
access methods: CS30-610-101
and electronic mail: CS20-610-101
equipment and techniques: CS15-612-101
impact on education: CS60-865-101
international: CS40-510-101
management and design system: CS25-611-101
nature and organization of transmission channels: CS10-411-101
planning considerations for: CS20-710-101
private telephone system: CS15-610-101
system performance: CS50-130-101
understanding of: CS10-710-101
Telecommunications Costs, Putting the Squeeze on: CS40-510-101
Telecommunications, How to Cut Costs and Improve Service of Your International: CS50-810-101
teleconferencing: CS20-710-101
telephone—
communications: CS20-325-101
cost controls: CS40-520-101
digital: CS15-612-101
equipment vendors: CS90-610-101
interconnect: CS20-150-101
network—
corporate: CS20-710-101
organization: CS15-710-101
traffic: CS60-160-101
private systems: CS15-610-101
use analysis form: CS30-410-101
Telephone System, The Evaluation & Selection of a: CS30-410-101
teleprinter leased channels: CS50-810-101
Teleprinter Terminals: CS90-120-101
teleprinters: CS15-120-101, CS15-150-101, CS20-210-101
Teleprocessing—The Modern Marriage of Computers and Communications: CS10-150-101
teleprocessing systems: CS10-150-101, CS10-230-101
teletypewriter terminals: CS15-060-101
telex, international: CS50-810-101
TELPAK network: CS25-611-101
A 10-Year Summary of Solutions to Packet-Switching Network Problems: CS25-411-101
terminal—
bank: CS20-409-101
communications requirements: CS20-235-101
control units: CS50-220-101
data communications: CS15-100-101, CS15-190-101, CS15-220-101, CS20-325-101
dialogue structure: CS25-221-101
display: CS15-150-101
intelligent: CS15-160-101
microprocessor controlled: CS15-150-101
number of in networks: CS20-210-101, CS25-120-101
operations: CS20-215-101
operators: CS20-205-101, CS20-210-101, CS25-220-101
-oriented networks: CS10-120-101
performance analysis: CS40-330-101
point of sale: CS15-180-201
programmable: CS15-160-101
real-time system: CS25-220-101
remote batch: CS40-330-101
special purpose: CS25-222-101
subsystem developments in 1980: CS60-220-101
system: CS15-050-101
teleprinter: CS15-120-101
traditional characteristics of: CS60-220-101
types of: CS15-060-101, CS20-215-101, CS20-510-101, CS25-320-101
types of usage: CS60-220-101
utilization of: CS50-510-201
vendors listing: CS90-120-101, CS90-150-101, CS90-170-101, CS90-180-101
Terminal Planning Factors, A Short Checklist of: CS20-205-101
Terminals, How to Calculate the Number of You Will Need in Your Communications Network: CS20-210-101
Terminals in the 1980's: CS60-220-101
Terminals in the 1980's—
a prototype advanced terminal system: CS60-230-101
terminals—
user oriented: CS20-220-101
third-party computer maintenance: CS35-710-101
Third-Party Maintenance Services, Understanding Independent: CS35-710-101
time division multiplexing: CS10-310-101
timesharing: CS25-605-101
Toward the Information Society—A Positive View: CS60-860-101
traffic, data communications: CS10-210-101, CS20-110-101, CS20-115-101, CS20-325-101
transaction systems: CS15-100-101
transborder data flow: CS50-815-101
transmission—
analog: CS10-410-101, CS15-710-101
channels: CS15-050-101, CS15-320-101
communications: CS20-121-101
data: CS10-320-101
data networks: CS10-650-101
digital: CS15-710-101
facilities: CS10-610-101, CS15-710-101, CS20-115-101
logical and physical: CS60-180-101
problems: CS10-230-101
rates: CS10-610-101
services: CS10-210-101
Transmission Channels, The Nature and Organization of Digital: CS10-411-101
Transmission Facilities, A Concise Summary of Public, Leased, and Private: CS10-610-101
Transmission Facility Characteristics: CS15-710-101
transmission facsimile: CS20-650-101
transmission systems, off line: CS25-605-101
transmission systems, on line: CS25-605-101
A Tutorial on Protocols: CS10-520-201
transparency, network/terminal: CS60-180-101
Two Possible Communications Futures: 1984 or the Goldfish Bowl: CS60-850-101
Typing and Printing at a Distance: CS15-120-101

U

UDLC: see Sperry Univac Universal Data Link Control
Understanding Independent Third-Party Maintenance Services: CS35-710-101
Understanding Satellite-Based Telecommunications Systems: CS10-710-101
user—
considerations of terminals: CS20-205-101
man-machine dialogue: CS20-220-101
service measures: CS50-120-101
terminal orientation: CS20-220-101
User Programmable Terminals: CS15-160-101, CS90-160-101
Users' Views of Transmission Facilities: CS15-711-101
The Uses of Message Switching in Network Design: CS10-620-101

Index

Using a Minicomputer as a Communications System Management Tool: CS50-330-101
utilities, network: CS25-110-101

V

validation of protocols: CS40-320-101
value-added networks: CS15-710-101
viewdata, description: CS60-860-101
voice/record leased channels: CS50-810-101
voice recognition: CS15-190-101
voice-response: CS15-100-101, CS15-190-101
Voice Response Data Terminals: CS15-190-101, CS90-190-101
Voice Response Equipment: CS90-190-101

W

Why Protocols? Why Standards?: CS10-510-101
Workflow—A Technique for Analyzing JES Systems: CS40-330-101
workflow analysis: CS40-330-101

X

X.25: CS25-810-101□

User's Guide

DATAPRO COMMUNICATIONS SOLUTIONS is an information service devoted entirely to the problems of communications equipment, systems, and networks. Communications, in its modern sense, is frequently restricted to mean data communications, which defines the purely digital world of terminals and computers but which excludes the richly expressive worlds of electronically augmented voice and video communications. The latest developments in communications indicate a distinct trend toward all-electronic multimedia systems in which every possible form of communications is exploited. The user is emerging as the star of this developing scenario because the electronic machinery of the media is becoming increasingly transparent to the user and thus far more accommodating to the user in terms of how easily information can be transferred from point to point. Much of the "easiness" of modern communications is the direct product of wider and faster channels among communications users coupled with computer-assisted channel control and routing techniques.

DATAPRO COMMUNICATIONS SOLUTIONS addresses many of the difficult adjustment problems that are surfacing in the wake of this growing merger between the separate disciplines of communications and computers and the interaction of these merging disciplines with the equally growing needs of an ever-larger body of users—users who are anxious to realize the benefits of teleprocessing, distributed processing, and many other new techniques but with as little technological pain as possible. DATAPRO COMMUNICATIONS SOLUTIONS answers these needs with concise, easy-to-read syntheses of the best currently available solutions to major communications problems.

PURPOSE OF THIS PUBLICATION

If your company does over one million dollars per year, you must consider using at least one inescapable product of modern technology—the computer. One million dollars is a convenient breakpoint because it is a general indicator of where certain clerical and other operations of a business can be profitably automated relative to the current costs of data processing hardware and software. The breakpoint can be slightly higher or lower and will be influenced somewhat by the nature of your company's business, but once beyond the breakpoint, the nature of your company's business is irrelevant to the need for a computer. Why? Because it has been amply and unarguably demonstrated that a company can operate more profitably (and thus more competively) with automated data processing facilities, and it is becoming practically impossible for a company to survive competitively without them. This dawning fact of economic and business survival is the dominant pressure in the rapidly swelling small-business computer market; but there are other ways one can have access to data processing power without necessarily buying a computer for each user.

User's Guide

The classic approach, one that started early in the 60's, is time sharing. Now called by many different names, the basic concept is simply one of distributing the unused portion of a central computer's power to remote users who can then "talk" to the computer through a terminal, which costs very little compared to the computer but which makes much of the computer's power available and usable even though the computer may be physically located several thousand miles away from the user. However, although terminals are quite adequate for casual processing tasks and are moderately adequate for certain small- to medium-sized tasks, they are not adequate for large, complex tasks. So, a modern trend, stimulated primarily by rapidly falling hardware prices, is to concentrate increasingly larger amounts of unfettered processing power at the remote sites and to link the sites together into a network of intelligent nodes. This approach is called distributed processing.

To restate the options: you can either buy a computer and use it as a dedicated central resource that only a few people can use; or you can buy/rent/lease computing power and distribute it among many users, each according to his special needs. Obviously, the second option is more desirable because it permits you to leverage, or get the most out of, a given investment in data processing power. But one significant ingredient is needed to make the techniques of timesharing and distributed processing work—communications.

The most desirable conceivable form of communications would be an invisible path between two or more separated points—invisible in the sense that the path does not contain any impediments to the easy, natural flow of information among the points—and this is the goal toward which all of modern communications technology is driving relentlessly. Unfortunately, the ideal "invisible path" is still several years in the offing because the "path" consists of evolving electronic techniques, microwave techniques, laser and optic techniques, hardware components, computers and computer software, and many other essentially disconnected disciplines that are being forcibly combined to satisfy the demanding pressures for universally available computational power.

ORGANIZATION

The service concept of DATAPRO COMMUNICATIONS SOLUTIONS is based on a set of problems distilled from the experience of Datapro's editorial staff with the needs of a broad spectrum of typical communications users. Each problem is addressed directly through a report that offers the best currently available solution to the problem. Each solution is carefully selected from among perhaps dozens of candidate solutions according to stringent criteria of clarity, applicability, and strength. The reader thus benefits both from the broad experience of Datapro's staff and from the well developed ability of the staff to recognize an optimum solution to a given problem. Each report is a problem/solution couplet. The problem is clearly defined in a prefatory statement to each report, and the body of the report is the selected solution. The reports are organized in nine major sections that reflect Datapro's perceptions of the problems a reader would typically encounter beginning with no communications system and gradually progressing through all the steps of basic familiarization, initial planning, design, acquisition, maintenance, management, and concluding with guidelines for future technology assessments. The reports are grouped as shown in Table 1 to address each of the preceding steps.

Step	Section(s)
Familiarization: Identify and explain the techniques and equipment that form the combined communications path	CS10 and CS15
Planning and design: How to analyze and relate your communications needs to the available techniques and equipment and how to develop paper designs for a workable system	CS20 and CS25
Acquisition and Maintenance: How to obtain the hardware and software identified by your design; how to install the system; and how to set up a maintenance subsystem	CS30 and CS35
Operations Management: How to manage the day-to-day problems of communications systems management	CS40
Systems Management: How to handle the larger problems of total system evaluation, applications, and internetwork relationships	CS50
Future Systems: How to assess and prepare for new technologies and new applications and how to evaluate the more far-reaching effects of pervasive communications facilities	CS60

Table 1. General organizational structure of DATAPRO COMMUNICATIONS SOLUTIONS

Auxiliary, nonsolution-type information is concentrated in Sections CS90 through CS98. Major communications equipment suppliers are listed in CS90; the significant communications standards and protocols are described in section CS93; and an extensive glossary of communications-related terms is offered in Section CS98. Section CS99 is reserved for the monthly newsbriefs supplied as part of the total subscription service, and Section CS05 explains the inquiry service, which is also part of the total subscription service.

Finally, a comprehensive index is given in Section CS01. The index is augmented by supplementary behind-the-tab indexes that are the first pages of each section CS10 through CS60. Each behind-the-tab index lists the Section reports by major problem topic category and provides a "See Also Report..." column offered to help the reader make connections to related reports throughout the service without having to resort to an index search.

HOW TO USE THIS SERVICE

The overall organizational structure of DATAPRO COMMUNICATIONS SOLUTIONS parallels the learning process for someone who enters communications as a novice (Sections CS10 and CS15) and who evolves into a sophisticated designer/user/manager of communications systems and networks (Sections CS50 and CS60). But the sectional breakdown of the service accommodates any level of reader sophistication because, although the reports are woven together loosely into a tutorial structure, each report essentially stands by itself. For example, a reader who is already familiar with communications techniques and equipment and who has already planned and paper designed a system may need some concrete advice on how to evaluate the relative economics of various design approaches. A quick scan of the behind-the-tab index to Section CS25 (Systems Design) yields the title "How to Evaluate the Economics of System Design Alternatives" (Report CS25-810-101) in the general problem category of Economic Analyses. The index entry is linked to other reports that explain how to negotiate contracts to provide the next link in the sequence of acquiring and assembling a system. This example provides a specific solution to a specific problem. The reader can also "browse" to establish points of reference for further planning. For example, a manager with an existing communications network may want to explore other possible applications of the network and to make some broad assessments of their utility and cost. A quick scan of Sections CS20 and CS50 will yield several titles, general and specific, for applications such as distributed processing, electronic mail, and Electronic Funds Transfer systems. These reports provide a convenient entry point for further investigation, and each report generally concludes with a comprehensive list of references that points the reader to a larger body of useful literature.

Finally, the reader is encouraged to use the Inquiry Service, which is a direct line to the Datapro editorial staff, for solutions not addressed directly by DATAPRO COMMUNICATIONS SOLUTIONS or for further, more specific solutions assistance to any problem covered by the service.□

Inquiry Service

DATAPRO COMMUNICATIONS SOLUTIONS is not just a two-volume set of books. It is, by contrast, a multi-part total information service of which the volumes of reports comprise just one part. This section of the COMMUNICATIONS SOLUTIONS service outlines another component of the total service, namely, the Inquiry Service. You will here discover what the Inquiry Service is all about, how it can help you solve your communications organization's problems, and how to use it.

DATAPRO COMMUNICATIONS SOLUTIONS is a total information service dedicated to helping communications management-level personnel solve the business- and management-oriented problems associated with organizing and running an automated information systems installation. As such, COMMUNICATIONS SOLUTIONS is a four-part service comprising 1) the two volumes of published reports, 2) monthly updates and supplements to these volumes, 3) monthly news and commentary through the Newsbriefs newsletter, and 4) the Inquiry Service.

The unique Datapro Inquiry Service is provided to our subscribers in order to make certain that no information need goes unsatisfied. If your required information cannot be located in the published reports, after having been researched with the aid of the Index, then you can use the Inquiry Service to get the required information directly from Datapro.

Typical kinds of information you might seek through the Inquiry Service include:

- Supplementary technical or background information to that provided in the published reports.
- Information on your problem that may not yet have been addressed by a specific communications SOLUTIONS report.
- Personalized professional advice regarding your specific problem at hand. This kind of dialogue can go well beyond any of our published reports in terms of its orientation and depth.

HOW TO USE IT

You can get your answers through the Inquiry Service in two ways: 1) by calling Datapro on the telephone; and 2) by using the unique Inquiry Service forms that follow and mailing us your queries.

TELEPHONE ACCESS: You can gain immediate and direct access to our unparalleled staff of in-house analysts, consultants, and editors simply by calling us at (609) 764-0100 and asking for the Director of Customer Services. Upon connection, you need simply state your problem and we'll either provide the required information while you're on the line, or we'll call you back, usually the same day, with your answers.

As a subscriber, you can use this telephone service freely. There is no limit on the number of times you may call during your subscription year. The service is included in your annual subscription fee. (If you should have a problem that entails original research to a number of distant information sources, we may ask you to pay only the cost of the toll calls.)

MAIL ACCESS: You can also get answers to less time critical problems by using the Inquiry Service forms that follow. To use these forms, simply jot down your question or problem, remove the form from the volume, and mail it to us, postage paid. The same group of in-house analysts will review your questions and respond, with your required information by telephone or mail, and again, usually on the same day that your form is received. These forms will prove

Inquiry Service

especially helpful to you when you want to describe a complex business-related problem that is simply more suitable for communication in writing rather than by telephone.

You will also find the following Inquiry Service forms convenient for use in general correspondence with us regarding comments, questions, and suggestions you may have concerning the DATAPRO COMMUNICATIONS SOLUTIONS information service. As a Datapro subscriber, all your needs will get immediate attention.

You will find six Inquiry Service forms in your initial volume. Since there is also no limit on this aspect of the service, you may write to us at any time for a new supply or forms as needed.

SUMMARY

To conclude, the Inquiry Service puts our vast data bank of management- and product-related information at your fingertips. Upon your request, this data is researched and put into perspective by the computer industry's largest and most experienced in-house staff of independent analysts and consultants. In that sense, the Inquiry Service is your key to this wealth of experience and data. We encourage you to use it. As the leader in our field, we are committed to getting you the information you need.□

Inquiry Service Form

Dear Subscriber:

Here's how you can get quick answers to your business- and management-level communications problems. Please state your problem or information need as concisely as possible in the space below, identify yourself at the bottom section of the form, and remove the postage paid form from the volume. Then just fold it, seal it and drop it in the mail. The information you need will be returned to you at once, either by telephone or by mail.

Name: ______________________ Address: ______________________

____

Organization: ______________________ Telephone: ______________________

☐ I'm a current subscriber. ☐ I'm evaluating the service for a possible subscription.

____

(signature)

PLEASE USE THIS SPACE TO ADD ANY COMMENTS OR OBSERVATIONS THAT MIGHT BE HELPFUL TO OTHER USERS OF PLUG-COMPATIBLE OR OFF-LINE PRINTERS.

FOLD HERE

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY CARD

FIRST CLASS PERMIT NO. 178 DELRAN, N.J. 08075

POSTAGE WILL BE PAID BY ADDRESSEE

datapro

DATAPRO RESEARCH CORPORATION
1805 Underwood Boulevard
Delran, New Jersey 08075

Attention: Telephone Consulting, Communications Solutions

FOLD HERE

Topic Index—Section CS10

General Topic	Report Title	Report No. CS10-	See Also Report
Technology Overviews	—A Brief Historical Perspective of Computer Communications	-120-101	
	—Teleprocessing—The Modern Marriage of Computers and Communications	-150-101	CS25-110-101
The Data Trans-Mission Channel	—The Special Requirements of Data Communications Traffic	210-101	CS15-220-101
	—The Basic Parameters of a Data Channel	-220-101	
	—General Solutions to Common Communications Interface Problems	-230-101	CS20-235-101
Multiplexing and Concentration	—The Strengths-and Applications of Digital Data Multiplexing	-310-101	CS15-320-101; CS40-410-101
	—The Power of Concentration	-320-101	CS15-420-101
Digital Transmission Techniques	—The Language and Techniques of Pulse Code Modulation (PCM)	-410-101	
	—The Nature and Organization of Digital Transmission Channels	-411-101	
Protocols	—Why Protocols? Why Standards?	-510-101	CS10-810-101
	—A Tutorial on Protocols	-510-201	CS10-810-101
Network Technologies	—A Concise Summary of Public, Leased, and Private Transmission Facilities	-610-101	CS15-710-101; CS15-711-101
	—The Uses of Message Switching in Communications Networks	-620-101	CS25-310-101
	—The Techniques and Applications of Packet Switching	-650-101	CS25-411-101; CS25-415-101
	—Operating Systems for Computer Networks	-660-101	
Wideband Transmission Facilities	—Understanding Satellite-Based Telecommunications Systems	-710-101	
Communications Control Techniques	—A Summary Overview of Communications Network Control Architectures and Protocols	-810-101	CS40-320-101; CS60-510-101

A Brief Historical Perspective of Computer Communications

Problem:

It may seem difficult to gain a historical perspective of a technology that is barely 20 years old, but if one considers that most of the other fruits of modern technology are scarcely 80 years old, then any history of technology must be considered on a different time scale than, say, the History of Western Europe.

The benefits gained by learning the history of anything are probably more comforting than real, although sometimes we might learn something useful from the records of false starts and mistakes. This report offers a brief chronicle of these records, but more important, it clearly identifies the evolutionary trends in computer communications. It analyzes their growth to the present and extrapolates their growth directions into the near future. This sort of broad "Where we were; where we are; where we're going" overview of a technology supplies a very important basic rationale for system planning because the system you build must ultimately merge with the mainstream of the technology and may help, in some small way, to shape tomorrow's history.

Solution:

HENRY FORD WAS WRONG

Of course, Henry Ford was wrong. History is not bunk. It just tends to look like bunk in the short range. Legitimately, historians must allow some time for the confusion of events to die away. They can then evolve theories about fading memories of the events. In his short story, "The Ugly Little Boy," Isaac Asimov[1] has a reporter say the following about a machine that recovers people from the past and makes them live in the present. "You can only reach out so far; that seems sensible; things get dimmer the further you go; it takes more energy. But then, you can only reach out so near." It is the same with history.

"Computer Communications—How We Got Where We Are" by Ivan T. Frisch and Howard Frank, Network Analysis Corp. from *Computer Networks and Communications,* edited by Robert R. Korchage.

Accordingly, some of the great historians still have only little to say about the computer revolution. Arnold Toynbee, for example, is still involved in the purely negative aspects of the revolution. His chapter on computerization[2] is called "Mechanization, Regimentation and Boredom"; this brings to mind some advice for fourteenth century magicians, "If you want to be a successful prophet, prophesy evil."[3] Daniel Boorstin, winner of the Bancroft Prize, the Parkman Prize, and the Pulitzer Prize for his penetrating series of books, "The Americans," is most fascinated by the gadgetizing of Americans: "When automation became widespread and electronic computers became almost as common as the adding machine, there were new cataclysms in the jobs of Americans and in their ways of thinking. By 1967, only a half-century after the first commercially successful billing machine, the annual American production of cash registers and computing machines totaled more than $4.5 billion.

A Brief Historical Perspective of Computer Communications

When precise and up-to-date information was available about the quantities of everything, businessmen and consumers could not help thinking quantitatively."[4]

The facts are right, but the impact is trivialized. This nearsightedness, being fairly general among historians, we therefore seek for the general history and impact of computer-communications elsewhere. We must search among the participants, namely ourselves, and among other commentators, who will be broadly classified as journalists. One must be wary of Marshall McLuhan's generalizations. After all, in their book, "War and Peace in the Global Village," McLuhan and Quentin Fiore attribute the age of chivalry to the invention of the stirrup.[5] Granted that McLuhan is not a master of understatement, one can still find truth in his estimation in the same book: "The computer is by all odds the most extraordinary of all technological clothing ever devised by man, since it is the extension of our central nervous system. Beside it, the wheel is a mere hula-hoop." One of the best journalist historians is James Martin. After all, he published a book in 1971 called, "Future Developments in Telecommunications," and much of this book, intended as almost science fiction for the year 1980, is a good history of the years 1971-1974.

SEPARATING THE USERS FROM HIS COMPUTERS

In 1939 Aikin and a team of IBM engineers at Harvard began the work that resulted in 1944 in the Mark I—the first automatic electromechanical digital computer. The first completely electronic computer was designed by Eckert and Mauchly at the University of Pennsylvania, for the Ballistic Research Laboratory at Aberdeen. The ENIAC (Electronic Numerical Integrator and Calculator) became operational in 1946. The history of computing in the 30 years since Mark I is a monumental one, which will require some new historians to record. For the present, we will try to simply indicate some of the trends and milestones in the more limited area of computer-communications, or computer networking or, in simpler terms, the process of separating the user from his computer. We will subdivide this process into two categories—terminal oriented networks and the area with the shorter history, but greater technical promise, computer-to-computer networks.

EVOLUTION OF TERMINAL-ORIENTED NETWORKS

The first computer network consisted of a computer with several cables attaching input devices. A majority of the networks in the world are still of this type. The need quickly arose to do more than just communicate with a computer 100 feet away, and remote terminals were added to the network. The networks were first extended to cover all of the buildings within an industrial complex on leased or specially constructed lines. The capability to dial into the main frame computer was then added, and the networking era began in earnest (see Figure 1).

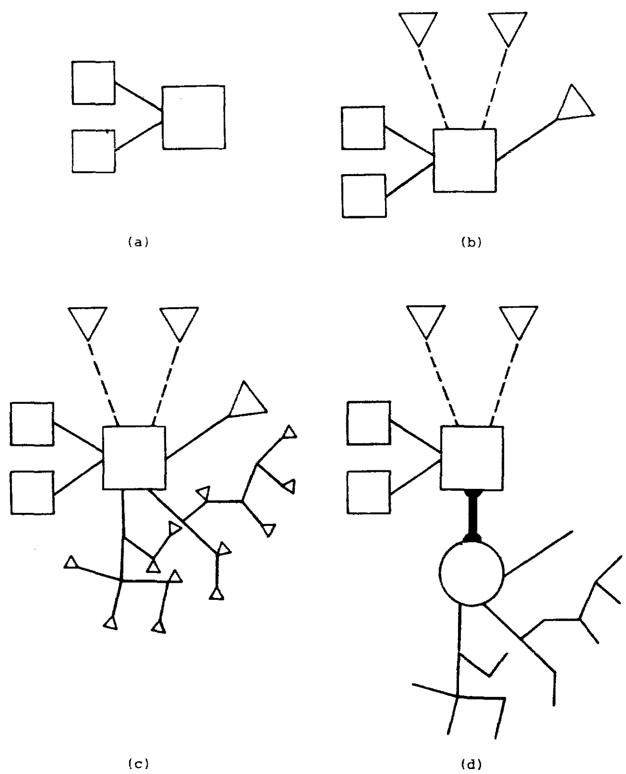


Figure 1. Evolution of terminal-oriented networks

As networks grew, their costs also grew, often quite rapidly. For example, as more and more demands were made on the system, the cost of the communications became a very significant fraction of the cost of the overall network. Originally, the computer represented the majority of the total system cost. But, as the network expanded, communications often exceeded 50 percent of the overall system cost. Therefore, efforts began to reduce this aspect of the overall cost. Innovations like multidrop lines, which allowed a number of different terminals to share a common line, were introduced to take advantage of all possible economies of scale. You might be able to lease a very low bandwidth line for, let's say, a thousand or fifteen hundred dollars per month. On the other hand, you could probably increase the capacity of the line by a factor of ten or more at a cost increase of only a factor of two. This provided sufficient capacity to allow sharing of the line by several terminals. But to do this, control mechanisms for selecting different terminals on the line and for protecting data had to be invented, and techniques for contention resolution and queueing were required.

A Brief Historical Perspective of Computer Communications

The next major difficulty encountered in building computer networks were the changes to the main frame software, which were found to be exceptionally difficult and costly. Thus, to reduce the time and cost of system development, devices called "front ends" were introduced. These allowed the communication functions of the computer network to be separated, by and large, from the processing function of the computer. Front-end use grew very rapidly, beginning in the late 1960's and was assisted by the introduction of low cost minicomputers. Today, front ends play an important role in network communications.

Next, the interesting observation was made that there was a cable between the front end and the computer. Since large networks always tend to get larger, the cable became longer and more communication equipment was required between the front end and the computer. As the front end increased its distance from the main frame, its name changed to that of "concentrator." In modern networks, concentrators may be thousands of miles from the computer. Their main function is to reduce communication cost by more effective communication line utilization. The next development was quite natural; another front end was added to the computer side of the network to complete the isolation of the computer from its network elements.

In Figure 2, we have a typical structure of a terminal-oriented network[6]. This particular network is called the NASDAQ System. "NASDAQ" stands for the National Association of Securities Dealers Automated Quotations System. This network was built in 1970 and became operational in 1971. Its function is to collect quotation information about the Over-the-Counter Securities market. Users distributed throughout the country receive responses to their input in five or six seconds. Responses contain information about the prices at which dealers are willing to sell or buy securities, and the exact bid and ask prices of each market maker who deals in a particular security. There are on the order of 1,700 terminals in this system at a

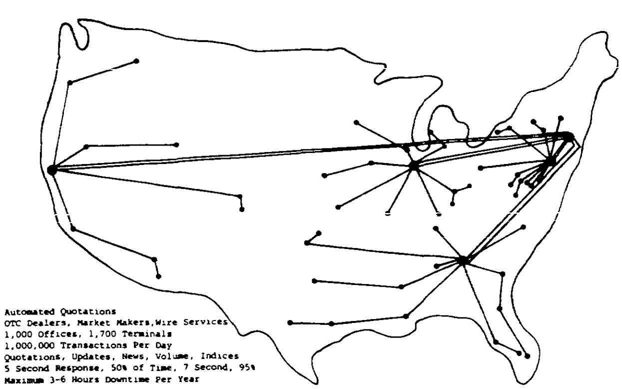


Figure 2. Simplified network diagram for the NASDAQ system

thousand different locations in about 400 different cities. The system has reduced the problem of getting the information about Over-the-Counter stocks from ten phone calls to a single network message. During active trading days, the NASDAQ System has handled more than one million messages a day.

MILESTONE TERMINAL-ORIENTED NETWORKS

There are almost as many terminal-oriented systems at present as there are computers, since almost every computer has terminals attached to it. And almost all these systems fit somewhere into the evolutionary pattern we have described. However, only a small number of these networks set milestones in either timing, structure, function or size. Those that have been major benchmarks fall into two general categories: special purpose networks—intended to serve a specific function for a selected set of users—and time sharing services—intended as a general utility for any user.

SPECIAL PURPOSE NETWORKS

Military

The military has been one of the leading users and pioneers of special purpose networks. Indeed, much of the technology developed for military purposes has been transferred and adapted for commercial use. The prime examples are point-of-sale systems, of which banking and airline reservation systems are pioneering areas. Other users such as educational institutions have also added major improvements necessitated by their particular requirements. Certainly a milestone in military systems and in computer communications development, in general, is the SAGE (Semiautomatic Ground Environment) system. Lest we forget in how many different ways this system was a pioneering effort, I will quote Ruth Davis:

"The first use of an automated display which permitted the user to exercise control over the information presented (and also to enter requests and information based on what was presented to him) occurred in the SAGE system (Figure 3). The significance of the introduction into this system of the light gun as a pointing device under the control of the display operator cannot be overemphasized. It was probably the one most important event which made possible the man-computer interaction deemed so essential at the present time. It occurred in 1952 utilizing the Whirlwind computer."[7]

But let us look at the computer communications aspects. The purpose of the system was air defense for the U.S. The results were benchmark efforts in computers, communications and computer communications.

A Brief Historical Perspective of Computer Communications

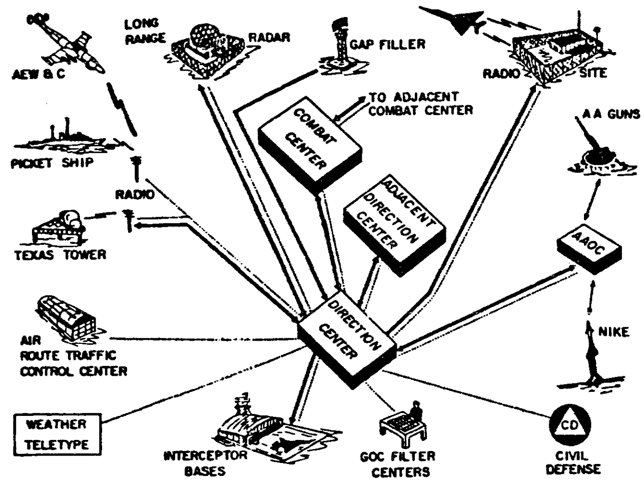


Figure 3. Typical sage sector. __ Data circuits. . . . . Other circuits

The Air Defense System Engineering Committee (ADSEC), a group formed by the Scientific Advisory Board at the request of the Air Force, evaluated the status of overall air defense in the 1950's. They recommended initial feasibility tests utilizing digital radar inputs to a central computer. This was to be accomplished by coupling the data-processing capabilities of the Digital Computer Laboratory to the radar data-transmission techniques of the Cambridge Research Center. Favorable results led to Project Charles and the establishment of Lincoln Laboratory in 1951 with a character to work toward a computer-based air defense system. Project Charles activities led to recommendations for a prototype test facility known as the Cape Cod System, which was established in 1952.

The New York Air Defense Sector became the first operational site in 1958. By 1963, SAGE Direction Center and Combat Centers had been installed at all continental stations. The system was designed in 1955 with IBM AN/F SQ-7 prototype computers, with SDC software at the central facilities. Each computer contained 58,000 vacuum tubes, consumed 1,500 KW of power and occupied an entire building floor.[8] Radars and information sources fed information to the centers, and the centers sent information to the interceptors and other weapons. Real time processing required key developments by many companies, small computer (not minicomputers) front-end processors, specification of 1600-baud data lines with better conditioning than voice grade lines, and redundant diverse routed paths for reliability.

Banking

The development of commercial systems such as banking could be done on a smaller scale and hence had less auspicious milestones. Certainly, the first of any system must be a milestone. The first banking milestone therefore sounds almost like an entry from the Guiness book of records. Telefile is described by Sackman[7] as the first online banking system in the world, linking the transactions of each of the three participating banks and their affiliated branches into a central data-processing system. This system grew out of automation feasibility studies initiated by the Howard Savings Institution of Newark, New Jersey in 1953. By 1956, system requirements were specified, two other banks cooperated in the venture, and the Teleregister Corporation was awarded the contract for developing and implementing the data-processing system.

The three main system requirements were as follows:

1. Online data processing at the teller window—for example, direct communication between the teller and the central computer for deposits and withdrawals.

2. High system reliability and accuracy commensurate with rigorous banking standards.

3. Uninterrupted continuity in banking service throughout the transition period from the initial manual system to the successor semi-automated system.

The system is a long way in scope from present broad purposed vast networks such as that of the Barclay Bank or that being considered by the Federal Reserve Board, but it was the beginning.

Airlines

One of the earliest large scale users of point-of-sale type systems has been the airlines (See Figure 4 for sample configuration). As Janet Taplin[9] has commented "American Airlines has been uniquely successful in its use of computers. Its SABRE I was the first on-line reservation system and represented a major breakthrough in terms of real-time computer usage". A joint research effort by IBM and American Airlines in the early 50's culminated in the SABRE system in the early 60's. The system consists of a central computer site with 2000 nationwide terminals multidropped to the central site.[10]

Education

One of the earliest and most ambitious educational networks is the Dartmouth Time Sharing System (DTSS), first placed in operation in 1964.

"It was . . . decided that exposure to computing and free availability of computing should become a standard part of the liberal arts educations at Dartmouth, an undergraduate college where only 25 percent of the students elect majors in the sciences and engineering.

A Brief Historical Perspective of Computer Communications

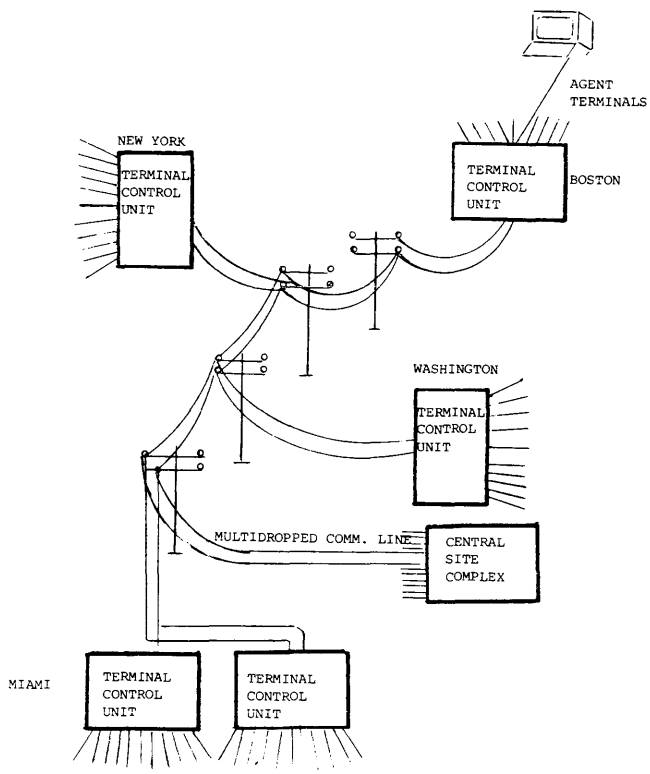


Figure 4. Airlines reservations system

. . . Against this background, it was recognized that the user-computer interface had to be simplified and harmonized with the educational environment if liberal arts students were to ingest a reasonable dose of sensible knowledge about computing. Two important consequences of this recognition were the decisions to bring the computer to the student via remote individual terminals (teletypes) and to devise an extremely simple user interface."[11]

The system evolved through several stages of hardware and software systems as well as communications. The use of DTSS by schools outside Dartmouth developed sporadically until given a major impetus in 1967-1968 by NSF Grants. The configuration in 1968 is shown in Figure 5.

TIME-SHARING NETWORKS

The emergence of time-sharing systems as general purpose on-line computing facilities is a development primarily of the 1960's. Some of the early experimental work took place at Project MAC at MIT; SDC under the aegis of ARPA; and RAND. By the mid-1960's, practically all computer manufacturers were marketing or developing some form of time-sharing facilities. A number of organizations now run commercially available time shared services. Among them are United Computing Services, Inc., Utility Network of America, and so on.

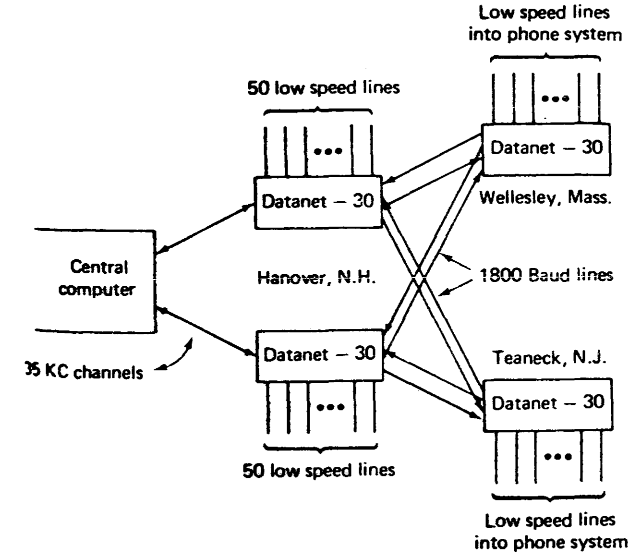


Figure 5. Dartmouth time sharing system interconnections between remote and local communications computers (1968)

The most significant networks are unusual in function, size and complexity.

One of the largest time-sharing network is run by General Electric (Figure 6).[12 13] The system evolved from GE's experience with the Dartmouth Time Sharing System and in 1965 used the operating system developed at Dartmouth.

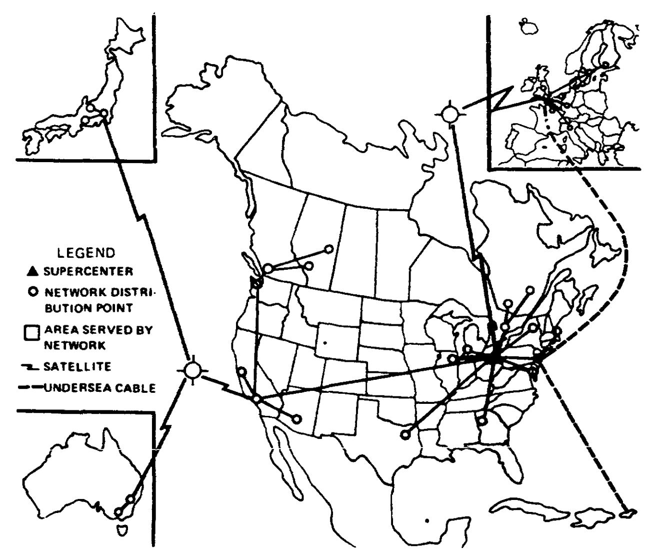


Figure 6. General Electric international network

The most sophisticated time-sharing networks currently in operation is TYMNET (Figure 7) owned by Tymshare, Inc.[12 14] The network employs 80 communications processors all over the U.S. accessing 26 host computers. The network configuration consists of a backbone of multiple rings, rather than a

A Brief Historical Perspective of Computer Communications

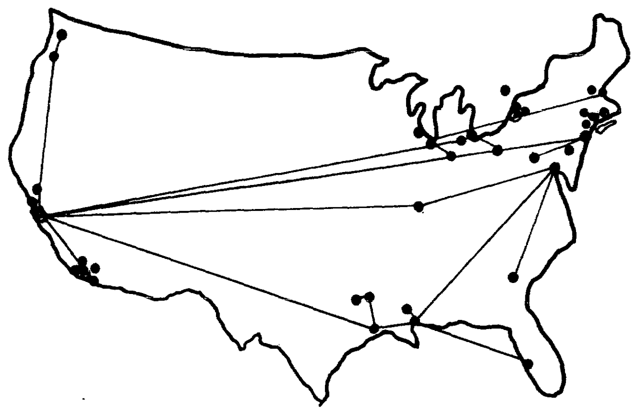


Figure 7. Tymnet

star, with other nodes connected in stars or straight runs. If one path to a computer is saturated or down, the network automatically switches to an alternate path. The network goes far beyond the concept of individual real time terminal users and services entire organizations such as major accounting firms and the National Library of Medicine.

COMPUTER-TO-COMPUTER NETWORKS

Parallel to the development of terminal-oriented systems, efforts were under way to allow computers to communicate directly with other computers in real time. The first step was, of course, to place two identical computers in the same building and to connect a cable between them. (Many of the computers being built today can be regarded as sophisticated computer networks in themselves.) To assist in this difficult task, devices very much like front ends were developed to handle the communications functions and other chores needed. Naturally, the communication lines became longer, necessitating communication hardware at the ends of the line.

A result of this approach is star-like networks (Figure 8) with a store-and-forward central switch. A significant network in this category is the AUTODIN System.[15] AUTODIN was built and is maintained and managed by Western Union for the U.S. Government.

An extension of this type is the ring computer network in which a front-end type device (often called a network interface processor) connects the network lines and the computers. Data for a computer is addressed to that computer and sequentially sent, link by link, in a circular fashion. At each step around the circuit, the data is interrogated by the interface processor, and when it finally reaches the interface processor connected to the destination computer, it is removed from the ring. Naturally, if a network like this is not planned very well, data may eventually circulate for-

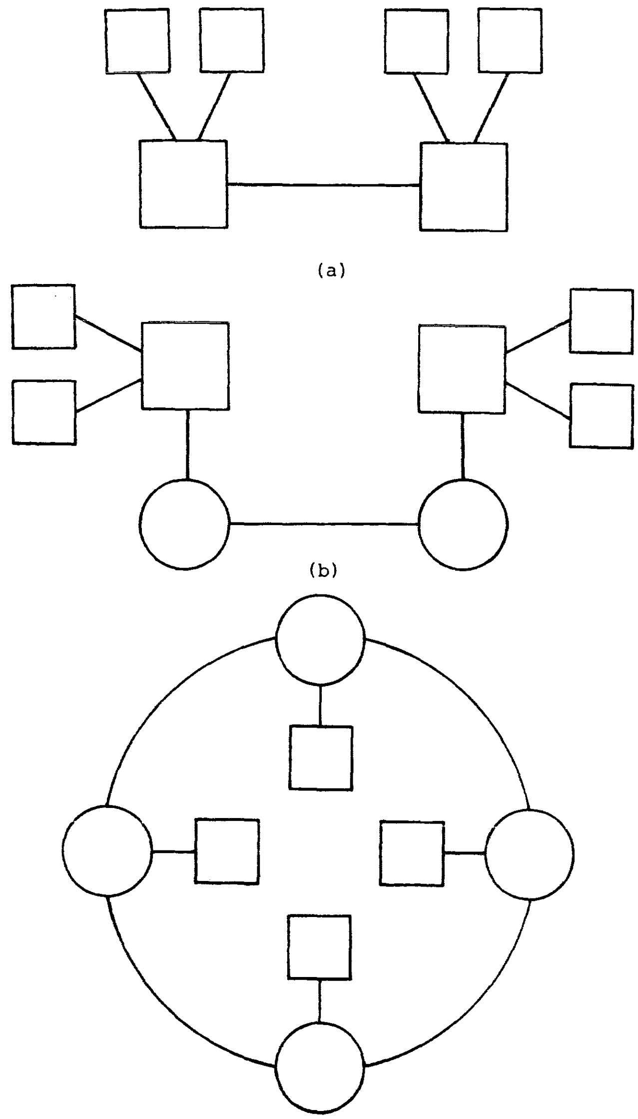


Figure 8. Evolution of computer-to-computer networks

ever. Thus, control devices to remove data which is "too old" from the network must be placed in the network. In addition, as such a network grows, its reliability can become very low because all elements along the ring must operate for the network to operate. Therefore, additional lines for redundancy and more flexible routing techniques must be added for effective operation.

A more ambitious type of system is called ARPANET.[16, 17] The concept of this system was to provide high flexibility by allowing any kind of interconnections and adaptive routing of information. In late 1969, the first four elements were installed on the West Coast. The network grew to about a 25 node system in 1971, to about a 40 node system in 1973, and today has well over 50 nodes (Figure 9). This network

A Brief Historical Perspective of Computer Communications

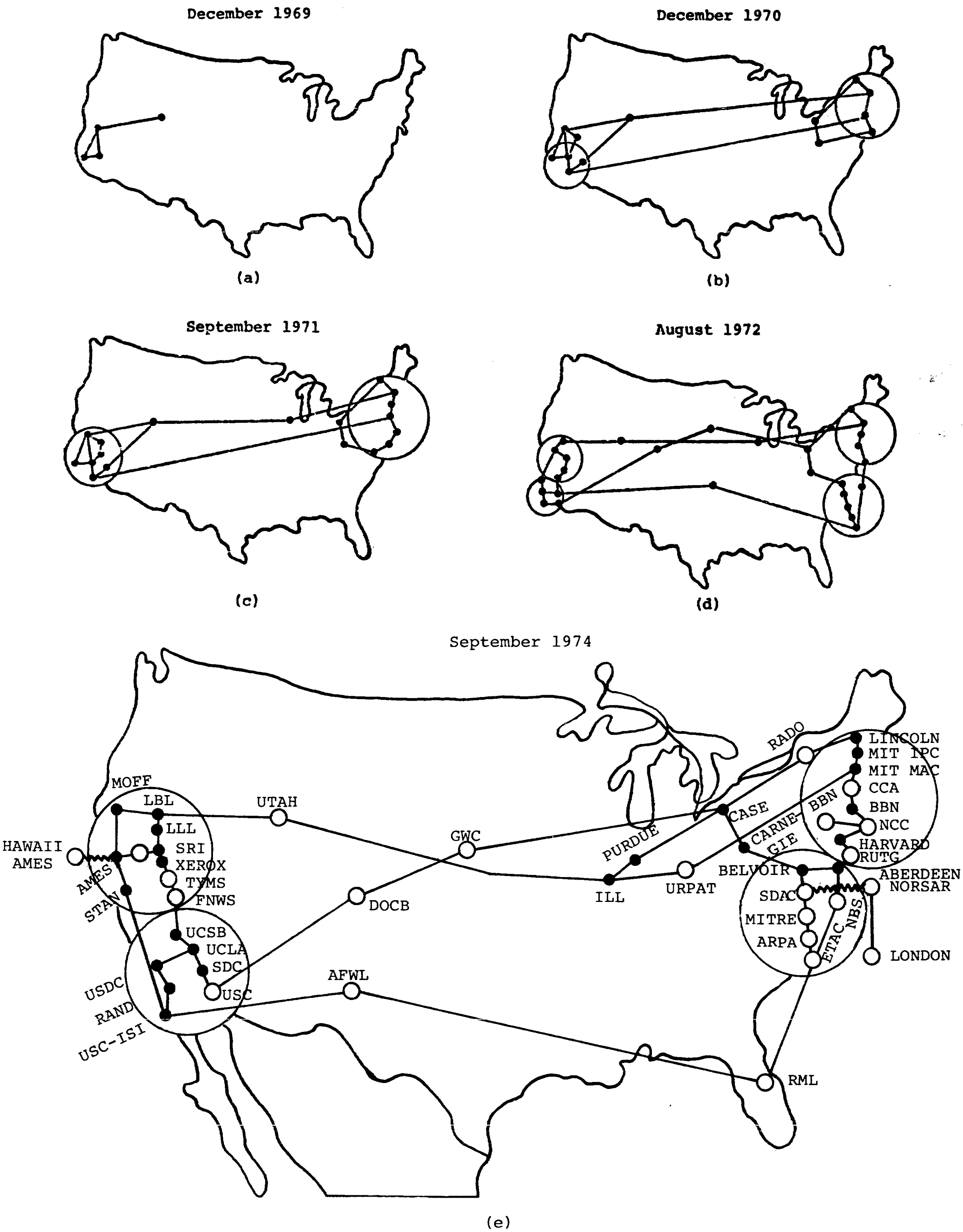


Figure 9. Geographical expansion of the ARPA network

is one of the first major applications of the new technology called "packet switching" in which data is broken up into blocks that are separately addressed and then allowed to make their way independently through the network from origin to destination. This type of network must handle the problem of controlling flows using a "distributed" control scheme.

The ARPANET significantly differs from the centralized system approach. In a centralized system such as NASDAQ, nearly all the controls reside in the central computer. If it cannot handle the flow, the computer will slow down the concentrators and do whatever else is necessary to prevent additional calls from being sent. In a distributed network, very sophisticated techniques of flow control and routing adaption in case of a line or node failure had to be developed. Packet switching is now viewed as a major addition to the technology of computer networking, and has already been applied to radio communications.[18] A number of other networks are now being built or designed based on the packet switching technology of the ARPANET, not the least of which is IBM's System Network Architecture (SNA), and the future of the field appears quite bright.

PROPHECY

Clearly an important part of the computer communications revolution has been the proposal and development of an incredible array of digital services. This includes new technical offerings and tariff structures by the common carriers dominated by AT&T[19] and Western Union.[20] A further development of crucial interest to the computer industry is the growth of the specialized common carriers including MCI, DATRAN, and a large number of regional carriers such as Western Tele-Communications. The picture is further enhanced by the addition of value added networks and satellite communication. These topics are covered in more detail throughout the body of this service.

Our mandate does not include prophecy—for evil or for good. But after all the only reason for knowing "How we got there," is so we can extrapolate to "Where we are going." Some things are certain. As Fano says "The 'Marriage' of computers and communication has been celebrated and consummated. But now the honeymoon is over, and the two partners are beginning to face the realities of their interdependence."[21]

Looking into the very near future, networks are planned that tend to combine the distributed network control concepts of ARPANET for computer-to-computer communications with the centralized NASDAQ-like approach for terminal-to-computer and terminal-to-terminal communications. These networks are an extension of the multidrop centralized net where now the terminal processor replaces the computer, and the backbone communications is then through a packet-oriented net like ARPANET. An example of this type of net is shown in Figure 10. This particular example is a sample design for a planned FAA Air Traffic Control Network. This network has 21 air traffic control computers at appropriate locations. It has a backbone communication network which is a simple loop like network. Emanating from the nodes of this network is an extensive terminal communications network, which is itself a collection of networks.

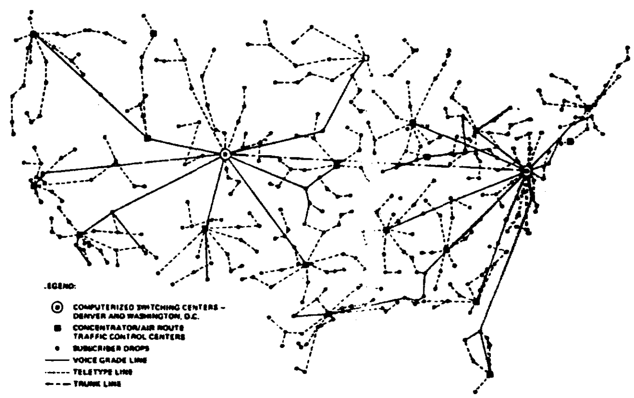


Figure 10. A computer plotted output for a data communication network of 500 locations

The growth of computer communication networks has clearly left the linear part of its presumed exponential growth. In-house systems or intercorporation facilities abound not only on paper but in actual implementation. In addition many more facilities are on the horizon. For example:

- In Canada, the Datapac Network is a nationwide, packet switched, shared, data network, which has been designed to become the basic Canadian network for data communications. There are four network nodes: Toronto, Montreal, Ottawa, and Calgary. These four nodes, or networks switching centers, will initially serve the entire country. By 1980, at least fourteen Canadian cities will have network nodes. After 1980, the network will continue to expand to meet Canada's data requirements.

- Also in Canada plans are being developed for CANUNET, Canadian Universities Computer Network, a packet switched network sponsored by the Ministry of Communications to link some 20 universities.[22]

- An international effort is planned by the Organization for Economic Cooperation and Development. The result is to be a European data communica-

A Brief Historical Perspective of Computer Communications

tion network between certain universities and research centers. This network, which will work on the "packet switching" principle, in reminiscent of the ARPA network. Secondary networks can be connected to nodal centers. Nodal centers will exist in Italy, France, Switzerland, the United Kingdom, and within the OECD administration. Norway, Sweden, Portugal and Yugoslavia have also joined the project.[23]

Beyond extrapolation we indeed enter the realm of prophecy. We can only list a few achievements we all know are here or on immediate horizon, make an obvious observation, and relate a personal experience.

First the list. The following developments are here:

- Minicomputers
- Programmable calculators
- Hand calculators
- Microprocessors
- Hand held radio transmitters
- Cable TV system for data transmission

Second, the obvious observation. Even without looking into the far future of hand held minicomputers on a chip or optical fibers, it is clear that computer networks will soon look nothing like they look now. Mobile users with hand held terminals dialing into vast networks of minicomputers and maxicomputers, with little difference between front ends and processors, is clearly possible.

Finally, a personal experience; as usual, one of us did his Christmas shopping on Christmas eve. He was at the counter at Macy's trying all the calculators, using one calculator to calculate the cost per feature on all the other calculators at the latest bargain price of overstocked Japanese calculators with Italian names. A woman standing next to him, silent for many minutes, finally got up the courage to ask the salesman what memory was used for on a calculator. He tried to explain several times and failed. Finally, he showed her how it was used to store an intermediate answer. A glow of discovery appeared on her face. For the first time after years of propaganda, advertising, and intimidation about computer memory banks she understood what memory was. A new American became intimate with the computer. This element of citizen acceptance of the computer when combined with the technical elements make a new revolution both inevitable and unpredictable.

Many others are, of course, actively speculating on the effect of the computer communications revolution on society. Some of this speculation is didactic. Says Peter Goldmark,[24] "What I propose is that the advances of telecommunications technology—satellites, cable TV, broadband circuits and similar devices—make it possible to attract future generations into the smaller towns of America beyond the commuting dependency range of the big city and suburbs and thus cut down on the excessive use of power." Some of the speculation is more ruminative. Says Paul Baran,[25] "The key man in the new power elite will be the one who can best program a computer, that is, the person who makes the best use of the available information and the computer's skills in formulating a problem. In a world where knowledge is power, and where communications mean access to power, he who can most effectively utilize this access will be in the driver's seat. Some persons (primarily computer programmers) claim that the richest man in the world in the year 2000 will be a computer programmer. This may sound outlandish, but few really good programmers laugh when they consider this assertion."

But the best appraisal is by Steward Brand,[26] humanist author of "The Whole Earth Catalog." In his essay, "Fanatic Life and Symbolic Death Among the Computer Bums," he sums it all up, "Ready or not, computers are coming to the people."

REFERENCES

[1]Asimov, I., "The Ugly Little Boy," *Nine Tomorrows*, Doubleday & Co., Inc., 1959.
[2]Toynbee, A. J., *Change and Habit: The Challenge of Our Time*, Oxford University Press, 1966.
[3]Thorndike, L., *A History of Magic and Experimental Science: Volume III and IV, Fourteenth and Fifteenth Century*, Columbia University Press, Second Printing, 1953.
[4]Boorstin, Daniel J., *The Americans: The Democratic Experience*, Random House, 1973.
[5]McLuhan, M. and Q. Fiore, *War and Peace in the Global Village*, McGraw-Hill, 1968.
[6]Frank, H., I. T. Frisch and R. Van Slyke, "Testing the NASDAQ System—Traffic and Response Time," *Proceedings of the Symposium on Computer-Communications Networks and Teletraffic*, Polytechnic Institute of Brooklyn, 1972, pp. 577-586.
[7]Sackman H., *Computer System Science and Evolving Society*, John Wiley, 1966.
[8]Enticknap, R. G. and E. F. Schuster, "SAGE Data System Considerations," *Transactions of the American Institute of Electrical Engineers*, January, 1959, pp. 824-832.
[9]Taplin, J., *A History of the Remote Access Computer Industry in the United States 1964-1972*, Master's Thesis, University of Pennsylvania, August 11, 1972.
[10]Knight, J. R., "A Case Study: Airline Reservations Systems," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1423-1430.
[11]Hargraves, Jr., Robert F., "The Dartmouth Time Sharing Network," *Computer Communication Networks*, N. Abramson and F. F. Kuo (Editors), Prentice Hall, 1974.
[12]Gaines, G. and J. Taplin, *Time Sharing Today*, Vol. 3, Nos. 1 and 2, May, 1972.

A Brief Historical Perspective of Computer Communications

[13]Mauceri, L. J., "Control of an Expanding Network An Operational Nightmare," *Networks*, Vol. 4, No. 4, 1974, pp. 287-297.
[14]Beere, M. P. and N. C. Sullivan, "TYMNET A Serendipitous Evolution," *IEEE Transactions on Communications*, Vol. COM-20, No. 3, June, 1972, pp. 511-515.
[15]Cox, J. E., "Western Union Digital Services," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1350-1356.
[16]Frank, H., I. T. Frisch and W. Chou, "Topological Considerations in the Design of the ARPA Computer Network," *SJCC*, May, 1970, pp. 581-587.
[17]Roberts, L. G. and B. Wessler, "Computer Network Development to Achieve Resource Sharing," *SJCC*, 1970, pp. 543-549.
[18]Abramson, N., "The ALOHA System," *Computer Communication Networks*, N. Abramson and F. F. Kuo (Editors), Prentice Hall, 1974.
[19]James, R. J. and P. E. Muench, "AT&T Facilities and Services," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1342-1349.
[20]Gerla, M., "Moving Bits by Air, Land and Sea," in this session.
[21]Fano, R. M., "On the Social Role of Computer Communications," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1249-1253.
[22]deMercado, J., R. Guindon, J. DaSilva and M. Madoch, "The Canadian Universities Computer Network: Topological Considerations," *Proceedings of the First International Conference on Computer Communication*, S. Winkler (Editor), October, 1972, pp. 220-225.
[23]Larsson, T., "Data Communication in Sweden and Some Aspects of the Situation in Europe," *Proceedings of the First International Conference on Computer Communications*, S. Winkler (Editor), October, 1972, pp. 17-25.
[24]Goldmark, P. C., "A Rural Approach to Saving Energy," *The New York Times*, Sunday, November 11, 1973.
[25]Baran, P., "On the Impact of the New Communications Media Upon Social Values," *Law and Contemporary Problems*, Vol. 34, No. 2, Spring, 1969, pp. 244-253.
[26]Brand, S., *II Cybernetic Frontiers*, Random House, 1974.
[27]Gaines, E. C., "Specialized Common Carriers Competition and Alternative," Telecommunications, September, 1973, pp. 17-26.
[28]Luther, W. J., "Conceptual Bases of CYBERNET," *Computer Networks*, R. Rustin (Editor), Prentice Hall, 1972.
[29]Schwartz, M., R. R. Boorstyn and R. L. Pickholtz, "Terminal-Oriented Computer Networks," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1408-1422.
[30]Worley, A. R., "The Datran System," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1357-1368.□

Teleprocessing—The Modern Marriage of Computers and Communications

Problem:

The most significant trend in communications is the convergence of communications and computer technologies toward a composite technology that permits processing to be done at a distance—teleprocessing—and that permits processing power to be distributed and interconnected through wider and easier communications links. This trend is being translated into the concrete applications realities of Electronic Funds Transfer Systems (EFTS), Electronic Mail, and Distributed Data Processing (DDP), to name a few. This report offers a broad introduction to the current scope of these applications and will help to clarify the connection between the basic technology and components discussed in this and the next section and the planning/design/management problems covered in later sections.

Solution:

APPLICATION BREADTH

As the saying goes, "We've come a long way." It was only in 1941 that telegraph paper tape information was first converted directly into punched cards to enter telegraph information into the computer. And it wasn't until 1954 that the Transceiver (a terminal attached to telephone lines) first transmitted punched-card data directly, without conversion to telegraph paper tape. But the late 1950s saw an explosion of techniques for remote computer usage. The SAGE Air Defense system, for example, drew digitized radar data from sites located over hundreds of miles to feed dozens of computer centers that were tied into a semiautonomous national computer network, involving one and a half million miles of communication lines and thousands of interactive display consoles. In 1962, SABRE, the first large, real-time, on-line airline reservation system, went into operation, linking 1200 reservation terminals to a central processing center.

Since then, teleprocessing has grown steadily. We see processors increasingly serving many remote users by remote job entry, with remote output of the results. We have seen the accumulation of large databases, which hundreds and even thousands of people can interrogate. We have witnessed the development of conversational techniques, whereby individuals can utilize remote processing power to develop new programs or procedures in a close interactive mode. We have seen, in general, an increasing sophistication in the tools made available to individuals for their use of data-processing resources in multiple locations.

Applications for teleprocessing systems are now extremely diverse. To illustrate:

1. On-line cash transaction applications and branch-to-central accounting in the banking industry
2. On-line freight loading, freight movement, and bill-checking information systems in the transportation industry

---

R.J. Cypser, *Communications Architecture for Distributed Systems,* © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

Teleprocessing—The Modern Marriage of Computers and Communications

3. On-line order processing and terminal-oriented bill of materials and inventory control in the manufacturing industry
4. On-line reservations in the travel and airlines industries.

These are from lists of many hundreds of important TP applications. The growth of such applications accelerates as costs come down and ease of use increases. These applications can be classified as follows:

- Conversational
- Inquiry/response
- Data entry
- Batch
- Application to application
- Sensor base

Each application has different characteristics and different requirements on an architecture for distributed systems.

Conversational Applications.

These applications are characterized by a series of rather short interelated messages. The amount of traffic is about the same in both directions. (It is said, therefore, to be "balanced.") A wait for a reply to each message is a normal mode.

Conversational program development, using a time-sharing subsystem, is one example of this class. Some of the simple inquiry system also can approach the conversational mode. Airline reservations are typically conversational within a reservation transaction. The size of messages is usually less than 100 characters (or 100 bytes of 8 bits each). Since the messages are small and the processing to generate each reply is relatively small, the overhead in both the communications and data-processing systems must be kept relatively low; otherwise, the achievable transaction rates may be severely limited by the overhead. In conversational applications, overall system response time is a prime requirement; response times less than three to four seconds are usually needed to keep the level of human efficiency high. Because the operator is continuously involved in the conversation, he or she can also be depended on to handle some recovery situations.

R.J. Cypser, *Communications Architecture for Distributed Systems*, © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

Inquiry/Response Applications.

These applications are typified by longer replies to short queries and by a greater independence of one request from another. The reply is typically four or five times longer than the inquiry, and the ratio can be much higher. (This traffic, then, is said to be "unbalanced.") The inquiry can be fairly complex, as the following examples indicate:
1. "Give me a list of all the subassemblies that use part number 5179,"
2. "Give me the projected reorder dates for the parts bly 224," or
3. "Give me the names of all the PhDs in chemistry who work at our plants in New York and Philadelphia."

The number of accesses to data storage and the sophistication of data-management services are usually greater in inquiry applications. The communications overhead and the time spent in communications are, therefore, a smaller percentage of the total round-trip overhead and processing time. Response time in the range of two to twenty seconds is still very important, but response times that are longer than three seconds are more acceptable than in conversational applications. As in conversational applications, consistency of response is important to user satisfaction.

Data Entry.

Data-entry and data-collection activities are often characterized by relatively long input messages and very short replies. The amount of host processing per message is minimal but may include validity checks, editing, and formatting.

Batch Applications.

These applications are typified by remote job entry and the distribution of voluminous output to one or more remote locations. The input may be limited to parameters to be used by the program that was previously stored in the central site: the program itself may be the input that then processes data kept at the central site; or both program and data may have to be the input. Both input and output may be voluminous. Turnaround time, in the order of minutes or even hours, is the response criterion for batch applications. In some cases, deadline-scheduling is used instead, stating that a particular job must be completed before, say, 8:00 A.M. tomorrow.

Batch applications enter the system in groups and are processed in a sequence that maximizes system throughput. They are expected to operate without direct operator involvement with a particular job. Therefore, a high degree of automatic recovery from

Teleprocessing—The Modern Marriage of Computers and Communications

communication error, without manual assistance, is required. Message lengths are, of course, likely to be very long and variable in both directions.

Application-to-Application.

These applications arise when there is considerable computational power at both ends of the communication line and the communications are directed by application programs or system service programs at both ends. This is becoming the case more and more as the cost of data-processing technology decreases and as it becomes more feasible to locate more function at each remote node.

The processor-to-processor traffic can be either balanced or unbalanced. It can have some of the characteristics of conversational, inquiry or batch applications, depending on the division of function between the two processors. There is a high requirement for automatic error detection and automatic recovery from errors, with an absolute minimum of operator intervention.

Sensor-base Applications.

The primary requirement in these applications is fast response time, measured from the instant that the sensing device requests the attention of the processor until the moment (after the processor has completed the processing of the transaction) when the total reply is delivered to the sensor. The message may be only a few bytes containing data from a single sensor, or hundreds of bytes containing data from many sensors. Because access to mechanical storage, such as tapes or disks, is time-consuming, all necessary data is kept in high-speed storage.

The time to interrupt the processor and begin processing the transaction must be very short and higher-speed transmission lines must be used, so that the total response time is measured in milliseconds.

CONFIGURATION BREADTH

Processing power may be either entirely in one location or distributed among many different locations. Most installations begin with one central processor and a few simple terminals and then expand to more distributed processing and more functional terminals. A manufacturing complex may, for example, grow to have twenty cluster controllers (that is, units for the control of a group of terminals), each with five to fifteen display units, keyboards, and/or printers. A single large reservation system spanning the nation may have several thousand terminals working through 50 or so concentrators (units that merge data from slow speed lines onto a higher speed line). The teleprocessing network may be restricted to a single building, or it may span a city, a state, or the nation. Groups of data-processing centers are also linked by communications systems; these networks may cover any area, even extending from country to country.

Data-processing centers may be dedicated to only one application, while in other cases, economy of scale leads to DP centers that serve multiple applications. Teleprocessing has provided the means for remote users to access both types of center. When the center contains integrated databases capable of serving a range of applications, the added value of the file makes it worthwhile for more users to access the center.

As applications multiply in number and size, databases (both dedicated and integrated) tend to develop at multiple data-processing centers. Since data is the raw material of data processing, a need often develops for access (occasional or periodic) to more than one database. This motivates users to demand interconnected networks of machines and applications. Thus, three factors, database development, remote access, and sometimes economy of scale, combine with lower cost technologies to advance the trend toward networking of data-processing facilities.

Processors may be interconnected in a local network (for example, via computer channel to computer channel) in a remote network via common-carrier or PTT facilities, or combinations of remote and local networks. Connecting networks may be tree structures (with a single path from the root to other elements), or they may be mesh structures (with alternate paths). The architecture for distributed systems must be able to accommodate a wide variety of such configurations and to facilitate changes in the configuration as the system matures.

MESSAGES AND TRANSACTIONS

A message is a single transmission of a user's data between two points. Some illustrative message sizes in today's systems are shown in Figure 1. If the input is an inquiry from a keyboard, it typically will be short, in the range of 20 to 50 characters. The response to the inquiry will often be in the form of a display, which can be quickly presented even though it amounts to anywhere from 100 to 1000 characters. The unit of work in batch applications may be tens of thousands of characters, but these are usually broken into smaller components (of perhaps 256 to 1000 characters) for transmission as a series of messages.

R.J. Cypser, *Communications Architecture for Distributed Systems*, © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

Teleprocessing—The Modern Marriage of Computers and Communications

Industry: Application:	Banking Data Collection	Airline Reservation	Broad-casting Inquiries	Public Utility Remote Job Entry
Characters/ message				
Input	200	20	20	100—1000
Output	60	100	400	
Lines Printed/job				100—2000

Figure 1. Illustrative message sizes and printout output

The message rates of different terminals range from one-tenth of a message per hour to hundreds of messages per hour. A message rate at the central processor may range from one-tenth of a message per second to hundreds of messages per second, depending on the type of messages. For example, the complexity of the message, in terms of the amount of computer processing involved, can vary from 100 to 1,000,000 units per message.

A typical airlines reservation message, for example, involves about 15,000 instruction executions and ten database accesses. The more complex messages may contain statements of a higher-level language (for example, COBOL or APL); they may contain macro instructions known only to the destination; or they may contain a request for a complex search of a database. Cases are known that involve over a million instructions and up to thirty database accesses per message. The amount of input and output data transmitted per message can vary by factors of a thousand or more.

Message size and the amount of message processing tend to vary inversely with message rates. On the other hand, for a given message rate, message size and processing tend to increase with time as the complexity and breadth of computer-aided operations increase. The growth in message rate and complexity results in a steady pressure for improved performance of both the data-communications and the data-processing systems.

A transaction involves a series of messages, in one or both directions, which together achieve a unit of work. The transaction is a characteristic of the application. For example, in the airline industry, a transaction might span the series of messages involved in reserving a seat in a flight reservation system. The transaction duration in this case is affected by the time for interaction between the passenger and the agent, as well as the time to process the transaction at the computer site.

R.J. Cypser, *Communications Architecture for Distributed Systems,* © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

In the banking industry, a transaction may consist of the entry of a customer's deposit and the updating of the customer's balance. In the retail industry, a transaction might be the process of recording item sales, or processing refunds, or verifying checks before accepting them as tender. In remote batch entry, the transaction may be a job step or an entire job, while in application program development, it may be the entry of a long series of programming statements and their trial execution.

In each case, the transaction time includes data-processing time and often human-interaction time. The time for actual data transmission may be a small or large fraction of the transaction time, depending on transmission speeds and the amounts of data-processing and/or human-interaction time.

The frequency and duration of transaction can vary over wide ranges. A rough approximation of this range is given in Figure 2 as a function of application class. A relative measure of the activity per terminal is given by the product of the two coordinates; in Figure 2, this is loosely expressed in erlangs, which, strictly speaking, is the ratio of mean service time to mean time between customer arrivals.

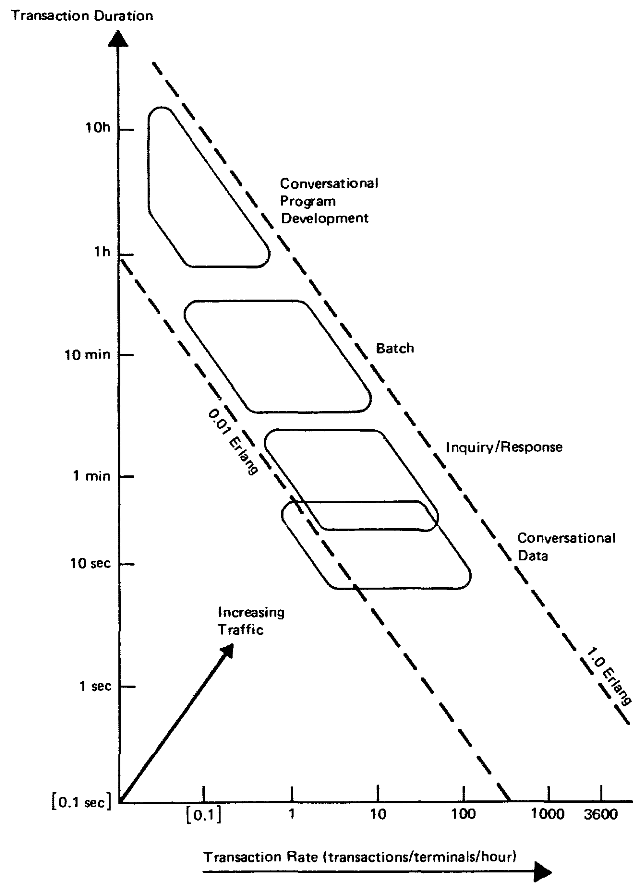


Figure 2. Spectrum of transaction characteristics by application class

Teleprocessing—The Modern Marriage of Computers and Communications

One or more transactions could be the subject of a distinct call on the switched network, with a physical connection established between two parties only for the duration of the transactions. On the other hand, the physical connection may be made on a long-term basis (for example, with a nonswitched line); then a logical connection may be made for two parties for only the duration of one or more transactions. In still other cases, if the disconnect/connect time is short enough, it is possible to have a physical connection only for each message within a transaction, disconnecting for the pause between messages.

ILLUSTRATIVE DISTRIBUTED SYSTEMS

The features of data-processing/communications systems thus span a wide spectrum that almost defies illustration. Nevertheless, we'll try to illustrate the fact that systems of today are distributed in two ways: first in a hierarchical fashion and then in a peer fashion. The degree of centralization and distribution will first be illustrated by examining typical systems in three industries:

1. Airline reservation systems using a centralized data base

2. A banking system, where the data base is centralized but some of the message processing is distributed

3. A retail system, where more of the processing is distributed

Then we will consider how, in addition, each of these systems might involve the use of peer processors and peer data bases.

Centralized Reservations.

One of the pioneering developments in on-line, interactive, data-base-oriented teleprocessing systems has been for airline systems. Out of the development of PARS (the Programmed Airline Reservation System) came a generalized Airlines Control Program (ACP) that was optimized for short standard messages, fixed formatted file records, and high transaction rates. Today, a typical ACP system might consist of 2000 to 5000 terminals. Some ACP systems also exist with a few hundred terminals, and 10,000-terminal systems have been envisaged for the near future.

Often, these networks span a large area, typically nationwide, connecting agents in the major cities of a country to a centralized database.Thus, any agent can sell, change, or cancel a reservation for any flight segment in the system and know that all information is accurate and current to that instant.

In airline reservation systems, long-distance communication lines are shared among many agents through the use of concentrators at key locations. Traffic to and from a number of agents is multiplexed by the concentrator onto a single long-distance line (see Figure 3). A number of these concentrators may all share a single 2400-bit/sec (bps) communication line to the central site; polling manages this sharing by allowing each concentrator, in turn, to use that line. The agent work stations may be locally attached to the concentrator or remotely attached via communication lines operating at 2400, 1200, or 148.8 bps. (A still higher level of sharing may be done by the telephone company, in which many such individual lines share a broadband transmission facility for the intercity and long-distance traffic. This sharing, however, is completely transparent to the subscriber).

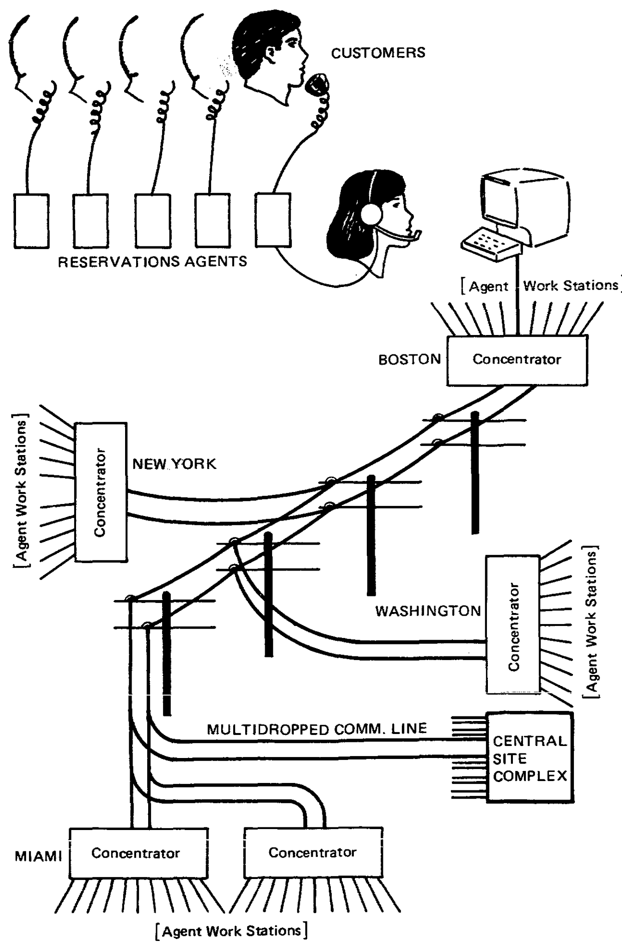


Figure 3. Centralized airline reservation system [*From "A Case Study: Airlines Reservation Systmes" by J.R. Knight, Proc. IEEE 60 (November 1972), pp. 1423-1431. Reprinted by permission.*]

R.J. Cypser, *Communications Architecture for Distributed Systems*, © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

Teleprocessing—The Modern Marriage of Computers and Communications

As an example, a system using an IBM S/360 Model 195 was designed to process 180 typical reservation messages each second, with the central processing unit operating at 85 percent utilization. The average response time was designed to be within two seconds, the response time at the 90th percentile to be within four seconds, and the average processing time per message to be less than 4.7 milliseconds.

Centralized Data and Distributed Processing.

Many financial institutions are using distributed programmable units to handle transactions locally. We will describe a hypothetical but representative system in which large numbers of work stations, spread over large areas, operate on-line in this type of distributed-function network.

Each work station typically is composed of the following terminal facilities:

1. A programmable keyboard
2. A reader of prerecorded information in magnetic stripes
3. An alphanumeric character display (for example, a 240-character gas display panel)
4. A receipt and journal printer (for example, a 30-character/second, 80-column printer)

Alternatively, a work station might be a higher-speed administrative line printer. A group of such work stations is managed by a programmable cluster controller, as shown in Figure 4. The work stations are connected to the programmable controller via private, on-premises loops at 1200, 2400, or 4800 bps, or via common carrier at—say, 1200 bps.

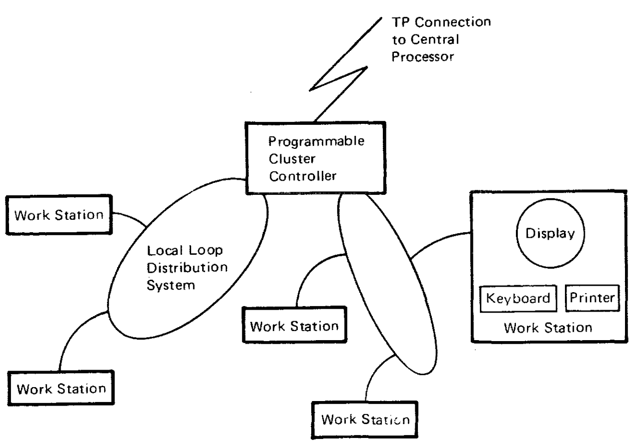


Figure 4. Multiple work stations attached to loops from a programmable cluster controller

---

R.J. Cypser, *Communications Architecture for Distributed Systems,* © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

The programmed cluster controllers execute application-oriented programs and store data pertinent to local operations. They can be programmed to act as an "electronic journal," maintaining local totals, logging transaction performed on attached terminals, and providing a detailed audit trail. They can also be programmed to capture transactions during off-line operation for later transmission to a central computer site. In one type of controller, a removable random-access diskette can store up to 560K bytes of data. In addition, certain members of that controller family have nonremovable disk storage of up to 9.3 million bytes. Many transactions, however, may also draw data in real time from the central database to which each programmable controller is connected. The programmable controllers are, in effect, local coordinators and preliminary processors for the operations at the multiple work stations.

Each of the programmable cluster controllers in a typical installation will be connected to a central host site via lines of 1200-4800 bps. In some applications, the central site contains the central data base that is updated in real time by certain transactions entered at each work station. Every transaction across the entire network thus can draw on information that is accurate up to that instant, regardless of the number and/or location of the transactions. An illustrative duplexed configuration for a central site is given in Figure 5, showing dual processors, shared disk storage, shared tape files, and shared communications controllers. Although an I/O device may be shared, only one processor, with the required amount of equipment dedicated to it, would be on-line at any given time.

Another part of the financial network may involve high-speed data collection during a brief period each day. Data is collected from the batch-processing centers at the dispersed locations to the above-mentioned central site. The batch-processing sites could be connected, via high-speed lines, to the central site. High-speed tapes, operating in the range of 470K-1250K bytes per second, would receive the batch input from these high-speed lines. The batch input from the dispersed locations provides the daily confirmation of the central data base, which then is incremented in real time during the day, as described previously.

Semiautonomous Distributed Processing.

Examples of distributed processing, where still greater autonomy is exercised at each processor, are found in the retail industry. Here programmable controllers operate autonomously, for the common types of transaction, in each store. With over a hundred thousand bytes of high-speed storage in the cluster controller, multiple applications can be run at

Teleprocessing—The Modern Marriage of Computers and Communications

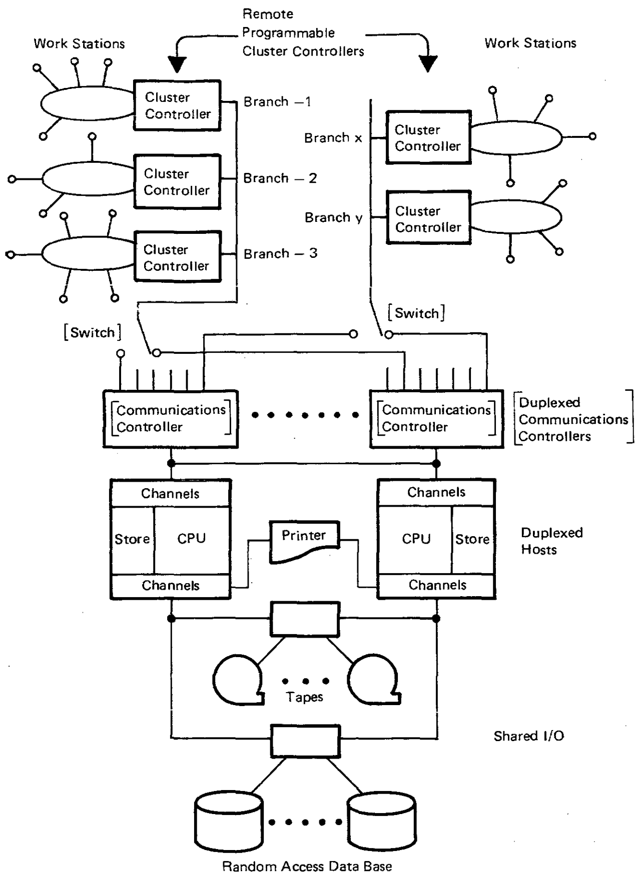


Figure 5. Illustrative duplexed central processing site and work stations on remote cluster controllers

the store level. For other types, an interaction with a central site is used. Let us examine one of these "in-store" systems.

Sales personnel use a "point-of-sale" terminal for sales transactions, credit authorization, and some inquiry functions. Data entry may be through a magnetic or optical wand, whose passage over a label reads the identity of the item, or through a numeric and function-key keyboard. Instructions to the operator and data being entered are displayed; data provided in response to an inquiry may be printed.

Cash transactions are handled solely by the interactions of the terminal and a programmed cluster controller located in each store. In this role, the programmed controllers operate autonomously. Credit and check-cashing authorization, on the other hand, involve a check against a master file at a central computer location. Also, once a day, another central computer application draws data from all of its connected controllers so as to establish register balances and conduct an overall sales audit.

R.J. Cypser, *Communications Architecture for Distributed Systems*, © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

Another set of applications concerns the flow of inventory, and relies on a few separate display terminals per store. Order entry is the creation of purchase orders and the input to the purchase-order data base. The receiving application controls the movement of merchandise received and checked. Accounts payable includes the entry of invoice data into the data base, the calculation of cost and retail sales dollars, and information verification. These types of application are executed partly in the controller and partly in the central processor. The interaction is from the display terminal via the same controller that handles the sales transactions to the central computer.

Let us take as an example a chain of stores located throughout several states. In this installation, a group of 20 department stores is being brought on-line, with one programmed controller in each store and a central computer to coordinate them all. In at least one case, several stores can share a single programmed controller.

In our example, terminals are connected to the programmed store controller via a 2400- or 9600-bps transmission loop. The controllers, in turn, are each connected to the central computer by a separate 4800-bps telephone line. Each programmed controller manages from 60 to 120 point-of-sale terminals plus a display terminal and a printer. These terminals may handle from 20 to 30 transactions per hour, and the programmed controller in a store may handle 2000 to 3000 transactions per hour during a peak sales period. Response times at a point-of-sale terminal probably average less than a second, and less than ten percent of the responses should take more than, say, 1.5 seconds.

Each credit authorization requires only one or possibly two messages to the central computer. However, transactions of the inventory-flow applications may involve four or five messages to the central computer per transaction. The central computer, then, must be capable of handling in the order of eight to ten messages per second during peak sales periods, even though all cash transactions are handled locally, using the in-store programmed cluster controller.

When the day's transactions are batched from all the store controllers to the central computer, the transmission must take place in a short time, say, 0.5-1.5 hours. The records for tens of thousands of transactions must be transmitted in this mode, and the central computer must be capable of handling an equivalent of 10 to 20 messages per second during this time.

Teleprocessing—The Modern Marriage of Computers and Communications

Multiple Peers

The preceding examples illustrate a hierarchical distribution of functions among three levels: the intelligent terminals, the programmed cluster controllers, and the central processing unit. Both data bases and processing capabilities can be so distributed.

Given this hierarchy of distribution, one can, in addition, have multiple servers that operate as peers. To illustrate, any of the central processing units in the above examples might be replaced by multiple CPUs and multiple databases. These might be at different locations. Different types of operation with such peers can be identified as follows.

1. Transaction routing to peer data bases. In some cases, the database is partitioned by geography or function, and separate databases are managed by different processors. These are peers of one another that can be coupled together. An illustrative configuration for systems with peer coupling is shown in Figure 6. It may be desirable that terminals at any location be able to access any database and that the terminal user be unaware of the database partitioning.

In such systems, if a request arrives at any CPU it should be rerouted automatically to the site where the appropriate database is located. With transaction routing, the routing to the correct database is based on a transaction code in the user's request. Similarly, a request from the database to any terminal can be routed to that terminal, via an intermediate host if necessary, using the terminal name in the request. In this example, the routing is achieved in an application-like program by examining the contents of the request that is provided by the user of the network.

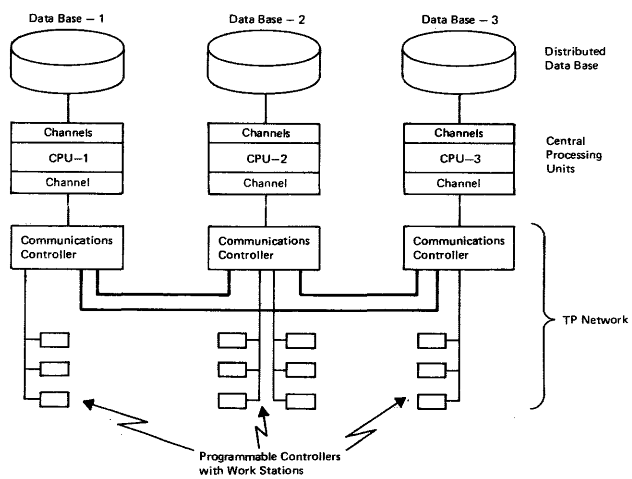


Figure 6. Configuration with peer-coupled distributed databases

---

R.J. Cypser, *Communications Architecture for Distributed Systems,* © 1978, Addison-Wesley, Reading, Massachusetts, Chp. 1, pp 3-19. Reprinted with permission.

2. Job routing to peer processing units. This is another form of transaction routing in which the work scope is a job (that is, an application program). As before, special fields within the user's request that accompanies the job can be used to achieve the routings. These fields can be interpreted by a so-called Job Entry Subsystem (JES), which functions as a pseudo-application program. JES performs the routing and coordinates the scheduling of jobs at multiple CPUs.

In one implementation, for example, the submitter of a job may specify the host upon which a job is to be executed and also the destination of the output resulting from job execution. A job may be entered into the network from any job entry station that is local to one of the hosts or from a remote terminal. A job may also be entered into the network via any of the internal job queues within any of the hosts. Jobs may be transmitted directly from an originating host to an execution host, without incurring store-and-forward overhead at intermediary hosts. When the job has been received at the execution host, it is queued to await execution. During execution of the job, output data sets are queued for transmission to the destination specified by the submitter of the job.

3. Transaction-routing network service. In the two cases cited above, the routing of the transaction (or job) is performed by a subsystem that operates as an application program external to the network. An alternative is to build the system so that the routing function is a part of the network services, even though examination of the content of the request is involved.

4. Connection to alternate peers. Quite a different approach is to build into the network architecture an ability to achieve logical connections to any program that may be located in any CPU without examination of the content of each of the user's requests. The connection (or session) usually pertains to the exchange of a series of bidirectional messages, which may proceed for a short or an extended length of time. Such a connection involves separate set-up messages to establish an initial connection. At that point, the user of the network specifies the name of the desired destination, for example, to which program subsequent messages will be sent. The subsequent dialogue employs addressing facilities that use headers supplied by transmission services of the network (rather than fields within the user's request).

More than one of these four types of operation may coexist in the same system. □

The Special Requirements of Data Communications Traffic

Problem:

Communications among people are usually through a voice medium, like a telephone, and occasionally through a visual medium, like television (mostly one-way), augmented by voice. Communications among computers and computer-related equipment are totally different. They are based on a form of pulse-type encoding generally called data, or digital, communications. The basic incompatibility between these two forms of communications has consistently been the central problem in computer communications systems because the most convenient and universal transmission facility, the telephone network, is designed for voice communications, not data communications. Many of the techniques and equipment discussed in later reports have been devised and built expressly to modify data communications traffic to accommodate the limitations of the telephone transmission facilities. Newer transmission facilities such as coaxial cables, microwave links, and particularly satellites and optical fibers are infinitely more compatible with the requirements of data communications, so much so that even voice traffic is gradually being digitized to take advantage of the considerable traffic control and routing efficiencies that can be realized in an all-digital medium.

This report explains the special requirements of digital communications traffic and is offered to set the stage for the reports that follow in this section.

Solution:

The use of communications with computers began in earnest in the late 1950's, about a decade after J. Presper Eckert and John W. Mauchly built the first electronic digital computer at the University of Pennsylvania in 1946. The initial use of computer communication was with teletypewriter-type connections on a point-to-point basis at speeds of about 45 bits/second. The introduction of modems, allowing digital information to be transmitted over the voice telephone system, allowed data transmission rates to be increased significantly and made the video terminal feasible. Multipoint lines were soon established over leased telephone circuits where a single line is shared by several terminals, and polling concepts were introduced. In polling, the computer interrogates each terminal in sequence to determine if there is information to be sent. The major advantage of the use of the voice or analog telephone system is the universality of its facilities, providing existing access essentially to every location in the more advanced countries at rates of up to 9600 bits per second.

Computer Technology Impact on Management by George A. Champine, Sperry Univac, Roseville, Minn. Chapter 5, pp 79-99.

The Special Requirements of Data Communications Traffic

The computer in the early systems provided all of the intelligence, and communications hardware was limited to line termination (conversion from telephone signal levels to computer logic) and multiplexing to handle many lines.

Growth of commercial communications systems was initially slow except for the airlines industry, which moved aggressively in developing real-time passenger reservation systems and in the process moved communications technology out of the laboratory and into products. These systems had a centralized database, with a host computer communicating with a few hundred to a few thousand terminals over leased telephone lines. The airlines passenger reservation systems were among the first to use failsoft techniques, not only in communications, but also in the mainframe via multiprocessors or standby systems with shared nonvolatile mass storage. The systems developed by the airlines were to remain the industry standard approach to both communications and failsoft systems for approximately a decade, with minor technical improvements such as terminal multiplexers and remote concentrators. The usage of communications has grown steadily until now about 80 per cent of new medium and large scale systems have data communications.

With the growth of communications usage came problems. Response times were long because of polling on multipoint, half duplex communications lines of limited bandwidth. Costs were high because dedicated lines were used even for small communications loads. Reliability was not good because a single failure in a multipoint configuration could cause the entire complex to fail. More importantly, the specific method of connecting terminals to a computer could preclude the use of the terminal on other computers or even other applications.

Beginning in about 1965, there has been increasing need for the establishment of networks meeting users needs better than public switched telephone systems, to provide improved response time, error rate, security, and cost. Large organizations were already doing this by establishing private networks, obtaining only transmission facilities from common carriers and providing all switching themselves through privately-owned equipment. Although this met the needs of the large user, each was tailor made at great expense, often underutilized, vulnerable to circuit failure, and often incompatible with other networks. This has given rise to a number of standards activities and value added network systems to overcome these problems.

A variety of differing user requirements exist for digital communications. These differ in transmit time, delay, and switching complexity as shown in Table 1.

Communications Description	Transmit Time	Delay	Switching
Batch data transfer	Tens of minutes	Not important	Simple
Time sharing	Seconds	Critical	Medium
Message switching	Minutes	Medium	Complex
Transactions	Milliseconds	Critical	Complex

Table 1. Data communications characteristics

In response to these requirements, the communications common carriers have developed the following classes of service:

- Leased line to a fixed location-batch data transfer or other long connect times
- Telephone network—where global access is required or usage is low
- Circuit switched—batch data transmission to many locations
- Packet switched—high transfer rate but low usage

The selection of the proper class of service for a given mix of applications is a complex function of call duration, distance, and data rates. In general, connect times on the order of seconds are best matched to packet switching (low speed) or leased line (high speed) whereas connect times on the order of minutes are better suited to telephone network (low speed) or circuit-switched (high speed).

The 1970's saw a natural evolution of communications technology away from the circuit switched or dedicated circuit to message switching and then to packet switching. The analog telephone system is well matched to voice communication, which is characterized by long connect time (minutes), low delay times (milliseconds), and low data rates (kiloHertz). Circuit switched systems, where a complete physical link is established and dedicated to a single user, met this need very well. Circuit switched systems also formed the basis for early data communications systems.

Because of the long time to establish a circuit and because unused bandwidth cannot be shared with other subscriber pairs, circuit switched systems are not economical for computer traffic of short duration. Computers need short connect time, high data rates, and in noninteractive applications, delays between transmit and receive can be minutes or even hours. Exploitation of the insensitivity to delay time in batch transmission led to message switching systems, in which complete messages were sent to store-and-forward nodes in sequence to obtain the desired routing. Network delay was determined by the number of nodes traversed and was typically

The Special Requirements of Data Communications Traffic

minutes to hours. Network utilization was much higher because of the sharing of circuits by many messages, and the cost per message was considerably reduced.

For interactive applications, the cost of message switching was attractive, but the delay time was intolerable. To improve the delay time while retaining the attractive cost per message, packet switching was introduced. Here, the message is broken up into a number of pieces (or packets) of fixed length. Each packet may be relayed through a node as it arrives without waiting for the complete message or being placed in mass storage, thereby obtaining much lower delay times, often on the order of one second.

Network systems also shifted the focus of communications away from the host computer. The same technology improvements that were helping host computers were also helping communications line equipment such as front-end processors, nodal processors, line concentrators, and modems; substantial reductions in price and improvements in functionality and performance have been obtained in these areas.

A computer communications system is made up of a number of elements. A host, in this context, is a computer whose function does not involve switching of data. A node is also a computer, but one that functions only as a data switch. In some designs, a single computer acts both as a host and as a node. Terminals are interfaces between humans and the host or communications system. Transmission links transfer data among the hosts, nodes, and terminals. A path is a series of end-to-end links that establishes a data route across part of the communications system.

Communications will be provided in the future, as now, by some combination of private (dedicated) systems, public value-added network systems, and common carriers. Independent of the relative dominance of these suppliers of communications, they will all use the same technology and provide roughly the same services to the end users, with the principal differences being cost and control.

The government regulation of communications has been at the forefront of discussion for some time and will not be addressed here. Instead, it will be assumed that in the long run, government regulation will not prevent the opportunities in the communications area from being exploited to obtain inherent opportunities in efficiency of operation and expansion of services.

The following sections describe the communications technology that may be expected in the next several years. The various transmission technologies are first discussed, followed by protocols and protocol standards. Reliability and security are then discussed as a foundation to a review of the kinds of communications systems expected in the middle 1980's. The Canadian DATAPAC public packet switching system, now in operation, is presented as a specific example of the kind of services and costs to be expected in the next few years.

TRANSMISSION SERVICES

Transmission services are provided today by four technologies; these are wire, coaxial cable, microwave, and satellite. In addition, a new technology, fiber optics, will soon be a significant factor.

Wire is used for communications rates up to 9600 bits per second; microwave and satellite links provide a variety of high and low speed channels. Coaxial cable has been used widely for cable television and incidently for local loops to interface to wideband microwave and satellite channels. Microwave systems are used not only by the common carriers for long distance circuits but are also used by large individual users for wideband communications over a limited area such as a city. These technologies are reasonably mature, although applications continue to improve, with the exception of wire, which does not show any potential for cost or performance improvement beyond 9600 bits per second. The new technology now developing is fiber optics.

Fiber Optics

Until about 1970, optical fibers then available had a great deal of signal attenuation, limiting their use in communications to a few tens or hundreds of feet. However, the bandwidth potential was almost unlimited, being in the gigabit/second range. Technological breakthroughs in the early 1970's have reduced the signal loss to about 5 dB per kilometer, lower than that of coaxial cable. Although fiber optic cable is currently substantially more expensive than coaxial cable, it should be less expensive than coaxial in the early 1980's with far greater bandwidth. Problems that still exist, in addition to cost, are the development of connectors, splicing methods, and tapes for multi-drop lines. However, promising approaches exist for the solution to each of these problems.

The fiber optic technology is rapidly maturing and can be expected to find widespread use in computer I/O systems and short haul telephone service. Several systems are now in field trials transmitting at 50 megabits/second, and experimental systems are operational at one gigabit per second. In addition to the cost and performance advantages, the fiber optic technology also eliminates computer electrical grounding problems, electromagnetic interference, is

The Special Requirements of Data Communications Traffic

about one-tenth the size of an equivalent copper cable, and can be bent at short radius. A 20 km (12.4 miles) link is now operational between Tokyo and Taito-ku in Japan carrying 3500 voice circuits over a cable of 24 glass fibers with a total diameter of 1.7 cm (0.7 inch.) Other systems are operational in Atlanta, Chicago, and Long Beach at data rates up to 44.7 megabits per second.

Coaxial Cable

Coaxial cable such as used for cable television is a relatively mature technology that offers considerable promise in providing a variety of digital and video information services to the home consumer. To date, significant technical/economic problems have inhibited any significant progress in this area which are tamperproof relative to two-way communications. The potential for this media is very large, however, and recently there has been an increase in development activity in this area, which should lead to products that can exploit the capability of this wideband channel capacity.

Satellite

Communications satellites have progressed rapidly in capability since their introduction in the late 1950's. Now, satellite communications cost is competitive with land line common carriers for distances over 800 km (500 miles). In future years, the break-even point between satellite communications and conventional microwave land circuits for large volume use is expected to be reduced to the 300-500 km (200-300 mile) range.

The early communications satellites used sub-synchronous orbits at altitudes of about 9600 km (6000 miles). At that time it was realized that the geo-synchronous orbit of 35,888 km (22,300 miles) would simplify the earth station requirements by allowing the use of fixed antennae rather than tracking antennae. Although there was concern about the half-second round trip delay and the lower signal strengths available because of the higher altitude, the geo-synchronous technique now dominates due to higher power transmitters, more sensitive receivers, and sophisticated echo-suppression techniques making possible full-duplex operation.

The principal advantages of satellite communications are in long haul service with wide bandwidth, where satellites provide lower cost and higher reliability. Raw error rates of one bit in 10^8 are common for satellite links compared with one bit in 10^5 for high quality terrestrial links. These raw error rates are greatly reduced in transmission systems by the use of error correction codes and retransmission techniques. A typical contemporary satellite can accommodate 12,000 high quality voice channels and can cover the entire 50 United States.

A significant recurring cost in the use of satellite communications is the land circuit connecting the ground station to the various subscribers. For large users, this land circuit may be eliminated by locating the ground station on the premises of the subscriber. Ground stations are still rather expensive—in the neighborhood of $500,000. However, the amortized cost is not large compared to what large corporations spend each year for data communications; for example, the Sperry Rand Corporation spends about $3 million annually for data communications alone, not counting voice, Telex, or facsimile. The cost of ground stations is expected to be reduced at a rate of about ten per cent per year for the next several years.

The satellite communications system of Sperry Univac is a typical example of a corporate satellite communications system of the 1980's. This system, shown in Figure 1, provides two 56 kilobit/second full-duplex channels between Sperry Univac headquarters in Philadelphia, Pennsylvania, and the location of the largest Sperry Univac operation in St. Paul, Minnesota. One channel is dedicated to computer applications, including batch and interactive data transmission. The other channel is allocated to a variety of communications functions including high speed facsimile, digitized telephone, and slow scan television to reduce the need for travel by using teleconferencing.

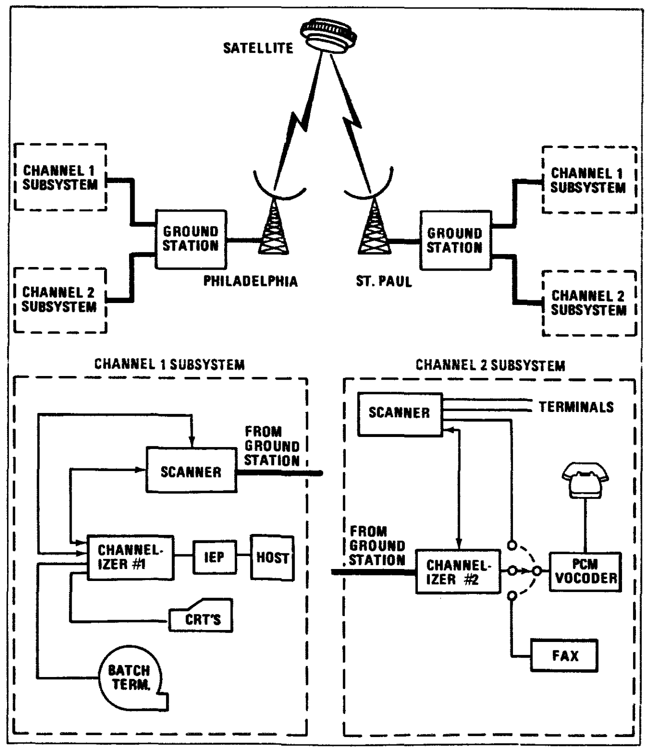


Figure 1. Satellite communications system

The Special Requirements of Data Communications Traffic

The ground stations, supplied by the American Satellite Corporation division of Fairchild Industries Incorporated, are located on the premises in each location and use a fixed 4.5 meter parabolic antenna operating in the 4-6 gigaHertz region. The satellite used is the "WESTAR" satellite provided by Western Union, located in geo-synchronous orbit at 99.5 degrees west longitude.

PROTOCOLS AND STANDARDS

For a long time, standards for interfacing end user terminals or host systems to data communications systems were almost nonexistent, so in this absence each computer manufacturer developed its own interface standards or protocols. Recently, progress has been made towards national and international standard protocols, driven by the very great demand for large scale communications systems often encircling the world.

These protocols exist at four levels, or layers, forming the basis of current network access standards:

- Physical circuit protocol (level 1)—This defines basic electrical voltage and timing standards as well as connector definition. Typical of standards in this area is X.21, a digital equivalent of the Electronic Industries Association RS449 standard, which defines data, clocking, and control leads including the size and configuration of pins in the connector. This is also under consideration as an American National Standards Institute (ANSI).
- Link Control Protocol (level 2)—This protocol, also called a "frame level" protocol, defines how a terminal interfaces to the network. This protocol provides a number of benefits in multiple access and error control. At this level, the International Standards Organization (ISO) has proposed a standard High Level Data Link Control (HDLC). HDLC is a bit-oriented protocol that overcomes many of the deficiencies of the earlier byte-oriented protocols by providing full-duplex operation and transparency to control characters. A variation on HDLC has been proposed by ANSI called Advanced Data Communication Control Procedure (ADCCP). Data link protocols like HDLC have a variety of uses in simple networks for transmitting data between terminals and computers.
- Packet Level Protocol (level 3)—This protocol defines how messages are identified to the network for routing and control. The most widely accepted standard in this area is the proposed X.25 standard, diagrammed in Figure 2. This standard consists of three levels, with X.21 at the first level; HDLC is defined as the second level to specify formats for frames of information containing source/destination address, message number, and various sequencing control bits in addition to the data. At the third level, more complex blocks of information are defined that contain additional addressing and sequencing information to allow interfacing to packet switching networks.
- Applications Level Protocol (level 4)—This protocol specifies how terminals and applications programs talk to each other. There has been little standards activity in this area.

The X.25 standard as it now exists is still somewhat controversial at the third level because of its complexity and certain problems in interfacing between networks. However, in the absence of any other standard, a number of organizations are proceeding on the assumption that X.25, or something very similar to it, will become the international standard.

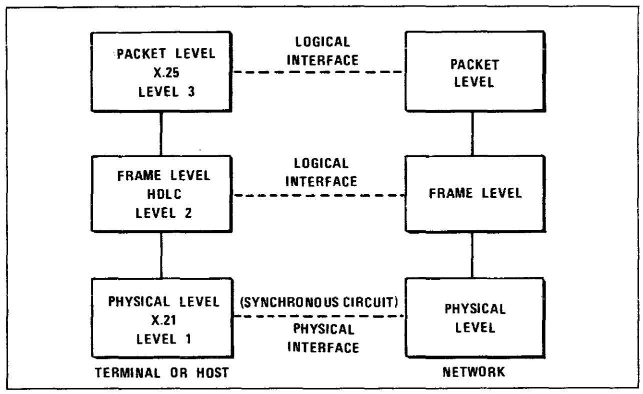


Figure 2. Levels of communications protocols

RELIABILITY

Communications has a unique set of reliability problems. In addition to the noise and parameter drift problems that are in common with other electronic equipment, the unique problems exist of 1) obtaining truly independent redundant facilities and 2) variation of four orders of magnitude in the reliability of some facilities relative to others.

A fundamental assumption in the redundant approach is that failures are independent; that is, a failure of one component will not cause the failure of another component. In the telephone system, these same redundancy techniques are attractive and widespread. However, a hidden pitfall is that many failures are dependent; that is, a failure in the primary component will also cause a failure in a redundant component.

The Special Requirements of Data Communications Traffic

One large U.S. airline went to extraordinary effort and expense to develop a fully redundant telephone network to support its very important passenger reservation system. What they did not realize was that both sets of redundant lines shared the same telephone poles the last mile before entering the computer center. One day a truck hit one of the (nonredundant) poles, and the entire system was off the air for several hours.

More commonly (and less visibly), "independent" telephone lines are often part of the same multipair cable. In particular, there is a very strong likelihood that all lines from a particular location (for example, a leased line and its leased or dial backup line) will share the same multipair cable for the full distance from the user location to the local switching office unless previous growth of traffic has caused multiple cables to be installed over this route. Aside from local loops, other sources of failure dependencies can include common equipment in switching offices (even for leased lines, which pass through switching offices without switching) and short haul or long haul carrier systems. In a recent case study of 35 "independent" lines for reliability, the multiple failure rate was 2.5 times as large as it should have been if the lines were truly independent.

These same 35 lines showed a reliability ratio of the best to the worst of 174 to 1, thus showing the very large variation in reliability of telephone equipment. Other studies have shown three or even four orders of magnitude ratios of reliability. The reason for this is that telephone equipment is designed for a 20-30 year service life at a time when there is substantial improvement in technology and large growth in the telephone physical plant. This has resulted in a wide variety of equipments being hooked together to form the installed plant. Also, the telephone system is subjected to a wide variety of operating environments and is among the most complex equipments ever designed and operated by mankind.

The net result of both problems is that designing a communications system by using averages for reliability can be very misleading, and availabilities at the 90 or 95 per cent confidence level should be used. Both problems also indicate that it is foolhardy to design systems which require ultra-reliable, real-time communications. Some form of local storage or other backup should always be used to be able to survive a communications outage.

SECURITY

The advent of data communications systems and the growing user dependence on these communications systems have brought with them the need for data security. The application of data security ranges from prevention of theft and fraud (principally in financial data) to prevention of industrial espionage. Essentially all organizations have data that must be protected from unauthorized access, both from people within the organization and from people outside the organization.

The methods of providing secure communications between hosts or between hosts and terminals involve encryption, with essentially all present activity focused on the proposed Data Encryption Standard.

The primary objective of data encryption is to prevent the disclosure of information to unauthorized individuals. However, in communications systems carrying very high value data such as financial information, there is also a need to detect and prevent message stream and modification, such as insertion or deletion of financial transactions.

The proposed Data Encryption Standard is a specific cipher (encryption) algorithm that transforms the input (plain text) into a cipher text based on a key. The plain text is encrypted in 64 bit blocks using a 56-bit key. There is a matching decryption algorithm that reverses the encryption algorithm when presented with the same key. Each bit of the cipher text in an output block depends on every bit in the input block from which it was generated, so a change of one bit in the plain text block or the key will result in changes to about 50 per cent of the cipher text. An important aspect of this algorithm is that it can be implemented inexpensively on a single semiconductor chip.

The proposed Data Encryption Standard appears to be resistant to conventional cryptanalysis, although it is breakable in a theoretical sense and has been said to be susceptible to breaking by the use of large, special purpose computers in a reasonable length of time (i.e., 24 hours).

Two rather different methods exist for utilizing encryption in a data communications system: link encryption, and end-to-end encryption. In link encryption, each link has a specific key, and a node uses that key to decipher messages coming in on that link or enciphering messages going out on that link. Since each link has a different key, each node deciphers an incoming message using one key and enciphers that message using a different key for retransmission. If an unauthorized party is able to break the key for a given link, all messages carried by that link are exposed.

In end-to-end encryption, the message is enciphered at the originating host and deciphered at the receiving host. The intermediate nodes cannot decipher a message, so the risk of exposure is less.

The Special Requirements of Data Communications Traffic

To prevent spurious connections to a host or terminal by an unauthorized agent (to prevent playing of recorded messages, or "spoofing"), each end must verify the other's identity by means of knowledge of a primary key. A user identifies himself to the host in plain text. The host uses this identifier to obtain the primary key appropriate to that user and sends a message to the user encrypted in that key to switch to a secondary key sent in the enciphered portion of the message. This secondary key is used for the remainder of the session. The primary key is used only for identification, thus limiting its exposure.

After the terminal has changed to the secondary key, it transmits a standard reply to the host. The receipt by the host of this standard reply in the secondary key verifies the identity of the user.

DATAPAC SYSTEM

The DATAPAC system, which became commercially operational in Canada in June, 1977, is typical of the data communications services that will be generally available to users in the early 1980's, based on systems now in development in Japan (DDX), Spain (CTNE), Germany (EDS), France (TRANSPAC), United Kingdom (EPSS), Scandanavia (NDN), and the United States (TYMNET, TELENET.) For this reason, the characteristics of DATAPAC as seen by the end user are reviewed in this section, along with a brief explanation of its method of operation and cost targets. Additional technical characteristics of a typical X.25 packet switching system are also given.

The basic service provided by DATAPAC is an X.25-compatible public packet switched communications system over a full duplex 50 kilobit/second coaxial cable and microwave facility to ultimately interconnect the 14 largest cities in Canada. The system, which is operated by the Trans-Canada Telephone System, supports two distinct user assigned categories of service- "priority", and "normal". The priority class is normally intended for inquiry-response applications for which low delay is a requirement, while the normal class is intended for bulk data transfer or remote job entry, which are not time critical. Each channel in the system can be up to 9600 bits per second.

A fundamental feature of all systems using X.25 is the virtual circuit. In contrast to a physical circuit, which maintains electrical continuity throughout a session whether data is being transmitted or not, a virtual circuit is a bidirectional association between a pair of terminals over which all data transfers take place in the form of packets. Transmission facilities are assigned only when data packets are actually being transferred. The high degree of sharing made possible by the use of virtual circuits and consequent multiplexing of facilities enables a substantial savings due to economy of scale. The characteristics of a virtual circuit inherent in X.25 and therefore DATAPAC are:

- Full duplex link
- Integrity of data
- Flow control, to match user-to-user speeds
- Sequenced data flow

In addition to the priority/normal class of service, the user options also include the ability to obtain a permanent circuit rather than a virtual circuit (to reduce queueing delays), reverse/normal charging, and closed subscriber group (for security).

Extensive services are provided by DATAPAC, including:

- Error Detection/Recovery—The sending node retains a copy of all packets sent until an acknowledgement is received. If bits are changed in a packet as detected by the cyclic redundancy code check bits, the frame is retransmitted. If a packet is lost, it is retransmitted; if a duplicate packet is received, it is discarded. If a link fails, alternate routes are selected.

- Flow Control—As packets are accepted by the receiving terminal, authorization is sent to the transmitting terminal to release more packets.

- Call Control—As requests for circuits are initiated, the validity of the request is checked. If a failure condition is detected, the call is aborted. After the failure is cleared, the interface is reestablished. When a call is refused or completed, the call is cleared.

- Directory Service—Based on the logical address of the receiving terminal, the appropriate routing is established to sequence the packet to the next proper destination.

The costs associated with transmission of a message through a packet switched system are a very complex function of bandwidth, number of nodes crossed, dwell time in the system, message length, and distance. However, many of these are fixed costs, and the only message parameter that strongly affects incremental cost is message length. Therefore charges in a packet switched system are generally not a strong function of distance. The DATAPAC tariffs as filed with the Canadian regulatory body are shown in Table 2.

The Special Requirements of Data Communications Traffic

Facility	Cost
Each virtual circuit	$2/month
Each permanent circuit	$10/month
Call set up	$0.005
Normal data packet (256 bytes, Vancouver to Montreal)	$0.95 per thousand
Priority data packet (128 bytes, Vancouver to Montreal)	$1.19 per thousand

Table 2. DATAPAC tariffs

Notice there is no charge for bandwidth or duration of call.

FUTURE COMMUNICATIONS SYSTEMS

The planning and implementation of communications systems in the future are going to be both easier and harder than today. It will be easier for those who want to implement rather conventional communications systems because they will be able to simply connect to one of the several digital value-added or common carrier digital systems now being developed. A variety of services will be available, including dial up, circuit switching, and packet switching with a high degree of data integrity, on the order of losing one bit per year on the average.

Communications systems planning and implementation will be more difficult for those who need a dedicated system to meet improved cost, performance, or reliability goals beyond those provided by the value-added or common carrier networks. The proliferation of network architectures, protocol standards, carrier services, hardware products in the form of front end processors, concentrators, network processors, and communications software will make the ~~design and implementation of~~ a custom communications system a formidable task.

No matter which approach is taken, value-added network, common carrier network, or custom design (the large organization may use all three), the technology options available will be the same, with the following exception. Since computers are used for both communications functions and data processing functions, it is often desirable to combine both functions on the same computer. Government regulations prevent common carriers from combining functions; however, private network systems using only transmission facilities from common carriers can combine functions and thus gain an added measure of efficiency. The following describes the system functions and components that will be used by any communications network in the next several years.

The greatest difference between communications systems in the past and in the 1980's is that at that time the communications system will have an existence of its own rather than being dependent on a host; so that the communicating system will continue to function even if one or more hosts fail. The host will be treated as an optional peripheral attached to the communications system. The network processor will perform the communications functions now performed by several, more specialized products, including:

- Front-end processor
- Remote/line concentrator
- Host (interface functions)
- Store-and-forward message switching
- Communications node

A front-end processor now interfaces the host processor to the communications system, often to a number of low speed dedicated lines to terminals. The front-end processor function is to perform highly repetitive functions of terminal polling, message assembly/disassembly, error detection/correction, and network support if the host fails. The network processor performs this same function independent of the nature of the communications load, be it many low speed lines, a few high speed lines, or a packet switching system. Where formerly it was necessary to multiplex many low speed lines into one high speed line, a concentrator was used; the network processor will also perform this function.

In order to interface terminals to a current communications system, a host must be present. With network systems, the network processor will interface terminals to the communications system with no host present. And network processors, as nodes in a communications system, will provide store-and-forward functions whether the information is text or data, message switched, or packet switched.

The configurations supported will be quite varied, to conform to the specific needs of individual users. The most common configuration will be hierarchical (which includes the star configuration) because of the match to people organizations, which are usually hierarchical. Also, star configurations as drops along a backbone communications link will also be common in larger organizations with decentralized operations. The more general network configurations will also be supported, using rather arbitrary interconnect of nodes with combinations of high speed, low speed, and redundant lines. The supported configurations will also allow interfacing among value added, common carrier, and private networks. Many systems will also allow interfacing of equipments (especially hosts) from various manufacturers, because most medium to large users have hosts from various manufacturers which must be

The Special Requirements of Data Communications Traffic

integrated. This will be accomplished by using either network standard protocols or by using software adapt modules at appropriate places. An example of a complex network system is shown in Figure 3.

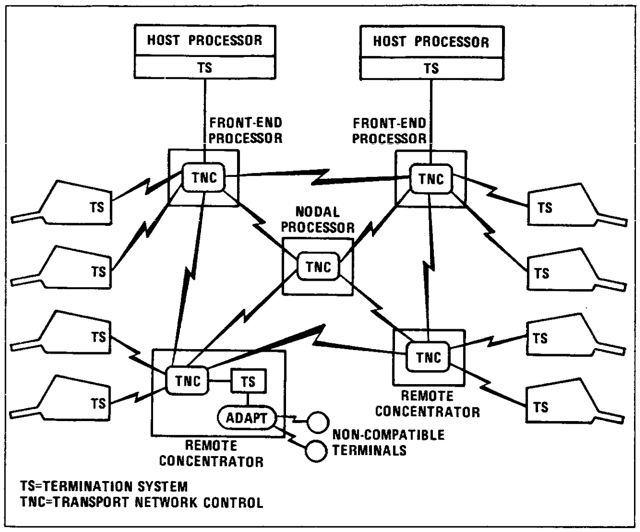


Figure 3. Complex network

The communications system will provide a very high degree of both data integrity and reliability in a manner transparent to the user. End-to-end (i.e., user to user) data integrity will be provided by a combination of:

- Cyclic redundancy checks (error detection codes for bursts)
- Checksums
- Holding packets/messages at transmission end until acknowledgement of correct reception
- Acknowledgement of correct reception
- Packet/message sequence numbers
- Error logging/reporting

Undetected error rates should be less than one bit per year. Reliability will be provided by redundancy at several levels. Redundant network processor configurations, often as multiprocessors, will be provided to enable failsoft operation. Multiple lines and multiple trunks will be used to avoid impact of link failure, with automatic switching by the network processor in case of an outage. The availability of such a system will depend strongly on the degree of redundancy used; therefore, availability will be a strong function of cost of the system.

The various modes of communications now used will be supported, including packet switching, message switching, dial up, and dedicated. It will be possible to mix these in various parts of the system, depending on specific local needs.

A number of services will be provided to the user by the manufacturer-supplied system software, thus relieving him of the very large software development to accomplish these, including:

- Session establishment
- Trunk selection
- Circuit load balancing
- Congestion control
- Directory service
- Circuit selection
- Flow control
- Encryption/decryption

There is also experimental work in progress to provide automatic host load-leveling. In this approach, individual hosts are allowed to "bid" on incoming tasks, depending on the backlog each has at the time and the efficiency of performing that kind of task. Although the feasibility of this approach remains to be validated, it appears to be attractive in a packet switching environment where transmission cost is relatively independent of distance.

Although the communications system software relieves the user of considerable software development, he still must provide the information necessary to configure the software. Because of the large number of options and the considerable complexity of communications systems, this will be a rather complex task, even after the desired system design has been developed. Also, installation, checkout, and acceptance testing of a large communications system will continue to be rather complex, although the total effort will be reduced by an order of magnitude relative to a custom system because of manufacturer supplied software.

The availability of LSI is already having a significant impact on the communications capabilities of new terminals in a number of areas. The increased intelligence available will permit incorporation of sophisticated protocols such as X.25 in the terminal so that it can be interfaced directly to a communications system rather than being attached to a host or a network processor. The terminal will be able to provide functions such as encryption and error detection/correction. Also, terminals in the next few

The Special Requirements of Data Communications Traffic

years are likely to have built-in modems as a low cost option.

A typical total corporate communications system of the mid 1980's is shown in Figure 4. Each location will use a variety of communications services, all digital, including voice, messages, facsimile, and computer data. The sources of the data will be telephone, electronic office work stations, and computer hosts or terminals. The various office locations are linked together by land lines (probably fiber optics) to a regional satellite ground station. A common carrier satellite is used to link together the regional centers, both domestic and overseas. □

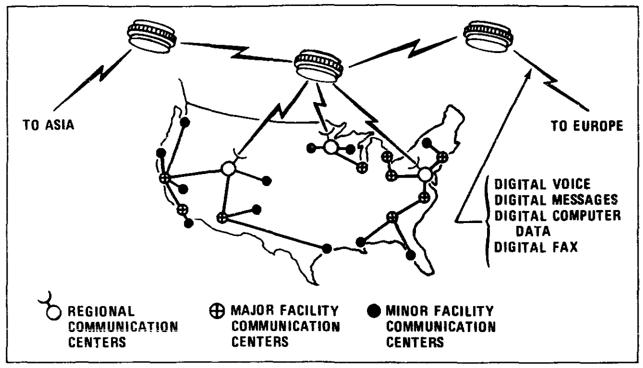


Figure 4. Typical corporate communications network

The Basic Parameters of a Data Channel

Problem:

The simple process of placing a signal on a data channel generates some very unsimple problems. First, the data channel itself must be thoroughly defined before you will be able to know what sort of signal you can put on it. And second, once the signal is on the channel, it is exposed to a host of secondary effects all seemingly intent on destroying the integrity, and thus the intelligence-carrying capability, of the signal. This report is not a direct solution to these problems but it is offered, instead, to clarify exactly what signal deteriorating problems will be encountered on a data channel. Most of the problem solutions will be at least partially incorporated into the equipment you will buy to implement your system, but since a data channel typically passes through many pieces of equipment, you must understand the basic dimensions and electrical parameters of the channel to account for its end-to-end capabilities and limitations.

Solution:

It is axiomatic that the most important part of a communications system is the channel. Its bandwidth limits the volume of information that may be transmitted within certain time limits. Any imperfections, such as interference and distortions, that may be present also affect the maximum possible volume of information and its accuracy; in addition, they contribute to the complexity of the terminal devices. A channel's cost is a major factor in the design and use of a communications system.

The balance of this discussion will cover the characteristics (bandwidth, interference distractions, distortion, and net loss) of communications channels in general, and their effects on information passed through these channels. Because most telephone circuits were originally designed for voice transmission, much of the terminology and technology is a holdover from voice transmission.

Modern Data Communication by William P. Davenport. Chapter 4. © 1971 Hayden Book Co. Reprinted by permission of Hayden Book Company, Inc., Rochelle Park, N.J.

BANDWIDTH

Frequency Defined

Electrical currents can be classified in two general categories—direct or alternating. Direct current (d c) travels in only one direction in a circuit, while alternating current (a c) travels first in one direction (+), then in the other (–).

Figure 1 shows graphically an alternating current starting at a zero value, going through its positive phase and returning to zero, then going through its negative phase and again returning to zero. This is one cycle of alternating current. Either half of the cycle is called a baud.

The Basic Parameters of a Data Channel

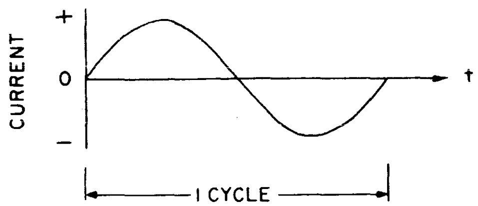


Figure 1. One cycle of alternating current

The number of cycles completed per second is the current's frequency, and is expressed in hertz.[1]

The Frequency Spectrum

Frequencies vary over an extremely wide range, beginning at zero and increasing progressively through acoustics, radio, infrared (heat), light, ultraviolet, X-rays, gamma rays, and cosmic rays. The acoustic range is from about 20 Hz to about 20,000 Hz and varies considerably from person to person. The radio range extends from about 14 kHz to over 10 million kHz.

Figure 2 indicates the disparity between the frequencies that can be detected by the human ear and those that can be transmitted over a telephone channel. The human voice (100 to 1100 Hz), however, falls mostly within the limits imposed by the telephone circuit.

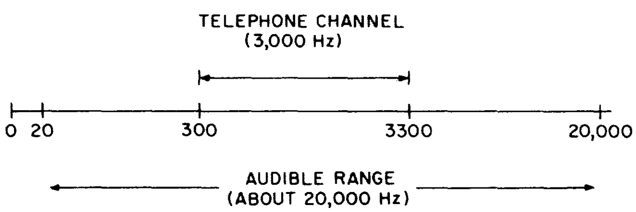


Figure 2. Frequency disparity between human ear and telephone channel

Frequency of a Data Signal

A data communication signal often consists of a range of frequencies. The frequency of a signal at any particular moment depends on the composition of the code being transmitted. For illustration, assume a character with the binary representation of 11110000 is transmitted. If binary 1's are given a positive voltage and 0's a negative voltage, only one cycle would have been transmitted during the time required to transmit the character: the line voltage would have gone from zero to a positive voltage (for the duration of the 1 bits) and then would have swung down through zero to a negative voltage while the 0's were transmitted.

---

[1]The term hertz, abbreviated Hz, has taken the place of cycles per second to promote international understanding.

On the other hand, if a character with the binary equivalent of 10101010 were transmitted, four cycles of current would have occurred during the same character-time. In fact, transmission of the second character would have resulted in the highest possible frequency for that particular signal, since it had caused the greatest number of transitions from one signal state (positive) to the other (negative). Thus the number of bits of transmission channel can carry per unit of time is directly related to the upper limit of its usable frequency range.

Bandwidth and Passband

Bandwidth is a measurement of the width of a range of frequencies. The telephone channel described in the above graph has a bandwidth of 3000 Hz (3 kHz).

A passband, on the other hand, is a slot at a certain place in the frequency spectrum that allows a particular range of frequencies (bandwidth) to pass. The passband of the telephone channel was defined by its limits of 300 and 3300 Hz. Notice that passband defines a particular slot in the frequency spectrum, while bandwidth defines a range of frequencies.

A major difference between the grades of available channels (teletypewriter, voice television, etc.) is their bandwidth. Bandwidth has a great deal to do with the quality of a received signal as compared with the signal originally transmitted. Intelligibility of the human voice, for example, requires a bandwidth of about 400 Hz. Articulation—the pitch and tonal qualities of a voice—requires about 1200 Hz of bandwidth, three times that of intelligibility.

Cutoff Frequencies

Since communications media often have many simultaneous conversations (or other information) imposed upon them, it is necessary to restrict each conversation to its own path. The electrical filters used for this purpose create a passband that allows frequencies within a certain range to pass through the circuit but will block all frequencies outside this range. The points at the upper and lower edges of the passband are called cutoff frequencies. (See Figure 3.)

DISTORTION

If it were possible to transmit a signal over a channel that had no imperfections, the signal would arrive at its destination exactly as it had been sent. Perfect channels, however, exist only in theory; thus signals become distorted during transmission.

Noise is an unpredictable phenomenon that is best described statistically. Distortion, however, has a fixed affect on a signal and is a function of each

The Basic Parameters of a Data Channel

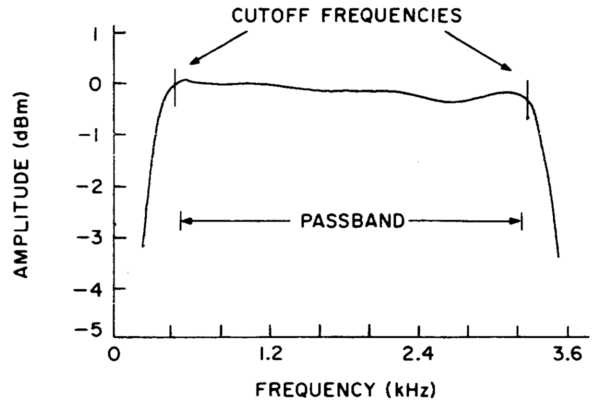


Figure 3. Passband created by using filters

individual channel. There are three types of distortion that a channel may impart to a signal: delay distortion, attenuation distortion, and jitter.

Delay (Phase) Distortion

To have some degree of understanding about the effects a channel imparts on signals it carries, it is helpful to examine the characteristics of a perfect signal and to compare these with the distortion signal that appears at the receiving end of a channel.

Signals can be constructed in many ways. In the following discussion, signals composed of alternating currents of various frequencies will be discussed. First, a simple single-frequency wave is shown, then, a complex two-frequency wave, and finally, the effects of channel distortion will be examined.

Simple and Complex Waves

A simple (single-frequency) wave is shown in Figure 4. The simple wave (signal) shown here can be represented by a rotating vector. Vectors of this type are assumed to rotate counterclockwise, with the reference point, zero degrees, at the right. One complete rotation of the vector represents one complete cycle of the signal. Since the waveform shows that one cycle of the signal is completed at 1 millisecond, the signal frequency is 1000 Hz. Therefore, the vector will rotate 1000 times per second.

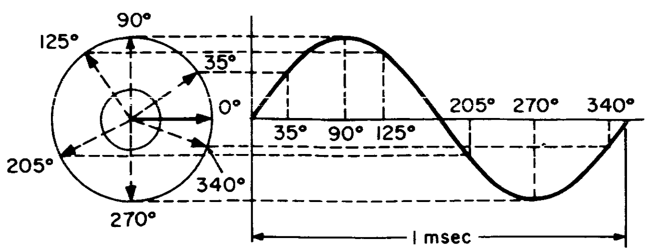


Figure 4. Representation of a simple wave

Complex waves result from combining two or more simple waves. The instantaneous value of a complex wave is the sum of the instantaneous values of the simple waves it comprises. Figure 5 represents a complex wave (A + B) and its two components (A and B) as the signal was transmitted and as it would be received if sent through a perfect channel. Effects of channel imperfections on signals are discussed in following paragraphs.

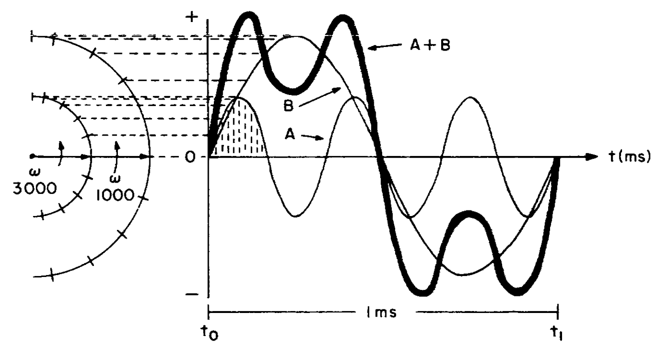


Figure 5. Plotting complex waves (A + B)

Propagation Time

A signal's propagation time is a function of both the nature of the channel through which it is travelling and the frequency of the signal itself. A channel with no external influences and no resistance would move a signal at a maximum of 186,000 miles per second (the speed of light). A microwave carrier might move the signal at 100,000 miles per second (0.01 ms per mile), and a pair of wires in a cable would pass the signal at about 14,000 miles per second (0.0714 ms per mile). An additional 1.2 ms is used whenever the signal passes through terminals that convert from carrier to wire (or vice versa), filters, and other equipment.

If the channel is distortionless, all frequencies will pass through it at the same speed. Under these circumstances, the frequency and the phase of any given signal will have a constant (linear) relationship with respect to time (see Figure 6). The interval of time between the instant that a signal is transmitted and the instant it is received is called phase delay, absolute delay, or propagation time. All signals have some phase delay, but the mere fact that a signal arrives later than it was sent poses few problems in data transmission.

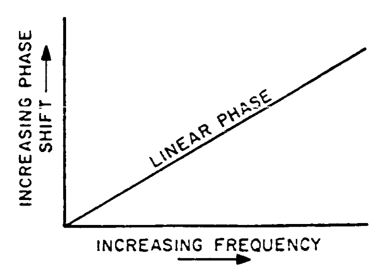


Figure 6. Phase delay through a distortionless channel

The Basic Parameters of a Data Channel

Envelope Delay (Phase Delay Distortion)

A cause of difficulty in data transmission is that the shift of phase with respect to frequency is not usually linear in most transmission media but has a curve similar to that of Figure 7. Under these circumstances, some frequencies of a complex signal will be delayed more than others during transmission, resulting in a distortion of the original signal. This phenomenon is called envelope delay, phase delay distortion, or merely phase distortion.

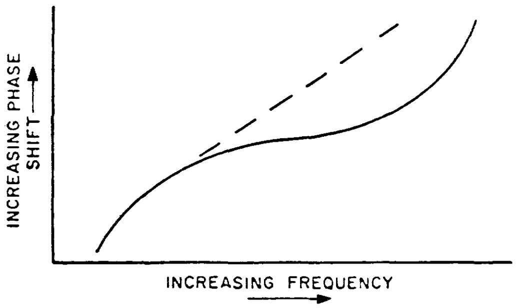


Figure 7. Phase delay of normal voice channel

Mathematically, envelope delay is defined as the first derivative of phase delay. This means that the shape of the envelope delay curve reflects the degree of change in the slope of the phase vs frequency curve. A linear phase shift vs frequency (dotted line of Figure 7) results in no envelope delay (dotted line of Figure 8), while a nonlinear phase shift vs frequency results in a distortion of the transmitted signal.

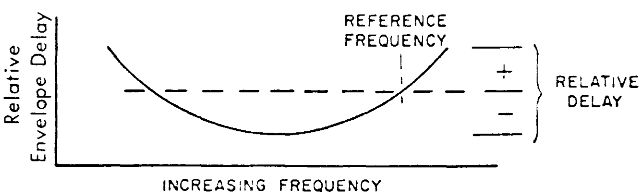


Figure 8. Relative envelope delay curve

If we were to transmit a binary signal at 1000 bauds, each signal element (baud) would have a duration of 1 ms. A relative envelope delay of only 1 ms between the mark and space frequencies, then, would cause the two frequencies to be superimposed at the receiver, obliterating the signal.

To understand the effects of envelope delay, compare the following three diagrams. The first, Figure 9, is the same as Figure 4 and shows the perfect complex waveform as it was transmitted at t = 0. The 1000Hz and 3000Hz components are in phase at t = 0.

The second diagram, Figure 10, shows the complex wave as transmitted at t = 250 ms. Although the two

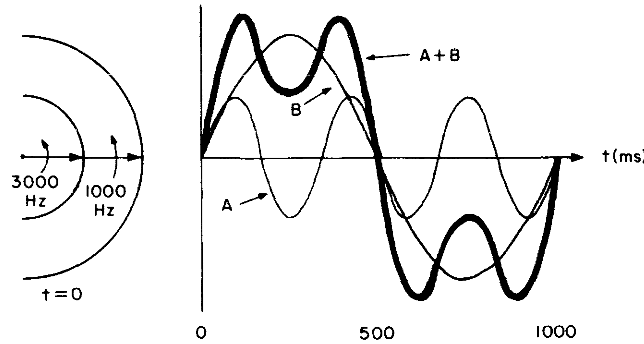


Figure 9. Perfect complex waveform as it was transmitted at t = 0

vectors are not superimposed as they were at t = 0, they are still in phase, since one vector is rotating three times as fast as the other. As long as they maintain this 3-to-1 relationship, they will be in phase with each other.

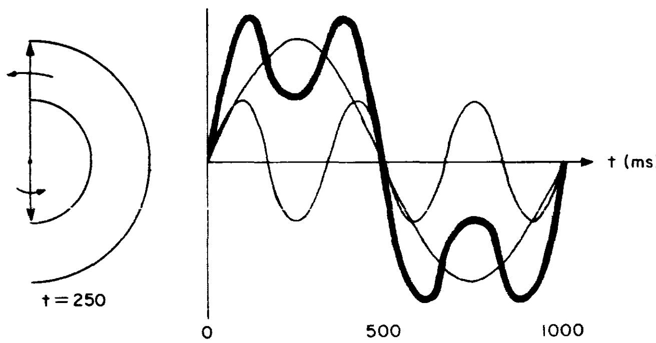


Figure 10. Complex waveform as it was transmitted at t = 250

The third diagram, Figure 11, shows what happens to a complex waveform when its components fall out of phase. The transmitted signal was the same as above, but the channel has delayed the low-frequency component, thus changing the resulting waveform.

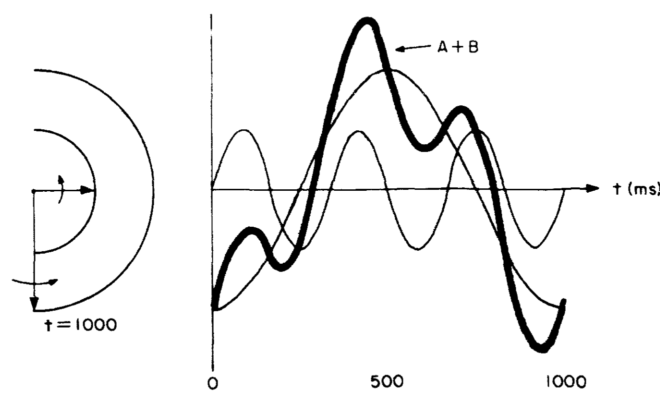


Figure 11. Complex waveform with its components out of phase

The Basic Parameters of a Data Channel

Delay Equalizers

Although envelope delay will always be present in communications circuits, it is important to flatten this delay across the frequency bandwidth of the channel to minimize delay distortion. This is done by adding "delay equalizer" networks to the channel circuits (see Figure 12). The equalizers introduce a delay inverse to that of the channel alone, and the cumulative result is a relatively flat delay across the channel's bandwidth.

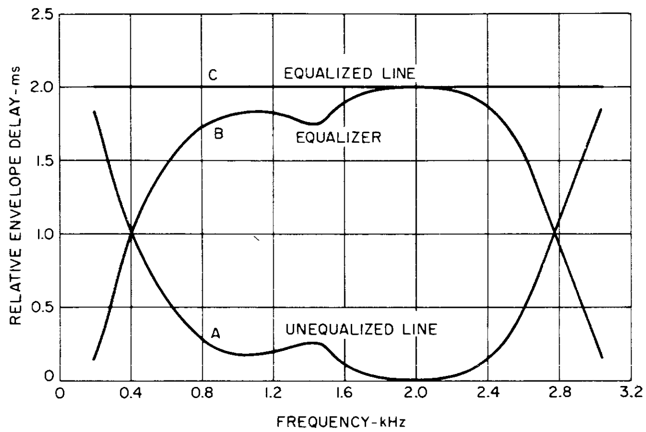


Figure 12. Flattening envelope delay by using delay equalizers (Courtesy of Rixon Electronics)

DISTRACTIONS

Distractions Defined

The term "distractions" is usually used to define extraneous audible signals that appear on the line during a telephone conversation. Distractions consist of crosstalk, echo, or both.

Crosstalk

Crosstalk is the "slopping over" of the contents of one telephone circuit to another circuit. This causes voices and music, which normally should not be present, to be audible in the second circuit.

The cause of crosstalk is the induction of currents by one circuit to physically adjacent circuits. Many circuits, composed of a pair of wires, pass through cables that may contain over two thousand such pairs but that are not more than 2½ in. in diameter! This proximity often provides enough inductive force to cause crosstalk. Factors that tend to increase the effects of crosstalk include the frequency and the strength of the source currents and the distance that the circuits parallel each other.

Echo

Echo is the return of your own voice during a telephone conversation. It usually happens only on long-distance calls and is caused by an electrical imbalance of the circuit.

If A is talking and if any of his speech energy is looped back toward him by an unbalanced electrical network at B, then A will hear his own voice as an echo (see Figure 13). A contributing factor is that most present circuits carry voice signals at less than 20,000 miles per second. If the distance and the speed of the circuit are such that any echo returns in more than 45 ms, then it will be noticed.

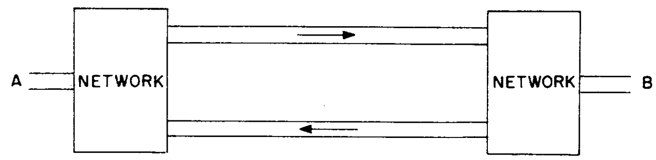


Figure 13. Simplified telephone circuit

Echo Suppressors

Echo suppressors were developed to eliminate this distraction, and they are in wide use today. As A begins to talk, his voice energy activates a relay which short-circuits the return path, thus blocking any echo. When A stops talking, the echo suppressor is quickly deactivated (10 ms) and B can begin talking (see Figure 14). This operation is the cause of the phenomenon you may have observed during a long-distance phone call, where all line noise stopped when you started talking, and you seemed to be talking to yourself. You may also have noticed that the distant party had tried to start talking before you were through, but you couldn't hear him until you stopped talking and the line seemed to "open up" again.

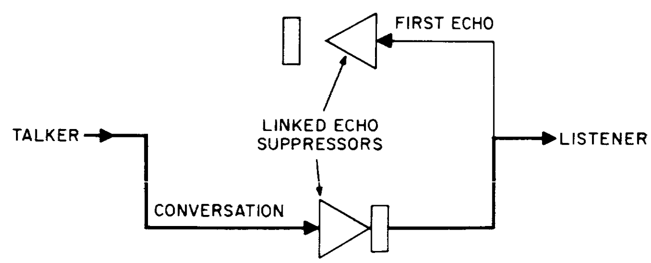


Figure 14. Principle of echo suppression (Courtesy of IBM)

Disablers

Echo suppressors work well for what they are intended to do, but they cause serious problems for data transmission. First, any business machine or data set using voice circuits equipped with echo suppressors must allow up to a second for turnaround time—the time it takes to reverse transmission directions. Second, and more important in many cases, is that if an echo suppressor is activated by a transmission from machine A, any attempt by machine B to send back an interrupt signal to A will be blocked. Disablers deactivate echo suppressors by applying a 2025 Hz tone to the line for about 300 ms (± 50 ms) when no other signals are being transmitted. Any interval of 100 ms or more will reactivate the echo suppressors.

INTERFERENCE

Interference Defined

While distractions are annoying to people during telephone conversations, they may or may not have an adverse effect on data transmission. Interference, on the other hand, often is a source of errors during data transmissions and can arise from induction, noise, or multifrequency tones.

Induction

Both magnetic and electrical induction can cause interference in parallel circuits. This is particularly true as the volume of data being transmitted increases, since data is applied more continuously to circuits than is voice. What difference does this make? Quite a bit, since the amount of induction depends largely on the cumulative strength of any signals present.

Suppose, for example, that a cable carried a voice conversation on each of its 100 pairs of wires and that each voice was well within required strength limitations—caused no induction, in other words. There are gaps in voice conversation, so that at any instant there may have been no more than 75 voices and a tolerable amount of noise actually present within the 100-pair cable. The 75 voices produced a limited and harmless inductive force.

Now let's substitute data transmissions for each of the 100 voice conversations. The gaps between words have been eliminated, so that now we have continuous energy eminating from the equivalent of nearly 100 circuits, rather than the equivalent of the 75 or so we had before. If there are electrical or magnetic imbalances within this group of circuits or if one or more circuits contain signals or noise of excessive power, the entire group of circuits can be affected to the extent that they are unable to carry error-free data.

NOISE

Channel noise consists of random electrical impulses. These unwanted signals are introduced by a variety of sources and are generally classified as either impulse noise or white noise. Noise is considered as interference when it causes errors in transmission.

Impulse Noise

Impulse noise is usually caused by the operation of machinery and switches and by electrical storms. It is characterized by its short but intense duration and its confinement to a limited portion of the frequency spectrum. Within the audio range it can be heard as sharp clicks or bursts of static. (See Figure 15.)

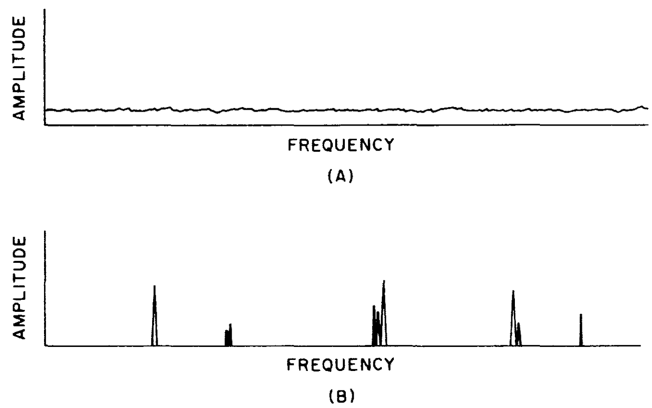


Figure 15. Amplitude and frequency distribution of (A) white noise, and (B) impulse noise

White Noise (Gaussian Noise)

White noise, on the other hand, has its energy spread out over a broad range of the frequency spectrum and is heard as the familiar background hiss on the radio or telephone. Its causes include powerline induction, cross-modulation from adjacent circuits, and a conglomeration of other random signals. One explanation for using "white" to describe this type of noise is that it causes the snowlike phenomenon seen on TV when the signal is weak.

Noise as an Error Source

Noise becomes bothersome when it exceeds a magnitude of about half that of a positive code element. This is because samples are taken of a signal, and if noise exceeds the decision level, the noise is interpreted as a positive signal (see Figure 16).

Effect of Noise on Channel Capacity (Shannon)

Since the unwanted signals that are noise have many of the same characteristics as an information-carrying signal, we must find some way of creating a clear distinction between the two. Fortunately, the power level (intensity) of noise is quite low on most circuits. If the power of the information signal is considerably

The Basic Parameters of a Data Channel

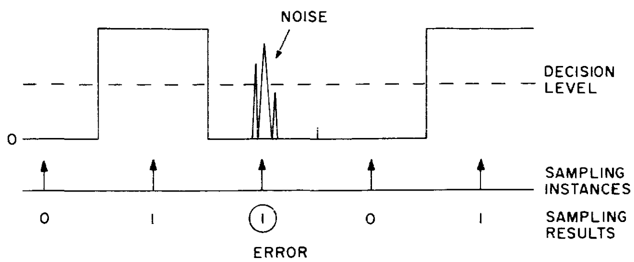


Figure 16. Effect of noise on a binary signal

above that of the noise, the receiving equipment can more easily distinguish between them. As the signal and the noise approach the same power level, while channel bandwidth remains constant, the signal has to exist for longer periods of time in each of its discrete conditions or states to enable the receiving equipment to distinguish it from the random states of the noise.

C.E. Shannon did some pioneering work in this area in 1949 and developed a theory stating that the theoretical maximum bit rate, C, through a channel of bandwidth BW and signal-to-random-noise-power ratio of S/N (where S = signal power and N = noise power) is given by this formula:

$$C = BW \log_2(1 + S/N)$$

The S/N power ratio indicates the relative strength of the signal to that of the channel noise. It is expressed either in ratio form (10^3 : 1) or decibels (dB). For example, an S/N power ratio of 10^3: 1 also could be expressed as 30 dB; a ratio of 10^2: 1 would equate to 20 dB, and so on.

If we had a perfect channel with a 3000-Hz bandwidth and an S/N power ratio of 10^3: 1, we could use the above formula and calculate the maximum bit rate of the channel:

$$\begin{aligned} C &= BW \log_2(1 + S/N) \\ &= 3000 \log_2 1 + 10^3) \\ &= 3000 \log_2(1001) \\ &= 3000 \times 10 \text{ (approx.)} \\ &= 30{,}000 \text{ bits per second (approx.)} \end{aligned}$$

Note that the coding and modulation methods are not described; they may be nearly impossible to achieve, and certainly would not be economical.

Noise Penalty for Multilevel Code Elements

In the presence of noise, a binary signal is more easily and accurately detected than one using several bits per code element. As the bit content (number of levels) of a code element is increased, a corresponding increase in the S/N power ratio must be made to maintain equal detection results relative to a binary signal. The above formula can be modified to give the required minimum S/N power ratio from a known bit rate and bandwidth.

$$S/N = 2^{C/BW}$$

Applying this formula to binary and multilevel signals will show the extent of the noise penalty required to permit the transmission of various multilevel signals.

The S/N power ratio for a binary signal must first be found to serve as a reference. Assuming a perfect 3000-Hz channel, Nyquist's rate of 6000 bps may be used, with the result that a minimum S/N power ratio of 3:1 is required:

$$S/N = 2^{C/BW} - 1$$

$$S/N = 2^{6000/3000} - 1 = 2^2 - 1 = 3$$

The decibel equivalent of a 3:1 S/N power ratio is:

$$\begin{aligned} dB &= 10 \log S/N \\ &= 10 \log 3 = 10\ (.48) = 4.8 \end{aligned}$$

In contrast to the binary system above, a ternary (three-level) system would require a higher S/N power ratio. The maximum bit rate of a ternary system through the ideal 3000-Hz channel is:

$$\begin{aligned} bps &= 2\ BW\ (\log_2 3) \\ &= 6000\ (1.58) = 9500 \end{aligned}$$

and the required S/N power ratio is:

$$\begin{aligned} S/N &= 2^{C/BW} - 1 \\ &= 2^{9500/3000} - 1 \\ &= 2^3 - 1 = 7 \text{ (approx.)} \end{aligned}$$

The decibel equivalent of a S/N power ratio of 7 is:

$$dB = 10 \log 7 = 8.5$$

The noise penalty of a ternary system relative to a binary system (in an ideal channel) is thus 8.5 — 4.8 = 3.7 dB. A quaternary system requires a minimum difference of 11.7 dB between the signal and noise power levels, and thus has a noise penalty of 11.7 — 4.8 = 6.9 dB above binary. (These are minimum requirements of an otherwise perfect channel and are shown here to indicate the extent of the noise penalty that is required to increase signal speed through a given channel.)

In addition to the limit that bandwidth and channel noise (reduced signal-to-noise-power ratio) impose on the bit-carrying capacity of a given channel, other

The Basic Parameters of a Data Channel

channel imperfections and limitations of present equipment impose a practical minimum S/N power ratio in the range of 10^2: 1 (20 dB) or more.

Multifrequency Tones

A relative newcomer to the classification of interference is a group of multifrequency tones. These tones, sometimes heard during telephone conversations, sound somewhat like musical horns on an automobile. The source of the tones varies from Touch-Tone dialers on telephones and teletypewriters to data-set signals, and their presence on a circuit is becoming more prevalent. Since tones are the language of data equipment, the presence of unwanted tones caused by induction is sometimes misinterpreted as valid data by various business machines or data sets that converse in these tones.

ATTENUATION

Definition

In communications, attenuation refers to the loss of power a signal suffers as it travels from the transmitter to a receiver. In other words, it is the power that is absorbed by the transmission medium.

Measurement

One of the most practical ways to measure attenuation is by applying logarithms to the ratio of the input (transmitted) power to the output (received) power. The common logarithm of a number is merely how many times 10 must be multiplied by itself to give that number. The logarithm of 100, for example, is 2, since 10 must be multiplied by itself twice: 10 × 10 = 100. The logarithm of one million is 6: 10 × 10 × 10 10 × 10 × 10 × 10 = 1,000,000. In the shorthand of mathematics, this would be written: log 1,000,000 = 6.

Since log 10 = 1 and log 100 = 2, it follows that log 47 will be a value somewhere between 1 and 2. (One would probably guess that it might be about 1.5, which is close, because it's 1.672.)

Decibels

The units that result from taking the logarithm of the ratio of input power to output power are called bels; one-tenth of a bel is a decibel. If an amplifier produced a power output, P_O, of 100 watts from an input, P_I, of only 1 watt, then its gain would be the results of dividing P_I into P_O, or a ratio of

$$P_O/P_I = 100/1 \text{ or } 100$$

This amplification can also be expressed by the logarithm of the ratio: log 100 = 2.

The result above would be called a gain of 2 bels, or 20 decibels (20 dB for short). Thus, a decibel equals 10 times the logarithm of the result of dividing the power output by the power input:

$$dB = 10 \log P_O/P_I$$

One further point: A decibel, alone, does not have an inherent value; it merely indicates the relationship between two degrees of power. Thus, a 10 dB loss gives a good idea of the attenuation of a signal over a circuit, but it gives no indication of the original strength of the signal.

Assume a signal being transmitted with 1.2 mW (milliwatts) of energy, and a channel that absorbs .6 mW. What is the dB loss of the signal as it passes through the channel?

Answer:

$$\begin{aligned} dB &= 10 \log P_O/P_I \\ &= 10 \log 0.6/1.2 \\ &= 10 \log 0.5 \\ &= 10\ (-0.3) = -3 \\ &= 3 \text{ dB loss} \end{aligned}$$

Attenuation Distortion

High frequencies lose strength more rapidly than low frequencies during transmission through a medium, thus a received signal can be distorted by an unequal attenuation, or loss, of its component frequencies. In the example in Figure 17, the high frequency has experienced more loss of strength through attenuation than the low-frequency portion of the signal.

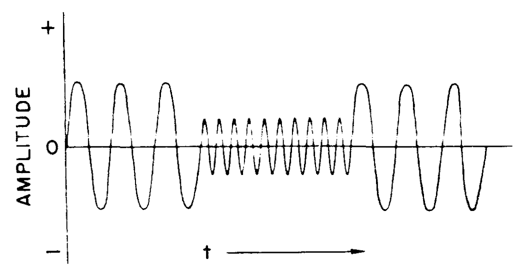


Figure 17. Effect of attenuation distortion

To overcome this attentuation distortion, attenuation equalizers are used. These are electrical networks that have frequency losses complimentary to those of the line. Thus, when added to the line circuit, they give a net result of equalizing the loss of all frequency components of a signal. (See Figure 18.)

The Basic Parameters of a Data Channel

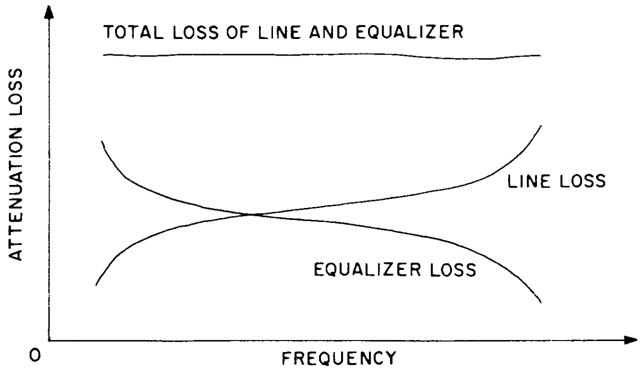


Figure 18. Attenuation distortion correction

NET LOSS

Definition

Net loss is merely the difference in power of a signal between its source and its destination, and is measured in dB.

dBm

Because the decibel, as a measure of the ratio of two powers, has no intrinsic value, it often would be helpful to have a known value to use as a reference to express the power level of a signal. The Bell System developed such a standard: 1 mW of power at 1000 Hz into a 600-ohm resistance impedance. This is called "zero dBm." If a signal has more power than this standard, it has plus dBm and if it has less than this standard power, it has minus dBm.

dBrn

A 1000-Hz tone with 1 mW of power could cause interference in a voice or a data circuit. To reduce this noise (interference) to a negligible level would require lowering it to about -90 dBm. Rather than refer to noise levels in negative terms (such as -40 dBm) the -90 dBm level was taken as a standard, and measurements made in reference to this standard are in terms of dBrn (dB above reference noise).

It should be evident by now that some sort of reference should always be given with an expression using dB. "A 10-dB gain" has meaning, whereas "a 40-dB signal" lacks definition.

JITTER

Definition

Since two frequencies are used in most cases to transmit a binary signal, the circuit characteristics will affect each in a different way. The combined effects of both attenuation and delay distortion results in peak distortion, or jitter. Jitter, then, is the variation in time of the received sequence of tone transitions as compared to the time sequence in which they were transmitted.

Discussion

If mark and space frequencies are both from a portion of the passband that has relatively flat attenuation and delay characteristics, jitter will be at a minimum. If the frequencies were affected similarly to those in Figure 11, however, jitter would probably be quite noticeable, or even troublesome.

To realize how jitter can become so much of a problem, consider that the function of the receiving terminal is to recreate squarewave pulses that duplicate those transmitted by the originating business machine. To do this, it must watch for zero crossings (transitions) in the incoming signal. As the level of the incoming signal crosses zero as it goes from a positive to a negative value, the output of the receiving terminal changes instantaneously to a negative d-c value. As long as the received signal stays on the negative side of zero, the output will be a steady negative d-c value. When the incoming signal again crosses zero in its transition to a positive state, the terminal will convert to a steady positive output.

If the transmitted signal becomes distorted during its course through the channel, the timing of its zero crossings at the receiving end will not correspond to those when it was transmitted. As the severity of the channel distortion changes, the zero crossings of the output of the receiving terminal will also fluctuate, causing jitter in the resultant squarewave signal. If the receiving terminal is deriving its timing from the received signal, or if the sampling is out of phase with the transmitter, errors in the interpretation of the signal by the receiving terminal are likely.

Jitter is measured by dividing A, the variation in received time of the transitions between frequencies, as compared to the transmitted time sequence, by B, the time required to transmit two bits (see Figure 19). Under these terms the maximum possible distortion is 50 percent.

$$\text{Percent jitter} = \frac{\text{Max. var. in transition time} \times 100}{\text{Time of two bits}}$$

$$= \frac{A \times 100}{2B}$$

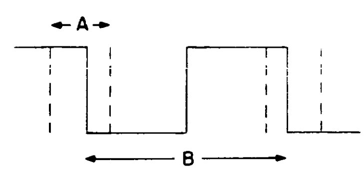


Figure 19. Calculation of percent jitter (Bell System Monograph 3580) □

The Strengths and Applications of Digital Data Multiplexing

Problem:

Multiplexing may seem like a bit of something-for-nothing electronic magic to the uninitiated, but it is really a very simple and unambiguous additive process. A multiplexed channel is like an initially empty four-gallon bucket that can be filled by one gallon from each of up to four separate sources (less is permitted). The electronic trickery comes in keeping the water segregated by source while the bucket is filled and then in neatly separating the water into its original four parts when the bucket is emptied. It's not quite as complicated as it sounds because electronic signals, at least in this case, are easier to deal with than water molecules.

The advantages of multiplexing are most obviously translatable into direct and very visible cost savings because a relatively small investment in multiplexing equipment can produce some astounding multiples in the information-carrying capacity of a data channel. The techniques and hardware of multiplexing are thus very valuable assets to anyone who is planning to put together a communications system.

Solution:

WHY MULTIPLEX?

Digital data multiplexers were invented to save money. They usually make it possible to save very large amounts of money, and they often pay for themselves in less than three or four months. The savings are realized primarily through the reduction of the number of transmission lines required to transmit a given number of data channels. Additional savings that can be achieved by multiplexing include a reduction in the amount of peripheral hardware, such as modems and computer front-end ports, and the substitution of leased-line facilities for the dial-up network.

From *Advanced Techniques in Data Communications* by Ralph Glasgal. © 1977 Artech House. Chpts. 1 and 2. Reprinted by permission.

A multiplexer is a hardware box consisting of central common logic and a variable number of channel cards. The process of multiplexing is simply the combining of several signal channels to form one composite data stream. Demultiplexing is the reconstitution of the original signal channels from the aggregate. That multiplexing is possible at all is due to the fact that the digital data channels normally used in communication systems have a greater capacity for transmitting data than do many of the terminals connected to them. A typical example is the ordinary voice line, either dedicated or dial-up—it has a bandwidth of about 3000 Hz, and yet it may be connected to a terminal such as the Teletype model 33, which transmits at a maximum rate of 110 baud. (In this report the term "baud" is used to define the signalling rate of asynchronous terminals and includes all the signalling elements, such as start, data, parity, and stop bits. All synchronous data

The Strengths and Applications of Digital Data Multiplexing

streams, regardless of format, are assumed to contain only data information bits or characters; therefore, the term "bits per second" is used.) In the last ten years, the development of multilevel encoding techniques in modems has made commonplace the transmission of data rates such as 4800, 7200, and 9600 bps; this improvement has enhanced the value of multiplexing techniques and, in particular, has given a boost to time division multiplexing (TDM) at the expense of the older frequency division multiplexing (FDM) method. FDM is an analog technique in which the sum of the bandwidths of each channel and the required guard bands between them cannot exceed the total bandwidth of the channel, which, in the case of voice-grade phone lines, is less than 3000 Hz. Typically, some eighteen teletypes can be frequency multiplexed, compared with 116 teletypes when using advanced modem and TDM equipment. In the future, however, the major disadvantage of FDM will be that a frequency division multiplexer generates tones and is therefore incompatible with the new all-digital transmission networks.

How Multiplexers Save Money

Figure 1 shows one of the simplest applications for multiplexers and illustrates how multiplexing can save such large sums of money. In the system of Figure 1(a), five IBM 2741's in one office in San Francisco communicate with five computer front-end ports in New York. By using a multiplexer as shown in Figure 1(b), four telephone lines have been

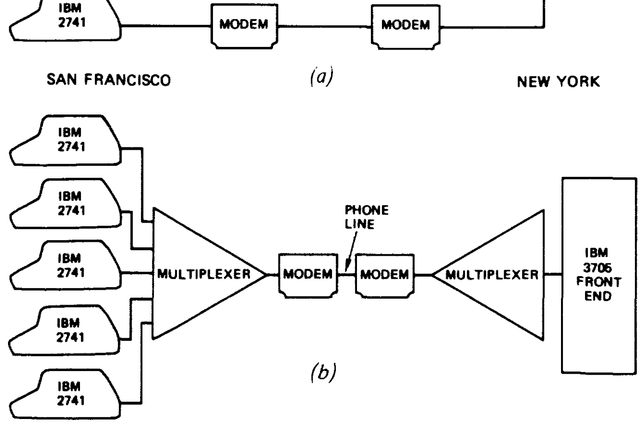


Figure 1. Multiplexers save money by reducing line and modem costs

eliminated. Since the cost of these lines would be approximately $2000 per month, a total savings of $8000 each month is achieved—enough money to pay for the multiplexer in the first month. Furthermore, instead of the ten modems previously required, only two are now needed. If one of the new digital network lines is used, even these modems would not be required, and the savings would be even greater.

Intangible savings are the simplicity of the cabling and the convenience of the centralized diagnostic control that a multiplexer can offer.

Time Division Multiplexing Fundamentals

Time division multiplexing is a method of transmitting several messages on the same circuit by interleaving them in time. A time division multiplexer apportions all the bits available on an aggregate transmission line to the many slower lines being multiplexed. Each aggregate line slot is sequentially occupied by data from one of the sources being multiplexed. Therefore, only one signal occupies the channel at any instant. This situation may be contrasted with that in FDM in which each signal occupies a different frequency band and all signals are transmitted simultaneously.

The best analogy to a time division multiplexing system is a freight train network. A train is assembled periodically in the freight yard with an engine and a string of boxcars. Cargo in the form of a data package is waiting on the loading platform and is loaded into boxcars that have been reserved for it. If there is no cargo ready for a given boxcar when the train pulls in, that boxcar goes out empty when the train leaves. At the end of the journey, the data packages in the boxcars are unloaded in order and are delivered to their proper recipients. Note that it is not necessary to address each shipment. The position of the cargo on the train is sufficient to identify its source and its ultimate destination. To make the system work efficiently, trains must run often enough to prevent an excessive accumulation of cargo on the loading platform, but yet not so often that a lot of boxcars make the trip empty. The train represents the multiplexer frame (headed by a sync character rather than an engine); the tracks are analogous to the transmission line; and the energy to power the train comes from the modem clock, which also sets the velocity of the train. The analogy is a valid one, and we will use it again in this report to illustrate other multiplexer features and options.

Asynchronous Data Multiplexing

Asynchronous data is defined as data from terminals that generate start-stop characters; Teletype models 28 through 40 and the IBM 2741 are just such terminals. They generate a character preceded by a start bit and concluded by stop bits. The job of an asynchronous data multiplexer is to recognize such characters and to assemble them into an aggregate high-speed data stream for transmission over a line to a demultiplexer which reverses the process. Figure 1(b) shows a typical basic system.

For reference purposes, Table 1 is a list of the asynchronous data speeds, formats, and character rates found in the normal multiplexing environment. TDM's must be capable of handling almost any combination of these speeds and codes.

Data Rate (Baud)	Equivalent in Characters per Second	Number of Data Bits	Type of Parity Bit	Minimum Number of Rest Bits	Code Name
50	6.67	5	none	1.5	Baudot
75	10	5	none	1.5	Baudot
110	10	7	even	2	ASCII
134.5	15	6	odd	1	IBM
150	15	7	even	1	ASCII
200	20	7	even	1	ASCII
300	30	7	even	1	ASCII
600	60	7	even	1	ASCII
600	66.6	6	odd	1	IBM
1050	140	5	none	1.5	Baudot
1200	120	7	even	1	ASCII
1200	171	5	none	1	Baudot

Table 1. Most common asynchronous character codes, speeds, and formats

Bit- vs. Character-Interleaved Multiplexers

In assembling a multiplexer train, the question arises as to what size to make the boxcars. The bigger they are, the more time it takes to collect a unit of cargo to fill one, and therefore the longer the delay in completing a shipment. But if the boxcars are too small, we may have to disassemble the cargo in order to load it into the car. To reassemble the cargo, we must send assembly instructions with it; these instructions take boxcars of their own, take time to interpret, and decrease shipping efficiency.

The usual choice is between using character- or bit-sized boxcars, and the decision is not an obvious one. In a character multiplexer, each boxcar is exactly one character in size. In a bit-interleaved multiplexer, each boxcar holds only one bit of a character, but extra bits have to be sent in order to know how to assemble the characters from these bits—so bit-interleaved TDM's are almost faster but less efficient than their character-interleaved relatives. In general, multiplexing by character permits more non-data bits, such as the start and stop bits, to be stripped from incoming characters before multiplexing.

In an asynchronous character demultiplexer, a character is offloaded intact and then is packaged easily for local delivery with a start and the required number of stop bits. It is clear where the start and stop bits should be inserted because each character arrives on the train as a discrete entity. In the bit-interleaved approach, it is necessary for the freight handler to keep a count of the bits received and to add the start and stop bits periodically where they belong. But he must be told where to start counting, which effectively means an extra bit must be sent; what is easier and more commonly done is to multiplex the start bit with the data bits.

Bit-interleaved TDM's require less storage (smaller loading platforms and lighter trains) and are better suited to synchronous data multiplexing; for asynchronous data, though, the bit-interleaving technique usually lacks both efficiency and option flexibility and is not suitable for use in intelligent or software multiplexing systems. In most of the discussion that follows, character interleaving is assumed.

There are two major types of systems in which the shorter delay of a bit-interleaved TDM is a critical factor—Telex and Echoplexing. In the case of Telex channel multiplexing, a rapid transmission of call setup signals, such as "Proceed to Dial" pulses, is essential, lest their late arrival be misinterpreted as "Trunk Busy" or "Disconnect" signals. Typical character-interleaved multiplexers have delays about twice as long as can be tolerated by most existing Telex networks. Both frequency division multiplexers and bit-interleaved time division multiplexers are fast enough to be used in Telex systems without ill effect. The bit-interleaved units can handle more channels but require special logic to first recognize and then multiplex and demultiplex Telex call setup signals with unusual formats, such as dial pulses and busy signals.

In Echoplexing, an operator at a typewriter keyboard depresses a key. The character generated is sent to the computer, which returns an identical character to the typewriter for printing. The process of echoing each character provides a simple method of checking for errors in the data entry process.

The Strengths and Applications of Digital Data Multiplexing

Normally the echo comes back so fast that the operator is not aware of the delay and can maintain a satisfactory data entry rhythm. When a TDM is interposed in such a system, an additional round-trip delay of as much as four character times is introduced (about .4 of a second) and some operators find this difficult to work with. FDM or bit-interleaved mulitplexers work well in this environment. Local loopback options on character-interleaved units offer only a partial solution to this problem since, although the delay is eliminated, the error-checking feature is lost as well. Intelligent character-interleaved TDM's can offer minimum delay echoplexing with the error-checking feature.

Maintaining Frame Synchronization

Frame synchronization is needed to maintain port-to-port multiplexer integrity and to define the character bytes. The synchronization system insures that what goes in on Channel One at one end comes out on Channel One at the other. There are non-dedicated systems where port-to-port integrity is not significant, since terminals being multiplexed identify themselves to the computer; but in most systems, some means of maintaining channel order and detecting a slippage in sync (and regaining it if there is a loss) are necessary. In all systems, character integrity is essential.

Returning to the train analogy—in order to avoid the necessity of putting address labels on each piece of cargo or sending bills of lading with each train, the source and destination of the cargo in each boxcar are preordained by the position of the boxcar in the train (counting from the engine). Trains are recognizable because each has an engine at its head. It would be quite wasteful if trains set out only occasionally, so on the multiplexer freight line, trains run continuously, caboose touching engine. With trains and freight cars continuously arriving and departing, engines must be clearly identifiable so that one can keep track of boxcar positions. In TDM's, engines are frame synchronization characters and are constantly searched for and monitored to insure that the data is distributed properly. Since engines do not carry cargo, it is inefficient to use short trains and many engines. Likewise, using too many sync characters or short frames is inefficient in time division multiplexing. Long trains mean that it takes longer to detect when the multiplexer is out of sync, which, in turn, means that a lot of data cargo would be misdelivered in the interim. Most TDM's detect loss of frame registration in about one second, depending on length of frame and aggregate data rate. Under average conditions about 10 characters per channel may be misdirected before the loss of frame sync is detected. This loss of sync delay parameter is usually only of significance in military and secure-data applications. Bit-interleaving units can usually detect the out-of-sync condition faster than character-interleaved units.

Multiplexing Efficiency

Since asynchronous terminals operate at all of the rates shown in Table 1, a multiplexer should be capable of multiplexing data at any or all of these rates simultaneously; when intermixing data rates and formats, efficiency should be maximized. If the boxcars on the train are all large enough to carry ASCII characters, then they are inefficient carriers of Baudot characters. Ideally, a multiplexer train should utilize a different size boxcar for each type of data it carries. Four different sizes are theoretically required for 5-, 6-, 7-, and 8-level codes, but few TDM's offer all four.

Even more important in maintaining efficiency in speed intermixing systems is the ability to apportion the boxcars properly among the data shippers. The train runs at a fixed speed determined by the trunk data rate. If one boxcar is assigned to each channel per train and exactly one data character per shipper was waiting to be loaded on each and every train, the efficiency would theoretically be maximized. In practice there must be just a few more boxcars than there are data characters to insure that even if cargo arrives a little faster than expected it can still be handled. Therefore, the normal situation is that all boxcars periodically make a trip either empty or take on a different kind of cargo. Figure 2(a) shows a simple frame that uses none of the sophisticated methods discussed. It has frequent sync characters, equal-size character slots, and uses one slot per channel. Note that a total of eight channels can be accommodated at a trunk rate of 2400 bps.

By adding more boxcars to the train and apportioning them as closely as possible among the data sources in proportion to the amount of cargo they have to ship, efficiency can be markedly improved. Almost all modern multiplexers use the proportional boxcar assignment method. Figure 2(b) shows such an efficient multiplexer frame. Note that the 30 character-per-second channel uses three times as many slots as does the 10 character-per-second user. For comparison, an FDM channel spectrum is shown in Figure 2(c). In most intermixed situations, the modern TDM is substantially more efficient than its FDM counterpart.

Cost-Effective TDM Systems

Once it has been decided to use a multiplexer to save line charges, the question arises as to what multiplexer features yield the most economical system. Some multiplexer characteristics that must be considered are capacity, intermix flexibility, channel

The Strengths and Applications of Digital Data Multiplexing

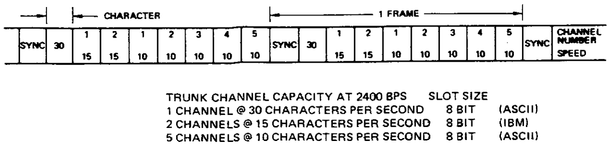


Figure 2(a). Simple frame format of character-interleaved time division multiplexer uses one slot per channel per frame and constant slot size

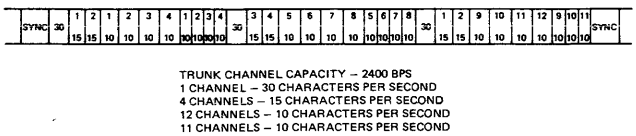


Figure 2(b). Typical character-interleaved multiplexer frame using proportional slot assignment and variable slot size to achieve high multiplexing efficiency

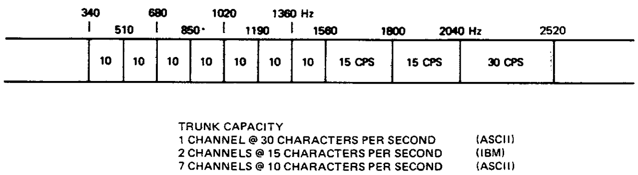


Figure 2(c). Typical FDM format for unconditioned voice-grade line

interfaces, adaptability, expandability, logic organization, mechanical convenience, field programmability, transparency to data and control, multiplexing delay, frame synchronization stability, and diagnostics.

Awareness of these basic characteristics and how they relate to the requirements of a multichannel network makes it easier to select the best multiplexer and options for a particular application. Figure 3 shows a typical point-to-point multiplexing system with several types of co-located and remote terminals being multiplexed.

To determine the amount of money that can be saved by point-to-point multiplexing, a comparison must be made between the cost of the system without the

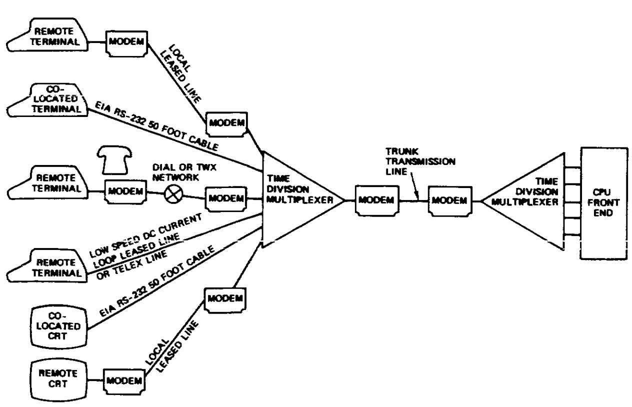


Figure 3. Basic two-point multiplexing system

multiplexer and the cost with it. The expected lifetime of the system is an important consideration in this calculation, since line costs are a recurring monthly expense, and the elimination of even one line can result in a substantial saving if a long enough period is involved. The following equation may be used to calculate the savings that obtain:

$$T(N-1)L - M - M_T + M_L - TLS = \text{saving, where}$$

T = expected lifetime of the system in months

N = number of channels being multiplexed

L = monthly line charge between the points being multiplexed

M = cost of multiplexer

M_T = cost of trunk modems (if any)

M_L = cost of channel modems no longer required

LS = monthly cost of low-speed access lines and modems to multiplexer

Some other factors, both monetary and non-monetary, to add to the above calculation include the cost and availability of maintenance for the multiplexers and their associated modems, the consequences of an outage on the one line between the TDM's or of the TDM's themselves, a possible requirement for a backup line, and the stability of the configuration during its expected lifetime (to avoid early obsolescence of the multiplexer and the cost of frequent reconfiguration).

As shown in Figure 3, virtually any data source can be time-multiplexed, including directly connected, dial-up, synchronous, asynchronous, TWX, Telex, and current loop terminals. The multiplexed data trunk may be transmitted via a modem and analog transmission line, via a digital long-distance line, or through a full duplex dial-up connection. Trunk line transmission is usually synchronous, but asynchronous transmission is also possible.

BASIC TDM SYSTEMS AND FEATURES

The communications manager should examine his data communications network to see if the savings that can be achieved through multiplexing apply to his network. If the network consists of several terminals at one or more locations communicating with a multiport computer at another location, the system is an excellent candidate for multiplexing. It is not necessary for all the terminals to be in the same room with the multiplexer. Terminals may be connected to nearby multiplexers via the voice,

The Strengths and Applications of Digital Data Multiplexing

TWX, or Telex dial networks, dedicated voice grade lines, or high voltage current loop telegraph lines. The major consideration in determining whether to multiplex or not is how much money can be saved. Additional factors, such as grade of service, availability of the system, reliability, and ease of maintenance of the resulting system, should also be taken into account. Once a decision to multiplex is made, a decision as to what type of multiplexer and what features it requires must be made. The following paragraphs are a guide to making this an informed decision.

Mechanical Features

The mechanical design philosophy used in building a time-division multiplexer is much more significant to the user than is the case with, say, a modem, a product that comes quite close to being the legendary black box. A poor mechanical construction technique can cost the user a significant amount of money or inconvenience when even minor reconfiguration or re-installation is required.

All TDM's have common logic, channel interface, and power supply sections. Virtually all multiplexers consist of one or more racks into which are plugged the common logic, the channel interface cards, and sometimes the power supply. Such plug-in card-cage construction makes it easy to effect repairs by means of card exchange. Where multiplexers differ mechanically is in the method used to interface low-speed channels to the central logic. Some TDM's use one interface card per channel. The virtues of this procedure are that when a card is extracted for repair, only one channel is interrupted, and in mixed interface systems, no odd number of channels need ever be left unused. TDM's with 2, 4, and even 24 channels per card are available and have the advantage of low cost and compactness. Multiple channel cards usually have fewer per-channel diagnostics, and any option or strappings must apply to all the channels on the card. Thus, a single two-channel card could not be used to service both an asynchronous channel and a synchronous channel.

Cable connectors are an important part of the cost of a multiplexer. These connectors may be mounted on the channel cards themselves or on the back plane of the card assembly. The advantage of the former method is that one need only pay for the connectors he plans to use and that different types of channel cards can have different types of connectors, saving the cost of cable adapters. This apparent price advantage is not always realized in practice, since several of the lowest-cost, small-system multiplexers mount the channel cable connectors on the back plane and yet are extremely price competitive. If the connectors are mounted on the cards, the cable must be removed when the channel card is repaired, an inconvenience that may be crucial in some environments. If the connector is at the rear of the card, then slack must be left to make it possible to remove the card from the front and then release the cable. In this way, rear access is not essential when making repairs. In at least one TDM, the cable connectors are on the front of the channel cards. While such placement is convenient for installing or moving cables between channels, it is often difficult to get clearance for such bulky cables in the front of standard cabinets, and the cable wires hide diagnostic controls and lamps.

Multiplexing Capacity

Three factors determine how many channels a given multiplexer can multiplex. The first is the mechanical capacity of the system—how many channel cards can be used with one set of central logic? Basic TDM units are available in such sizes as 4, 9, 12, 16, 20, and 24 channels. These units may then be expanded, usually by adding more channel-card cages. The limit in such units is the channel capacity of the central logic and the power supply capacity. In some cases, expansion is achieved by interconnecting complete multiplexers, each with its own common logic and power supply. The former approach is less costly for large systems but requires a large initial investment even when starting small. The latter method becomes unreliable, bulky, and expensive very quickly as the number of channels grows. For example, a 48-channel system would require four complete TDM units if constructed from 12-channel basic units and would therefore be four times as likely to fail as a 48-channel multiplexer using one set of common logic and four expansion rack assemblies.

The second limitation on multiplexing is the data rate at which the aggregate data stream is clocked—a figure usually determined by the modem or facility used, typically on the order of 2400, 4800, 7200, or 9600 bits per second. A multiplexer running at 4800 bps can multiplex twice as many characters per second as one running at 2400 bps.

The third factor in determining multiplexing capacity has to do with the flexibility of the multiplexer logic in frame establishment. Can slots be assigned in exact proportion to the character rates of the data sources? Must start and stop bits be multiplexed? Can smaller aggregate stream characters be used for Baudot or IBM coded data sources? How much overhead is used for frame synchronization characters? The following formula may be used to determine quite closely how many channels of asynchronous mixed-speed data can be multiplexed in a typical multiplexer:

$$(A - \Sigma_i S_i)\ .97 \geq N_1C_1L_1 + N_2C_2L_2 + N_3C_iL_3 + \ldots N_iC_iL_i$$

The Strengths and Applications of Digital Data Multiplexing

where A = aggregate stream data rate in bits per second

S_i = rate in bits per second of any synchronous data being multiplexed

N_1 = number of channels being multiplexed at each speed and code

C_i = rounded-up character rates of the data sources. If C_1 is the highest character rate in the system, then $C_2, C_3, \ldots C_1$ must be rounded up to the nearest whole fraction of C_1, i.e., 1/1, 1/2, 1/3, 1/4, 1/5, etc.

L_2 = number of bits per character multiplexed on the aggregate line for a particular C_i. May be 6, 7, 8, or 9, depending on the sophistication of the particular multiplexer.

The above equation gives good results for most of the multiplexers on the market. It can be made exact for a particular multiplexer by determining the slot size (L) and the intermix fractions (how much the slower channel rates must be rounded up to fit the frame) for the particular TDM being considered. In an older multiplexer without intermix capability, all the C_i's and L_i's are equal to the highest C_i or L_i in the system and are all the same. For example, if A = 2400 bps and we wish to know how many 110 baud channels can be multiplexed with four 300 baud channels in a multiplexer with a fixed high-speed slot of 9 bits per character, no intermix capability, and no synchronous data being multiplexed, then, from the formula:

$$(2400-0)\ .97 = 4 \times 30 \times 9 + N_2 \times 30 \times 9$$

$$N_2 = 4.6 = 4 \text{ channels}$$

Now let us repeat the calculation for a modern multiplexer with efficient speed intermix capability. In this case we assume that a mixed 8- and 9-bit slot size is possible and $C_2 = 10$ (from Table 1)

$$(2400-0)\ .97 = 4 \times 30 \times 9 + N_2 \times 10 \times 8$$

$$N_2 = 15.6 = 15 \text{ channels}$$

The ratio of C_1 to C_2 in this case is 1/3, a whole number fraction, so no rounding up is required.

As a last example, let us check the frame shown in Figure 2(b). In that example:

A = 2400 bps
S = 0

$C_1 = 30$	$N_1 = 1$	$L_1 = 8$
$C_2 = 15$	$N_2 = 4$	$N_2 = 7$
$C_3 = 10$	$N_3 = 12$	$L_3 = 8$
$C_4 = 10$	$N_4 = 11$	$L_4 = 6$

$$2400 \times .97 \geqslant 1 \times 30 \times 8 + 4 \times 15 \times 7 + 10 \times 12 \times 8 + 10 \times 11 \times 6$$

$$2328 \geqslant 240 + 420 + 960 + 660$$

$$2328 \geqslant 2280$$

Since the equation is true, it is possible for a multiplexer to multiplex this combination of channels at 2400 bps if it has a variable slot size, uses an 8-bit slot for ASCII data, and can handle speed ratios of 1/2 and 1/3.

Frame Geometry

In order to establish efficient multiplexing frames, it is necessary to fix a method of frame programming. The bits or characters available must be divided up and assigned to the low-speed channels so that no channel has too few slots, which would cause a loss of data, or too many, which would mean a loss of efficiency. There are as many types of multiplexer frames as there are multiplexer manufacturers; this diversity — plus the variations in frame sync characters — is why TDM's of different manufacturers cannot talk to each other. Quite often even different models of TDM's from the same manufacturer are not compatible. Certain universal principles, however, are common to all TDM's.

First, an overall frame length is established. The length of this frame usually is set so that the frame repetition rate is just slightly faster than the character rate of the slowest channel to be multiplexed. Sync characters mark the beginning of each frame. It is possible to skip sync characters between some frames, but the basic concept is not altered. The frame is then divided into timeslots; sometimes the timeslots are subdivided into character slots. Each manufacturer has different names for these subdivisions (e.g., sections, segments, frame units, bytes, timewidths, fractions), but the principle is the same. Each channel to be multiplexed is assigned to as many slots in the frame (boxcars) as it requires.

The assignment process usually must be done to minimize buffering requirements. For example, in Figure 2(b) the 30-character-per-second channel is assigned three slots and placed at even intervals in every frame. If they were placed adjacent to each other, it would be necessary to accumulate and store up to four characters for that channel instead of two —the additional storage would increase the peak multiplexing delay and the logic cost.

The Strengths and Applications of Digital Data Multiplexing

Frame Programming

Once an efficient frame is established on paper, it is necessary to program the multiplexer by setting the frame length and, if applicable, the trunk modem data rate, and by assigning each channel to the proper slot or slots (analogous to establishing the length of the train, speed of the train, and order of the boxcars).

Multiplexer programming varies with the manufacturer and is done with straps, plugs, cable connectors, wire wrapping on the back plane, and read-only memories. Ease of programming is an extremely important property of a multiplexer. Programming methods that require special parts from the factory or rewiring of logic nests can be both expensive and quite irritating if used in systems where reconfiguration must be done frequently. TDM's that can be reprogrammed in the field by the user are the most desirable, but if special tools or soldering are required, they may not be truly field programmable by semi-skilled computer-room technicians. It is also desirable to be able to change the speed or code of one or two channels without affecting other on-line channels or having to power down the TDM.

When non-volatile ROM's are used as repositories of the frame geometry, it is possible to use several of them, each with a different frame program, and to switch manually or remotely from one program to another as desired. But each time the TDM is reconfigured, a new ROM must be ordered from the manufacturer. Read-write volatile memory has also been used to program multiplexers, but a logic glitch, power shutdown, or failure in such systems requires a reloading of the program. An advantage of the alterable frame memory is that a computer can electronically reprogram the multiplexer system.

Logic Organization and Redundancy

There are two possible approaches to multiplexer design. In the first, almost all buffering and processing is done in common logic; therefore, the channel cards can be simpler and less expensive. In the second approach, each channel card has its own character buffers, frame program counters, and diagnostics; thus, the central logic can be simple and inexpensive. The latter technique has the advantage of decreasing the probability that a logic failure would cause all lines to go down at once. It also makes spare cards or a redundant common logic system less costly. However, such a design approach requires extremely complex channel cards and increases the number of logic devices in a system of even moderate size, thereby significantly increasing the cost, dimensions, and power consumption of the unit. The fact that a very large, and therefore vulnerable, power supply becomes necessary in the channel-card logic approach counteracts its advantages. Furthermore, if only the amount of down time per channel is considered, the system that would have the highest reliability and least amount of down time is the system with the fewest logic elements per channel (per-channel logic plus common logic divided by the number of channels multiplexed) and the least-stressed power supply.

Doing as much as possible in common logic usually results in a relatively inexpensive channel card; therefore, this approach is most economical for large systems. It does mean, however, that while single channel failures are rare, common logic failures are more probable. In this type of system, redundant central logic is a good investment if complete system failure cannot be tolerated. In the first approach, redundant common logic buys little in a system where 95% of the hardware is on the channel cards.

Where redundant central logic is offered, the question arises as to how the central logic can monitor its own failures and switch over. No such system is infallible, but the switchover is backed up by remote and local manual controls. Various signals, such as loss of frame sync, key timing pulses, and the continuous pseudo-multiplexing and demultiplexing of a test character, can all be used to detect a central logic failure and to initiate automatic switchover.

Frame Synchronization Criteria

Barring a logic failure, the only way a multiplexing system can lose frame synchronization is if the modem or device that is clocking the TDM skips a cycle with respect to the data. With good modems, such an event is extremely rare, since the circuits used in such equipment can maintain bit and clock integrity for many seconds even if the transmission line is cut. A multiplexer should not go through frequent resync cycles simply because an occasional sync bit is received in error due to line noise.

A time division multiplexer is, therefore, protected against going into constant resync cycles by logic that looks at many characters or bits over a period of time before deciding that the system is truly out of sync. Most multiplexers have sync systems whose send and receive directions are independent. Thus a remote multiplexer can search for sync on its receive side while still outputting a normal data stream on its send side. This ability makes it possible for the out-of-sync TDM to send a signal to the other mux telling it that it is out of sync.

Overspeed Compensation

Multiplexers reclock asynchronous data when they demultiplex it. It is possible that the demultiplexer clock is marginally slower than the clock used to

The Strengths and Applications of Digital Data Multiplexing

generate the data at its source or that the source is running fast and outputting slightly more characters per second than the demultiplexer is expecting. In the train analogy, this situation means that more cargo is arriving than can be removed from the unloading platform. When the platform is full, the next boxcar cannot be unloaded and its cargo is destroyed. To prevent these losses, there are several possible solutions. One is to provide forklifts and extra unloading equipment. Offloading can now proceed fast enough so that cargo never accumulates; a disadvantage is that cargo may be delivered slightly faster than it is really wanted. In electronic terms, the speed-up is accomplished by increasing the data rate clock enough to insure that all data can be demultiplexed and outputted as fast as it is received. Alternatively, just the rest bit between characters can be shortened to achieve the same effect. However, some terminals are sensitive to such a speed-up of the data or abbreviation of rest bits.

A better, more sophisticated speed compensation method is to note when the loading platform is filling up with cargo and then to accelerate the shipping operation for just as long as it takes to empty the platform. This dynamic method has the advantage of not requiring a continuous amount of fixed speed-up and is particularly useful where multiplexers are used in tandem. Where TDM's are used in series over many shorter links, each succeeding multiplexer must cope with any speed-up in the multiplexer ahead of it by speeding up the demultiplexed data still more. The dynamic speed compensation method reduces the magnitude of any speed-up effect and permits any number of such TDM's to be used in tandem.

A third method (common in other communications equipment but uncommon in multiplexers) is to send a message to the shipper whenever the loading platform is full, requesting a short halt to shipments until the backlog is cleared. This method introduces no distortion in data or rest bits and causes no cumulative speed-up in tandem multiplexer links.

Data Transparency

Virtually all time division multiplexers pass all the characters in the code set used by the data source they are multiplexing. However, in multiplexers that do not multiplex start and stop bits, some means must be found to distinguish between an all-space character and a spacing line and between an all-marks character and a marking line. Or, in other words, the demultiplexer must be told when not to insert rest bits on a spacing line and when not to put start bits on a marking line. The messages that transmit this type of information are usually called control bits or characters, and multiplexers usually transmit many other kinds of control and diagnostic information. Since efficiency requires that control information for a given channel be sent in the same slot as data, the trick is to transmit control data without confusing it with data or interfering with data transmission. Control characters are often distinguished by adding a bit to each character that indicates whether the character is data or control. Control characters are sent whenever no data is ready to be sent. This method is effective but decreases efficiency by 10 to 20%, depending on the character length. A more efficient method can be used with codes that have a parity bit by making normal parity, data and reverse parity, control. Trouble can arise under noisy line conditions, since data characters may be distorted to look like control characters and thus be deleted, or control characters may become data characters and be outputted. However, experience in thousands of systems has shown that these effects are rare and no more disruptive than the errors that occur on unmultiplexed transmission lines.

Adaptive Channel Multiplexers

An option common in the newer multiplexers permits a variety of terminals operating at different rates and codes to access a multiplexer via the dial network. The same phone number is dialed by all users of the system, and the callers are assigned by the telephone equipment to whichever multiplexer channels are free. Since there is no way to determine which terminal will be operating with which channel card, it is impossible to program the channels in advance. The multiplexer must automatically adapt to each caller as the calls are answered. Computer front ends have the same problem when operating in dial-up systems but are now available with adaptive features so that they are fully compatible with an adaptive multiplexer.

There are several methods used to detect the speed of the incoming data. In most cases, a special predetermined first character must be sent by the terminal. This character is detected at some idle speed and is decoded to determine the speed of the terminal that sent it. The multiplexer can then shift from the idle speed to the terminal speed. This character or another control character can then be sent to the remote multiplexer to tell it what the speed of the channel is. Once the multiplexers have adapted, a character can be sent by the terminal or by the multiplexer to cause the computer front end to adapt. Between calls, the multiplexers are reset to the idle speed, based on the lowering of a control signal such as "Carrier Detect."

Another adaptive method is to measure the bit period of the received data to determine its speed, but this method is only reliable if the speeds expected are substantially different and if the first character chosen not to mimic legitimate slower-speed data. A dis-

The Strengths and Applications of Digital Data Multiplexing

advantage of both these types of adaptive multiplexers is that the terminal sign-on protocol for dialing via the TDM is not usually the same as it is for terminals calling a computer directly. This discrepancy can cause operator confusion in circumstances where a terminal normally makes calls to more than one computer, some via multiplexers and some not. A more significant disadvantage lies in the fact that the presently available multiplexers do not reprogram the frame slots adaptively. Therefore, all adaptive speed channels in the mux must be programmed so that enough high-speed line slots are available to service the fastest terminal expected to call in. The result is an average inefficiency that reduces the maximum number of channels that can be multiplexed.

As an example, if forty terminals in a city can dial a local multiplexer and the terminals are mixed 10-, 15-, and 30-characters per second, all the channel cards must be programmed for frame slots of 30 cps. Thus, if a 2400 bps trunk is assumed, only 9 channels can be multiplexed at once, even though the 9 channels might be 10 cps and in theory 29 channels could be handled at this rate. Theoretically, adaptive framing is possible, but at present it appears to be impractical to implement in hard-wired multiplexers. For these reasons, adaptive system users seldom attempt to mix 1200 baud and 110 baud terminals adaptively.

Channel Card Types

A multiplexer, if it is to be truly universal, must interface to as wide a variety of data sources as possible. The most common data interface is the one defined in EIA RS-232. A channel card is needed to connect to terminals and modems using this interface.

Another interface is the high voltage current loop interface commonly used by Baudot coded 50 and 75 baud terminals. Several 2-wire, 4-wire, polar, or neutral current loop interfaces are possible, so a current loop channel card, to be flexible, must have many strap options.

Channel cards may also be provided with built-in 103 or 202 equivalent modems for direct connection to low-speed leased or dial-up lines. Most multiplexers also use different channel card types for asynchronous and synchronous data. Data interfaces compatible with the 303 modem and a bipolar interface compatible with the CSU (Channel Service Unit) of the DDS network are needed in some wideband systems. Mil Standard and Overseas specifications (CCITT) may require still more varieties of channel costs.

Since there is no uniformity in the type of connectors all the various channel card types require, adapter cables are often necessary. The majority of multiplexers use the female 25-pin EIA RS-232 specified connector, so an adapter with a male connector at one end and spade lugs, barrier strip terminals, or some other connector at the opposite end is normally required for all non-RS-232 channels. Those TDM's that can mount the connectors directly on the channel cards do not require such adapters, but such multiplexers are rare.□

The Power of Concentration

Problem:

The difference between multiplexing and concentration is a subtle but very important distinction between treating data as meaningless strings of tagged characters (multiplexing) or treating data with some understanding of the contents and purposes of each string (concentration). A full-blown concentrator in all its glory is a powerful, dedicated computer that can compress data, accommodate data of different speeds, correct errors, convert codes, and perhaps even do some preprocessing to partially synthesize data for a computer. In fact, the newest concentrators are indistinguishable from front-end processors and communication processors. However, you can build concentration into your system from very humble beginnings and can increase your "power of concentration" through a fairly easy succession of upward-compatible equipment, particularly if you begin to think now in terms of microprocessor-based equipment.

Solution:

There is probably more confusion about the distinction between multiplexing and concentration than there is about almost any other topic in data communications. The word "concentrator" is often loosely applied to time division multiplexers, but in this report the term "concentrator" is reserved for devices that feature contention and that process or modify data in some fashion. A concentrator with contention is, in its simplest form, a switch with enough skill to connect on demand any of M inputs to a lesser number of N outputs (M-to-N concentrator). Therefore, a system with six CRT's and only four computer ports will need a concentrator where M = 6 and N = 4. Figure 1 illustrates such a basic concentrator. In this system, each CRT contends with the others for the available computer ports on a first-come, first-served basis. By contrast, a pair of time

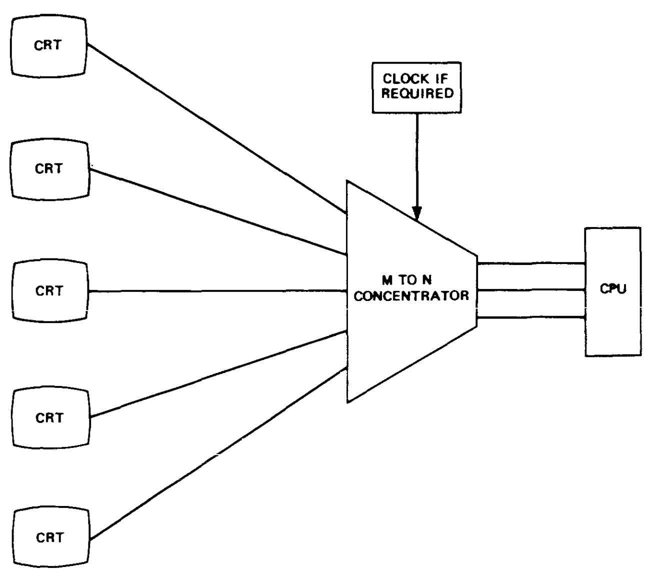


Figure 1. Basic system using 5-to-3 concentrator

"The Power of Concentration" from *Advanced Techniques in Data Communications* by Ralph Glasgal. © 1977 Artech House. Reprinted by permission.

The Power of Concentration

division multiplexers connects all input terminals to the computer all the time.

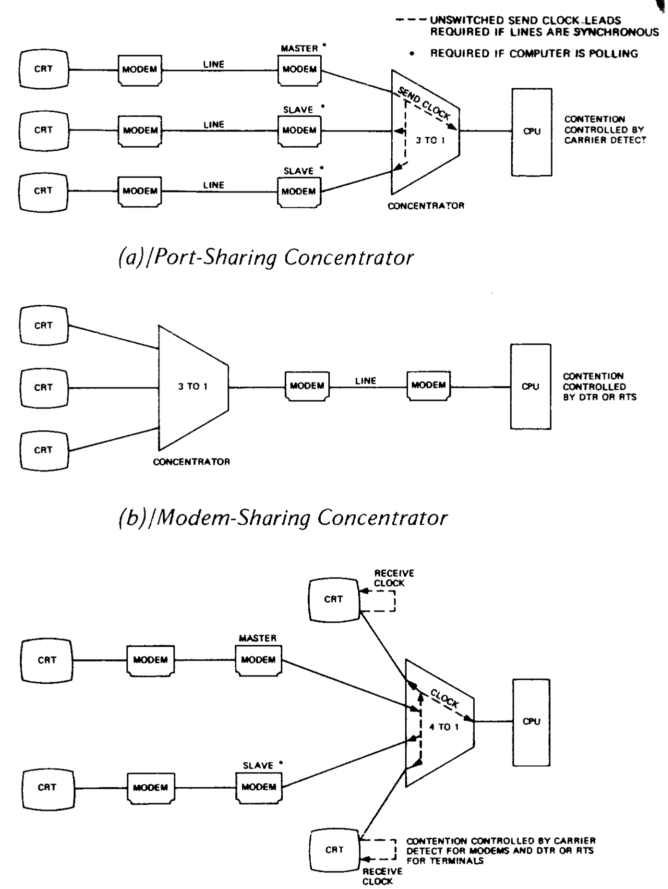
THE PORT-SHARING CONCENTRATOR

In the simplest type of concentrator, several inputs contend for only one output (N = 1). The most common application for such a device is in port- or modem-sharing usage. In Figure 2(a), the remote terminals communicate with the computer via modems and a concentrator. The presence of "Modem Carrier" indicates to the concentrator which terminal has initiated a call and is to be through-connected. Alternatively, the computer can initiate the connection by polling the remote terminals, and when one responds by telling its modem to send "Carrier", the concentrator knows which line has responded and can connect it to the computer port. Note that port- or modem-sharing concentrators usually have a broadcast feature that permits a computer polling signal to be sent out on all lines simultaneously. Figure 2(b) shows the same concentrator used to permit several terminals to access the same modem. In this case, modems and transmission lines are saved, as well as computer ports. Combinations of local and remote terminals may also share a computer port, as illustrated in Figure 2(c). In two of these systems, particular attention must be paid to the clocking if synchronous modems and terminals are involved and if the broadcast mode is needed. In the case where a local terminal is connected directly to the computer via the concentrator, provision must be made to clock the terminal and computer, as illustrated. These applications usually require special cable adapters or straps in the concentrator to implement, and may also require clock drivers if the modem is unable to support all the clock loads directly.

Simple port-sharing devices usually do not use memory and therefore do not delay or store data. Essentially, transfer connections are made instantaneously. Such concentrators are speed-transparent and therefore need no speed or code programming.

ANALOG CONTENTION

Figure 3 shows a telephone-line sharing device. In this device, several phone lines can share a single modem and computer port. The advantage of doing contention on the line side, rather than the digital side, is that fewer modems are required, and dial back-up is easily implemented. This form of contention can only be used in polling systems, and all modems must be of the carrier-controlled type, since the presence of carrier is what causes a line to be connected to the modem. In the send direction, all the lines broadcast the same signal.



(a)/Port-Sharing Concentrator

(b)/Modem-Sharing Concentrator

(c)/Hybrid-Sharing Concentrator

Figure 2.

COMBINING CONCENTRATORS AND MULTIPLEXERS

Concentrators can be placed between a multiplexer and a computer front end, but it is usually better to concentrate before multiplexing, if posssible, in order to reduce the transmission bandwidth and save as many TDM channel cards as possible. Figure 4 shows an interesting application using dial network

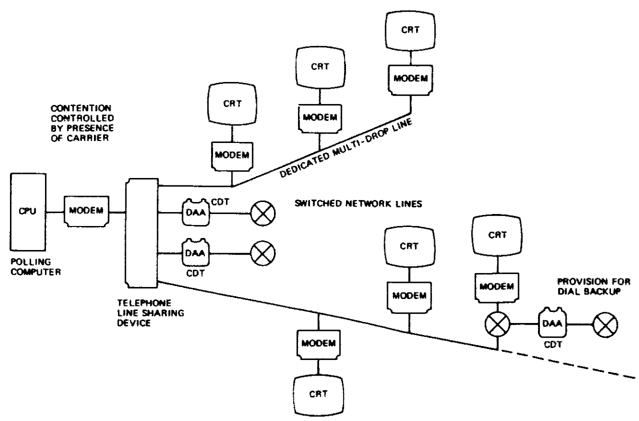


Figure 3. Analog contention unit serves modems in polling systems

rotaries, multiplexers, and several simple 4-to-1 concentrators. The system illustrated permits any terminal in any city to dial a local call and be connected to the computer if a port is free. Once all the ports are occupied, the caller receives a busy signal. The intercity contention is performed by the concentrators at the computer. The intra-city contention is handled by phone company free-line-hunting equipment. Note that the modems used at the remote sites must have the capability of looking busy to the central office in response to an EIA level control signal (out-of-service pin 25) from the multiplexer. This signal is generated by the concentrators and applied to the unselected inputs.

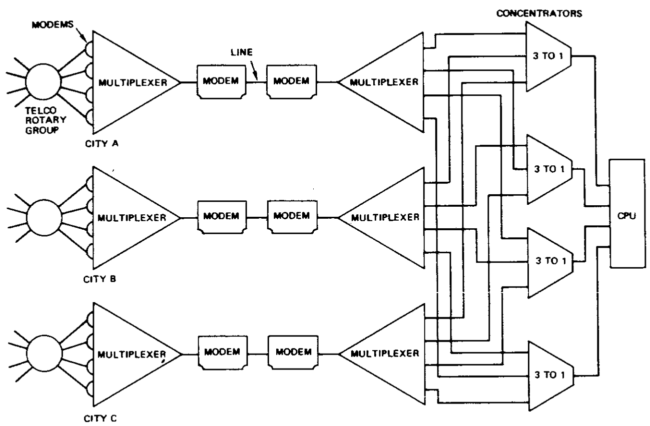


Figure 4. A combination of multiplexers and concentrators provides economical interstate and intra-city contention

LARGE-SCALE M-TO-N CONCENTRATORS

So far we have discussed concentrators with only a single port (N = 1). Such concentrators have little common logic and cost much less than most modems. Where N is greater than 1, a more substantial device is required.

Let us consider the steps required to service a request for a connection in a typical concentrator. The line requesting service first raises a control signal, such as "Ring Indicator" or "Carrier", if a modem, or "Data Terminal Ready" or "Request to Send", if a terminal. The M-to-N concentrator is continually scanning its input lines, looking for an appearance of these signals. When a request is found, the address of the channel is noted in a memory. A search is now made for a free computer port by scanning a memory containing the present status of the ports. When a free port is found, its address is matched with the address of the input line and stored in memory. Now each time an active input line is scanned, data and control signals can be transferred. Most M-to-N concentrators use circulating shift registers as memory elements, and the number of transfers of data per second are limited by the speed at which the memory can be scanned and by the number of lines in the system. Sampling times between five and seventy microseconds are typical; the faster units can operate well at data rates of up to 9600 bps. The sampling technique eliminates the need for speed/code programming in concentrators and permits either synchronous or asynchronous data to be switched.

Almost all M-to-N concentrators also have provisions for forming several subcontention groups. This facility permits lines or ports to be grouped by speed or code. Where synchronous data modems are concentrated, provision must be made to transfer clocks as well as data and control signals. If synchronous data terminals are concentrated, a means of clocking the terminals and computer ports must be provided. This is most conveniently done if the concentrator has an oscillator and a clock bus system built in.

Current loop lines may also be concentrated if the concentrator can recognize an idle-to-active line data transition. The problem here is to determine when disconnect is desired. This problem can be solved, in some cases, by using an EIA interface on the computer side so that the computer can initiate a disconnect by lowering "Data Terminal Ready." Such a concentrator may also be used as an interface converter by setting M equal to N.

Large M-to-N concentrators usually contain displays that make it possible to see which channels are connected to which ports. Switches are also provided to take input lines or output ports temporarily out of service.

MORE SOPHISTICATED CONCENTRATORS

While simple contention-type M-to-N concentrators that respond to control signals have many applications, they are naturally limited to concentration functions that can be performed without examining or reading the data stream or generating characters to be outputted. Some useful functions that one might like a data-inspecting concentrator to perform include port partitioning based on speed or code and port or computer selection based on application. The ability of a concentrator to generate data characters permits "Busy" messages to be sent to locked-out terminals. Both hardware- and software-based concentrators provide all or some of these data-stream dependent features. Elaborate software concentrators go so far in this direction that they become circuit switches or terminal controllers and perform such functions as polling, data compression, backward or forward error correction, and code conversion. Discussion of such devices, which are loosely called remote concentrators, is beyond the scope of this report.

The Power of Concentration

To perform the data-derived functions enumerated above, the sophisticated concentrator must determine or be told the speed and code of the lines contending. As in adaptive multiplexers, a special first character, usually carriage return or circle D, must be sent by the terminal. The multiplexer or concentrator samples the character, using a convenient clock such as 300 bps or 600 bps and decodes the result—which will be distorted in a predictable fashion if the data is actually, for example, 300, 150, 134.5 or 110. Once the speed is known, the code can be inferred. Additional characters then can be used to indicate alternate codes or to select a particular computer port group. Of course, once the speed and code of a terminal are known, it is possible to send the terminal a "Busy" message. Software-based concentrators are much more flexible in this regard and can send "Busy" messages that include a number representing a terminal's position in a queue. Priority schemes of various types are also available so that special channels can be put at the head of the queue of even interrupt a connection already established. Channel usage reports can also be generated.

The ultimate in data-controlled software concentrators, as in multiplexers, is the concentrator that can adapt automatically without requiring special characters. Such units have almost no limitations on the speeds and codes that can be adaptively concentrated or routed and do not require terminal operators to use complex sign-on procedures.

By combining concentration with multiplexing, remote concentration operating over a single communications line becomes possible. Such combinations are less expensive and more flexible if done using software techniques, and the distinction between concentrators, multiplexers and remote terminal controllers becomes blurred.

CONTENTION CONTROL AND INTERFACING

An M-to-N concentrator, like a multiplexer, must appear to be either a modem or a terminal, depending on the device it is attached to. On the output port side, its connectors are usually configured to look like a modem, so that the computer cables may be directly attached. On the line side, the unit looks like a terminal so that modems may be directly attached. Unfortunately, the application of these units varies a great deal, and quite often local terminals contend directly with remote terminals on dial-up or leased lines. In the modem-sharing application, as shown in Figure 2(b), terminals seek a single modem, and so the concentrator must look like a modem to the terminals and like a terminal to the modem.

In most concentrators, the same twisted cable adapters are often required. However, the problem in concentrators is compounded by the fact that a control signal is required to make and break the connections. If a modem on a dial-up line is requesting service, it will raise "Ring Indicator." If the modem is on a dedicated line, it will raise "Carrier Detect" to initiate a connection. If a terminal requests service, it will raise "Data Terminal Ready," or if DTR is on all the time, a switchable "Request to Send" may have to be used. To drop a connection, the computer port may lower "Data Terminal Ready," or the connection may have to be disconnected on the line side by a modem lowering "Carrier Detect," or, if half duplex, "Data Set Ready." A terminal may disconnect by lowering DTR or RTS. If data sources are mixed, it may be necessary to use a combination of these methods to operate. Most concentrators rarely make adequate provision to accommodate all these variations, and special cabling or strapping is often required. Some full-duplex terminals in non-polled systems that have no variable control signals can only be concentrated by turning off power to them or to their associated modems, or by using a software concentrator.□

The Language and Techniques of Pulse Code Modulation (PCM)

Problem:

It appears that the day of the universal, high-speed all-digital data path is not only inevitable but is probably only just a few short years away. The term "all-digital" means that all of the data in the path is encoded in a pulse-type format. There are several pulse encoding techniques that modulate selected pulse characteristics to introduce an information-carrying capability into the data path. For example, the pulse amplitude is modulated in a technique called Pulse Amplitude Modulation (PAM); the pulse width is modulated in a technique called Pulse Depth Modulation (PDM); and pulse position relative to a fixed-position master pulse is used in a technique called Pulse Position, or Pulse Phase, Modulation (PPM). The PAM, PDM, and PPM techniques are very economical—each pulse represents a complete datum—but their economy is offset by their susceptibility to noise, which makes them particularly error prone. A more promising technique, which grew out of NASA's work in deep-space telemetry, is called Pulse Code Modulation, (PCM). PCM is less economical—it usually uses eight pulses to do the same job as one pulse in the other techniques—but PCM pulses are far easier to handle electronically because the pulse heights, widths, and positions are kept constant, which sharply reduces the noise and detection problems.

PCM is directly compatible with the encoding schemes (ASCII, EBCDIC, et all) used inside most computers and can also handle any kind of analog data source, like voice. An analog source is sampled at a rate roughly twice its highest-frequency component, and each sample is converted into its 8-bit (e.g.) binary equivalent. An 8-bit binary number yields 256 combinations, which is enough resolution for all but the most complex analog sources. True, PCM techniques have a poor ratio of packing density for a given bandwidth compared with the other pulse techniques and a very poor ratio compared with purely analog transmission techniques, but the onrushing development of increasingly inexpensive, super-wideband media, like satellites and optical fibers, is slowly but surely making the "wastefulness" of PCM a tolerable exchange for the switching and other control advantages offered by all-digital systems.

The Language and Techniques of Pulse Code Modulation (PCM)

This report offers a comprehensive overview of the theory, applications (with examples), and the advantages/disadvantages of PCM. It is important to recognize and understand the mechanics of this significant trend toward all-digital communications systems.

Solution:

Pulse code modulation (PCM) is a method of modulation in which a continuous analog wave is transmitted in an equivalent digital mode. The cornerstone of an explanation of the functioning of PCM is the sampling theorem, which states: If a band-limited signal is sampled at regular intervals of time and at a rate equal to or higher than twice the highest significant signal frequency, then the sample contains all the information of the original signal. The original signal may then be reconstructed by use of a low-pass filter.

As an example of the sampling theorem, a nominal 4-kHz channel would be sampled at a rate of 8000 samples/s (i.e., 4000 x 2).

To develop a PCM signal from one or several analog signals, three processing steps are required: sampling, quantization, and coding. The result is a serial binary signal or bit stream, which may or may not be applied to the line without additional modulation steps. At this point a short review may be in order so that terminology such as mark, space, regeneration, and information bandwidth (Shannon and Nyquist) will not be unfamiliar to the reader.

One major advantage of digital transmission is that signals may be regenerated at intermediate points on links involved in transmission. The price for this advantage is the increased bandwidth required for PCM. Practical systems require 16 times the bandwidth of their analog counterpart (e.g., a 4-kHz analog voice channel requires 16 x 4, or 64 kHz when transmitted by PCM). Regeneration of a digital signal is simplified and particularly effective when the transmitted line signal is binary, whether neutral, polar, or bipolar. An example of bipolar transmission is shown in Figure 1.

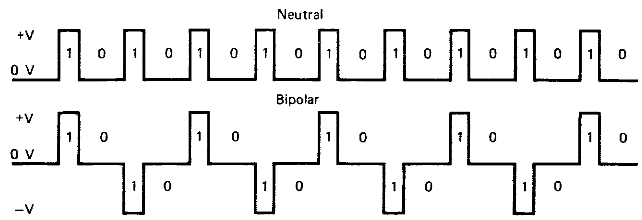


Figure 1. Neutral versus bipolar bit streams. The upper drawing illustrates alternate "1"s and "0"s transmitted in a neutral mode; the lower illustrates the equivalent in a bipolar mode

Binary transmission tolerates considerably higher noise levels (i.e., degraded signal-to-noise ratios) when compared to its analog counterpart (i.e., FDM). This, plus the regeneration capability, is a great step forward in transmission engineering. The regeneration that takes place at each repeater by definition recreates a new digital signal; therefore noise, as we know it, does not accumulate. However, there is an equivalent to noise in PCM systems generated in the modulation-demodulation processes. This is called quantizing distortion and can be equated in annoyance to the listener with thermal noise. Regarding thermal-intermodulation noise, let us compare a 2500-km conventional analog circuit using FDM multiplex over cable or radio with an equivalent PCM system over either medium.

	FDM/radio/cable	PCM/radio/cable
Multiplex	2,500 pWp	130 pWp equivalent
Radio/cable	7,500 pWp	0 pWp
Total	10,000 pWp	130 pWp equivalent (not dependent on system length, see Sec 11.2.6)

Error rate is another important factor. If we can maintain an end-to-end error rate on the digital portion of the system of 1×10^{-5}, intelligibility will not be degraded. A third factor is important in PCM cable applications. This is crosstalk spilling from one PCM system to another or from the send path to the receive path inside the same cable sheath.

The purpose of this report is to provide a background of the problems involved in PCM and its transmission, including the several PCM formats now in use. Practical aspects are stressed later in the report, such as the design of interexchange trunks (junctions)

From *Telecommunication Transmission Handbook* by Roger L. Freeman, Chpt 11, pp 27-29 © 1975 John Wiley & Sons. Reprinted by permission of John Wiley & Sons, Inc.

The Language and Techniques of Pulse Code Modulation (PCM)

and the prove-in distance,* compared with other forms of multiplex or VF cable. Finally, long-distance systems are focused on as well as PCM switching as a method of extending (or shortening) prove-in distance.

Consider the sampling theorem given above. If we now sample the standard CCITT voice channel, 300-3400 Hz (a bandwidth of 3100 Hz), at a rate of 8000 samples/s, we will have compiled with the theorem, and we can expect to recover all the information in the original analog signal. Therefore a sample is taken every 1/8000 s or every 125 μs. These are key parameters for our future argument.

Another example may be a 15-kHz program channel. Here the lowest sampling rate would be 30,000 times/s. Samples would be taken at 1/30,000-s intervals or at 33.3 μs.

The PAM Wave

Most practical PCM systems involve time division multiplexing. Sampling in these cases does not involve just one voice channel, but several. In practice, one system to be discussed samples 24 voice channels in sequence; another samples 32 channels. The result of the multiple sampling is a PAM (pulse amplitude modulation) wave. A simplified PAM wave is shown in Figure 2, in this case a single sinusoid. A simplified diagram of the processing involved to derive a multiplexed PAM wave is shown in Figure 3.

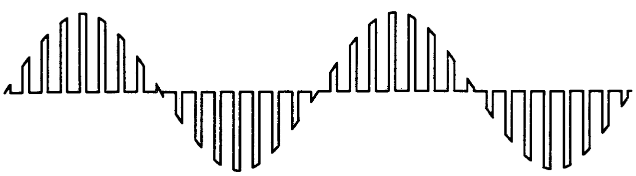


Figure 2. A PAM wave as a result of sampling a single sinusoid

If the nominal 4-kHz voice channel must be sampled 8000 times/s, and a group of 24 such voice channels are to be sampled sequentially to interleave them, forming a PAM multiplexed wave, this could be done by gating. Open the gate for a 3.25-μs period for each voice channel to be sampled successively from Channel 1 through Channel 24. This full sequence must be done in a 125-μs period (1 x 10^6/8000). We call this 125-μs period a frame, and inside the frame all 24 channels are successively sampled once.

Quantization

The next step, it would appear, in the process of forming a PCM serial bit stream, would be to assign a binary code to each sample as it is presented to the coder. For instance, a binary code with four discrete elements (a four-level code) could code 2^4 separate and distinct meanings or 16 characters, not enough for the 26 letters in our alphabet; a five-level code would provide 2^5 or 32 characters or meanings. The ASCII is basically a seven-level code allowing 128 discrete meaning for each code combination (2^7 = 128). An eight-level code would yield 256 possibilities.

Another concept that must be kept in mind as the discussion leads into coding is that bandwidth is related to information rate (more exactly to modulation rate) or, for this discussion, to the number of bits per second transmitted. The goal is to keep some control over the amount of bandwidth necessary. It follows, then, that the coding length (number of levels) must be limited.

As it stands, an infinite number of amplitude levels are being presented to the coder on the PAM highway. If the excursion of the PAM wave is between 0 and + 1 V, the reader should ask himself how many discrete values are there between 0 and 1. All values must be considered, even 0.0176487892V.

The intensity range of voice signals over an analog telephone channel is on the order of 60 dB. The 0-1

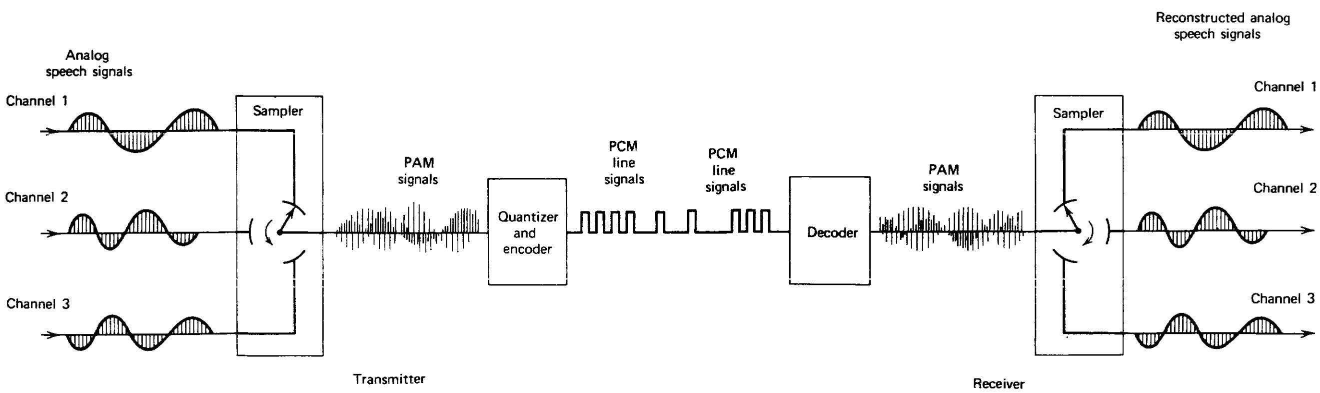


Figure 3. A simplified analogy of the formation of a PAM wave. Courtesy GTE Lenkurt Demodulator, San Carlos, Calif.

*The point at which a PCM system becomes viable economically.

The Language and Techniques of Pulse Code Modulation (PCM)

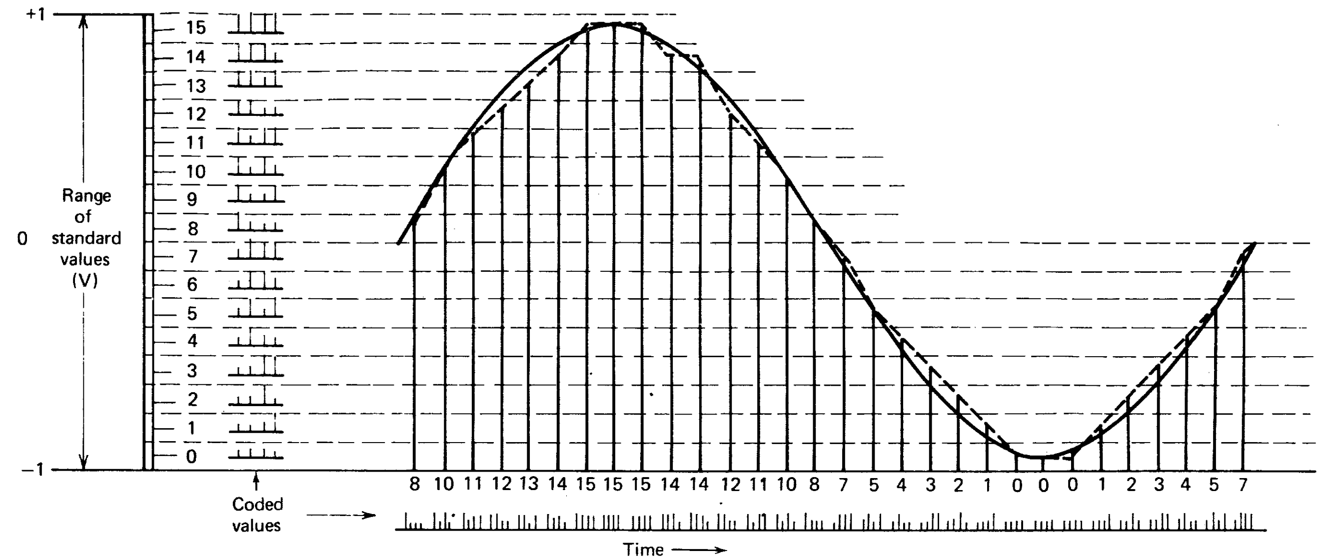


Figure 4. Quantization and resulting coding using 16 quantizing steps

range of the PAM highway at the coder input may represent that 60-dB range. Further, it is obvious that the coder cannot provide a code of infinite length (e.g., an infinite number of coded levels) to satisfy every level in the 60-dB range (or a range from — 1 to + 1 V). The key is to assign discrete levels from —1 V through 0 to + 1 V (60-dB range).

The assignment of discrete values to the PAM samples is called quantization. To cite an example, consider Figure 4 in which 16 quantum steps exist between — 1 and + 1 V and are coded as follows:

Step Number	Code	Step Number	Code
0	0000	8	1000
1	0001	9	1001
2	0010	10	1010
3	0011	11	1011
4	0100	12	1100
5	0101	13	1101
6	0110	14	1110
7	0111	15	1111

Examination of Figure 4 shows that Step 12 is used twice. Neither time it is used is it the true value of the impinging sinusoid. It is a rounded-off value. These rounded-off values are shown with the dashed line in the figure that follows the general outline of the sinusoid. The horizontal dashed lines show the point where the quantum changes to the next higher or next lower level if the sinusoid curve is above or below that value. Take Step 14 in the curve, for example. The curve, dropping from its maximum, is given two values of 14 consecutively. For the first, the curve is above 14, and for the second, below. That error, in the case of "14," for instance, from the quantum value to the true value, is called quantizing distortion. This distortion is the major source of imperfection in PCM systems.

In Figure 4, maintaining the —1—0—+ 1 V relationship, let us double the number of quantum steps from 16 to 32. What improvement would we achieve in quantization distortion? First determine the step increment in millivolts in each case. In the first case, the total range of 2000 mV would be divided into 32 steps, or 187.5 mV/step. The second case would have 2000/64 or 93.7mV/step.

For the 32-step case, the worst quantizing error (distortion) would occur when an input to be quantized was at the half-step level, or in this case, 187.5/2 or 93.7 mV above or below the nearest quantizing step. For the 64-step case, the worst quantizing error (distortion) would again be at the half-step level, or 93.7/2 or 46.8 mV. Thus the improvement in decibels for doubling the number of quantizing steps is:

$$20 \log \left(\frac{97.7}{46.8}\right) = 20 \log \text{ or } 6 \text{ dB approximately}$$

This is valid for linear quantization only. Thus, increasing the number of quantizing steps for a fixed range of input values reduces quantizing distortion accordingly. Experiments have shown that if 2048 uniform quantizing steps are provided, sufficient voice signal quality is achieved.

For 2048 quantizing steps, a coder will be required to code the 2048 discrete meanings (steps). We find that a binary code with 2048 separate characters or meanings (one for each quantum step) requires an 11-element code or

The Language and Techniques of Pulse Code Modulation (PCM)

$2_n = 2048$; thus n = 11

With a sampling rate of 8000/s per voice channel, the binary information rate per voice channel will be 88,000 bps. Consider that equivalent bandwidth is a function of information rate; the desirability of reducing this figure is therefore obvious.

Coding

Practical PCM systems use seven- and eight-level binary codes, or

$2^7 + 128$ quantum steps
$2^8 + 256$ quantum steps

Two methods are used to reduce the quantum steps to 128 or 256 without sacrificing fidelity. These are nonuniform quantizing steps and companding prior to quantizing, followed by uniform quantizing. Keep in mind that the primary concern of digital transmission using PCM techniques is to transmit speech, as distinct from digital transmission, which dealt with the transmission of data and message information. Unlike data transmission, in speech transmission there is a much greater likelihood of encountering signals of small amplitudes than those of large amplitudes.

A secondary, but equally important aspect, is that coded signals are designed to convey maximum information considering that all quantum steps (meanings, characters) will have an equally probable occurrence because practical data codes assume equiprobability. When dealing with a pure number system with complete random selection, this equiprobability does hold true. Elsewhere, particularly in practical application, it does not. One of the worst offenders is our written language. Compare the probability of occurrence of the letter "e" in written text with "y" or "q.") To get around this problem, larger quantum steps are used for the larger amplitude portion of the signal, and finer steps for signals with low amplitudes.

The two methods of reducing the total number of quantum steps can now be labeled more precisely:

- Nonuniform quantizing performed in the coding process.
- Companding (compression) before the signals enter the coder, which now performs uniform quantizing on the resulting signal before coding. At the receive end, expansion is carried out after decoding.

An example of nonuniform quantizing could be derived from Figure 4 by changing the step assignment. For instance, 20 steps may be assigned between 0.0 and + 0.1 V (another 20 between 0.0 and —0.1 etc.); 15 between 0.1 and 0.2 V, 10 between 0.2 and 0.35 V, 8 between 0.35 and 0.5 V, 7 between 0.5 and 0.75 V, and 4 between 0.75 and 1.0 V.

Most practical PCM systems use companding to give finer granularity (more steps) to the smaller amplitude signals. This is instantaneous companding compared with syllabic companding. Compression imparts more gain to lower amplitude signals. The compression and later expansion functions are logarithmic and follow one of two laws, the A law or the "mu" (μ) law.

The curve for the A law may be plotted from the formula

$$Y = \frac{AX}{(1 + \log A)} \qquad 0 \leq X \leq 1/A$$

The curve for the mu law may be plotted from the formula

$$Y = \frac{\log (1 + \mu X)}{\log (1 + \mu)} \qquad -1 \leq X \leq 1$$

A common expression used in dealing with the "quality" of a PCM signal is the signal-to-distortion ratio (expressed in decibels). Parameters A and μ determine the range over which the signal-to-distortion ratio is comparatively constant. This is the dynamic range. Using a μ of 100 can provide a dynamic range of 40 dB of relative linearity in the signal-to-distortion ratio.

In actual PCM systems, the companding circuitry does not provide an exact replica of the logarithmic curves shown. The circuitry produces approximate equivalents using a segmented curve, each segment being linear. The more segments the curve has, the more it approaches the true logarithmic curve desired. Such a segmented curve is shown in Figure 5.

If the mu law were implemented using a seven (eight)-segmented linear approximate equivalent, it would appear as shown in Figure 5. Thus, upon coding, the first three coded digits would indicate the segment number (e.g., $2^3 = 8$). Of the seven-digit code, the remaining four digits would divide each segment in 16 equal parts to further identify the exact quantum step (e.g., $2^4 = 16$).

For small signals the companding improvement is approximately

A law	24 dB
mu law	30 dB

using a seven-level code.

The Language and Techniques of Pulse Code Modulation (PCM)

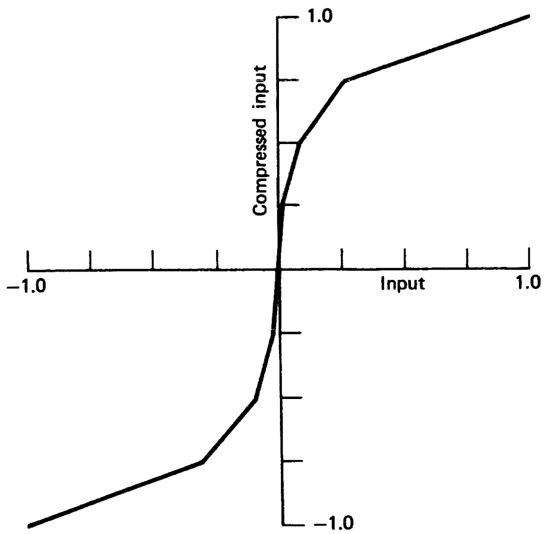


Figure 5. Seven-segment linear approximate of the logarithmic curve for the mu law (μ = 100) Copyright © 1970, by Bell Telephone Laboratories

Coding in PCM systems utilizes a straightforward binary coding. Two good examples of this coding are shown in Figures 7 and 10.

The coding process is closely connected to quantizing. In practical systems, whether using A law or mu law, quantizing uses segmented equivalents as discussed above and shown in Figure 5. Such segmenting is a handy aid to coding. Consider Figure 6, which shows the segmenting used on a 24-channel PCM system (A law) developed by Standard Telephone and Cables (UK). Here there are seven linear approximations (segments) above the origin and seven below, providing a 13-segment equivalent of the A law. It is 13, not 14 (i.e., 7 + 7), because the segments passing through the origin are colinear and are counted as one, not two segments. In this system, six bits identify the specific quantum level, and a seventh bit identifies whether it is positive (above the origin) or negative (below the origin). The maximum negative step is assigned 0000000, the maximum positive, 1111111. Obviously we are dealing with a seven-level code providing identification of 128 quantum steps, 64 above the origin and 64 below.

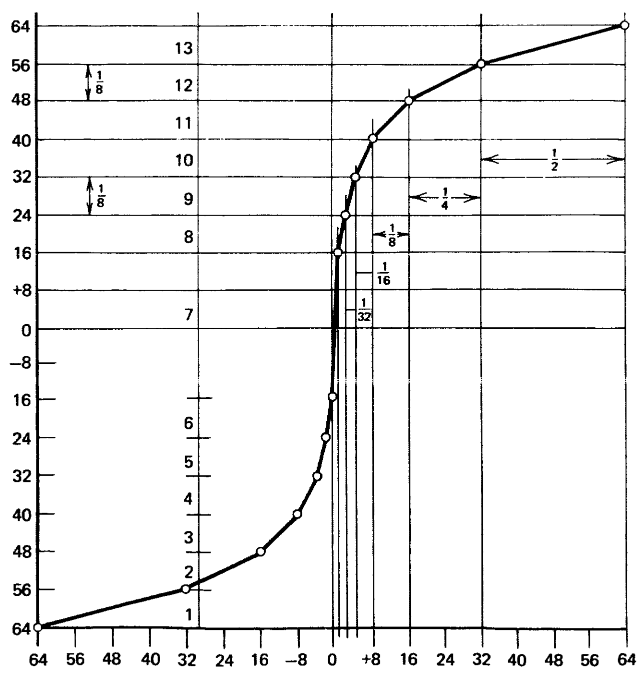


Figure 6. A 13-segment approximation of the A-law curve as used on the 24-channel STC system (PSC-24B). The horizontal ordinate represents quantized signal levels. Note that there are many more companding values at the lower signal levels than at the higher signal levels

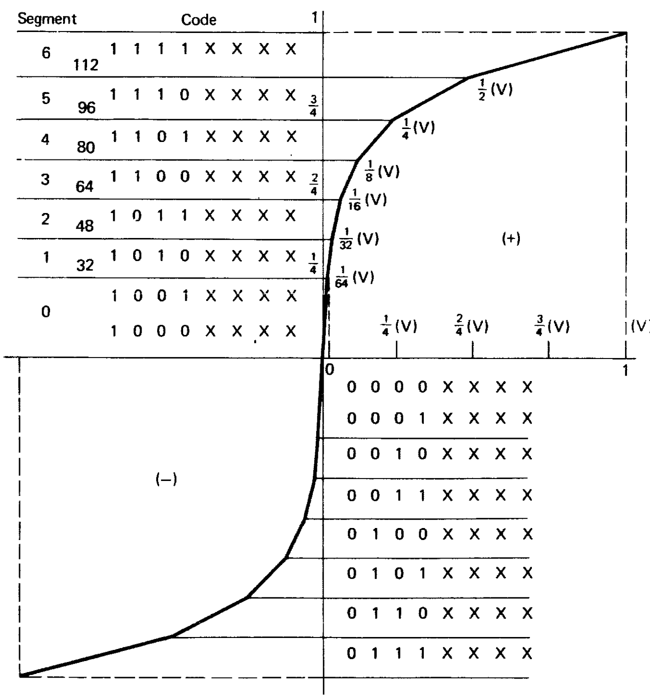


Figure 7. Quantization and coding used in the CEPT 30 + 2 PCM system

The 30 + 2 PCM system also uses a 13-segment approximation of the A law, where A = 87.6. The 13 segments (14) lead us to an eight-level code. The coding for this system is shown in Figure 7. Again, if the first code element (bit) is 1, it indicates a positive value (e.g., the quantum step is located above the origin). The following three elements (bits) identify the segment, there being seven segments above and seven segments below the origin (horizontal axis).

The first four elements of the fourth + segment are 1101. The first "1" indicates it is above the horizontal axis (e.g., it is positive). The next three elements indicate the fourth step or

The Language and Techniques of Pulse Code Modulation (PCM)

0—1000 and 1001
1—1010
2—1011
3—1100
→ 4—1101
5—1110 etc.

Figure 8 shows a "blowup" of the uniform quantizing and subsequent straightforward binary coding of Step 4; the Figure 8 shows final segment coding, which is uniform, providing 16 ($2^4 = 16$) coded quantum steps.

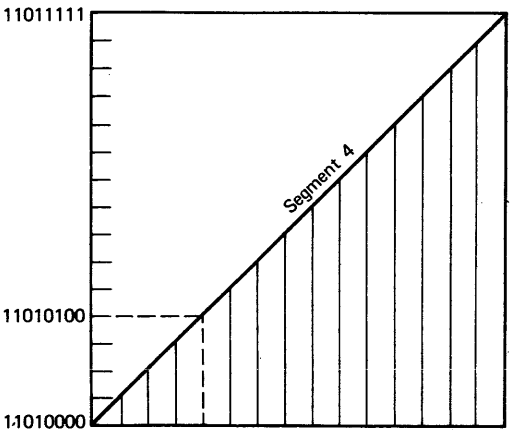


Figure 8. CEPT 30 + 2 PCM system, coding of Segment 4 (positive)

The North American D2 PCM system uses a 15-segment approximation of the logarithmic mu law. Again, there are actually 16 segments. The segments cutting the origin are colinear and counted as 1. The quantization in the D2 system is shown in Figure 9 for the positive portion of the curve. Segment 5 representing quantizing steps 64-80 is shown blown up in the figure. Figure 10 shows the D2 coding. As can be seen in this figure, again the first code element, whether a "1" or whether a "0," indicates if the quantum step is above or below the horizontal axis. The next three elements identify the segment, and the last four elements (bits) identify the actual quantum level inside that segment.

The Concept of a Frame

As shown in Figure 3, PCM multiplexing is carried out in the sampling process, sampling several sources sequentially. These sources may be nominal 4-kHz voice channels or other information sources, possibly data or video. The final result of the sampling and subsequent quantization and coding is a series of pulses, a serial bit stream that requires some indication or identification of the beginning of a scanning sequence. This identification tells the far-end receiver when each full sampling sequence starts and ends; it times the receiver. Such identification is

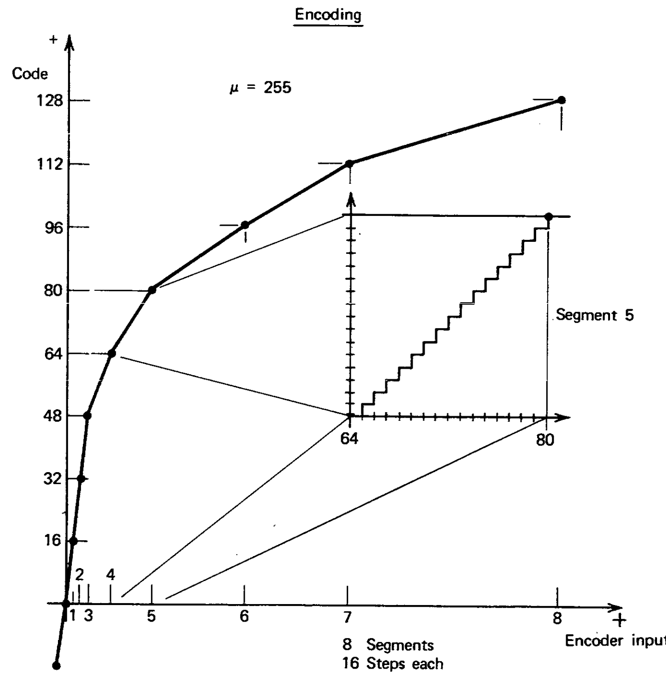


Figure 9. Positive portion of the segmented approximation of the mu law quantizing curve used in the North American (ATT) D2 PCM channelizing equipment. Courtesy ITT Telecommunications, Raleigh, N.C.

called framing. A full sequence or cycle of samples is called a frame in PCM technology.

Consider the framing structure of several practical PCM systems: the ATT D1 System is a 24-channel PCM system using a seven-level code (e.g., $2^7 = 128$ quantizing steps). To each 7 bits representing a coded quantum step 1 bit is added for signaling. To the full sequence 1 bit is added, called a framing bit. Thus a D1 frame consists of

$$(7 + 1) \times 24 + 1 = 193 \text{ bits}$$

making up a full sequence or frame. By definition 8000 frames are transmitted so the bit rate is

$$193 \times 8000 = 1{,}544{,}000 \text{ bps}$$

The CEPT* 30 + 2 system is a 32-channel system where 30 channels transmit speech derived from incoming telephone trunks, and the remaining two channels signal and synchronize information. Each channel is allotted a time slot and we can speak of time slots (TS) 0-31 as follows:

TS	Type of Information
0	Synchronizing (framing)
1-15	Speech
16	Signaling
17-31	Speech

*CEPT = Conference European des Postes et Telecommunication.

The Language and Techniques of Pulse Code Modulation (PCM)

Code Level		Digit Number 1	2	3	4	5	6	7	8
255	(Peak positive level)	1	0	0	0	0	0	0	0
239		1	0	0	1	0	0	0	0
223		1	0	1	0	0	0	0	0
207		1	0	1	1	0	0	0	0
191		1	1	0	0	0	0	0	0
175		1	1	0	1	0	0	0	0
159		1	1	1	0	0	0	0	0
143		1	1	1	1	0	0	0	0
127	(Center levels)	1	1	1	1	1	1	1	1
126	(Nominal zero)	0	1	1	1	1	1	1	1
111		0	1	1	1	0	0	0	0
95		0	1	1	0	0	0	0	0
79		0	1	0	1	0	0	0	0
63		0	1	0	0	0	0	0	0
47		0	0	1	1	0	0	0	0
31		0	0	1	0	0	0	0	0
15		0	0	0	1	0	0	0	0
2		0	0	0	0	0	0	1	1
1		0	0	0	0	0	0	1	0
0	(Peak negative level)	0	0	0	0	0	0	1*	0

Figure 10. Eight-level coding of the North American (ATT) D2 PCM system. Note that actually there are really only 255 quantizing steps because steps "0" and "1" use the same bit sequence, thus avoiding a code sequence with no transitions (i.e., "0"s only)

In TS 0 a synchronizing code or word is transmitted every second frame occupying digits 2-8 as follows

0011011

In those frames without the synchronizing word, the second bit of TS 0 is frozen at a 1 so that in these frames the synchronizing word cannot be imitated. The remaining bits of time slot 0 can be used for the transmission of supervisory information signals (see Appendix C on signaling).

The North American (ATT) D2 system is a 96-voice channel system made up of four groups of 24 channels each. A multiplexer is required to bring these four groups into a serial bit stream system. The 24-channel basic building block of the D2 system has the following characteristics:

255 quantizing steps, mu-law companding, 15-segment approximation, with $\mu = 255$, 8-element code

The frame has similar makeup to the D1 system, or

$$8 \times 24 + 1 = 193 \text{ bits per frame}$$

The frame structure is shown in Figure 11. Note that signaling is provided by "robbing" Bit 8 from every channel in every sixth frame. For all other frames all bits are used to transmit information coding.

Quantizing Distortion

Quantizing distortion has been defined as the difference between the signal waveform as presented

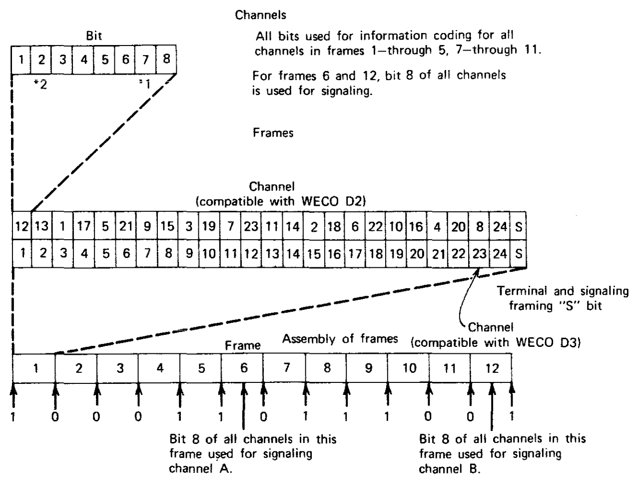


Figure 11. The frame structure of the North American (ATT) D2 PCM system for the channel bank. Note the bit "robbing" technique used on each sixth frame to provide signaling information. Courtesy ITT Telecommunications, Raleigh, N.C. Notes (1) If bits 1-6 and 8 are 0, then bit 7 is transmitted as 1. (2) Bit 2 is transmitted as 0 on all channels for transmission if end-to-end alarm. (3) Composite pattern 000110111001, etc.

to the PCM multiplex (codec*) and its equivalent quantized value. Quantizing distortion produces a signal-to-distortion ratio (S/D) given by

$$\frac{S}{D} = 6_n + 1.8 \text{ dB (for uniform quantizing)}$$

where n is the number of bits used to express a quantizing level. This bit grouping is often referred to as a PCM word. For instance, the ATT D1 system uses a 7-bit code word to express a level, and the 30 + 2 and D2 systems use essentially 8 bits.

With a 7-bit code word (uniform quantizing)

$$\frac{S}{D} = 6 \times 7 + 1.8 = 43.8 \text{ dB}$$

Practical S/D values range on the order of 33 dB or better† for average talker levels using A-law logarithmic quantizing. At the lower limit of commercial speech levels, where noise is most noticeable, the equivalent contribution of psophometrically weighted noise that the PCM segment contributes in a PCM-analog hybrid system is on the order of 100 pWp. This figure does not increase with PCM system length and assumes only one complete segment. Each time the system decodes to analog voice and recodes to PCM, another 100 pWp must be added.

The 100-pWp figure does relate to the design of the codec and depends on such things as the number of quantizing steps and the type of quantizing employed. It also depends greatly on talker volume. Very loud

*The term codec is introduced, meaning coder-decoder and is analogous to modem in analog circuits.
†Using eight-level coding.

talkers may suffer clipping with voice peaks outside the quantizing range or in the section of the quantizing curve where quantizing steps are very large. Very low level talkers likewise tend to show a degraded S/D ratio. For instance, if we have an equivalent sinusoid input to a codec of —33.5 dBm, we then could expect an S/D ratio of about 33 dB with a resulting noise floor of — 66.5 dBm (33 + 33.5), which equates to about 130 pWp of equivalent noise. However, quantizing noise tends to annoy the telephone user more than conventional noise.

Idle Channel Noise

An idle PCM channel can be excited by the idle Gaussian noise and crosstalk present on the input analog channel. A decision threshold may be set that would control idle noise if it remains constant. With a constant level input there will be no change in code word output, but any change of amplitude will cause a corresponding change in code word, and the effect of such noise may be an annoyance to the telephone listener.

One important overall PCM design decision to control idle channel noise is the selection of either the μ or A values of the logarithmic quantizing curve used. The higher the values of these constants, the more finely granulated are the steps (quanitizing steps finer) near the zero signal point. This tends to reduce idle channel noise. Care must also be taken to ensure that hum is minimized at the inputs of voice channels to the PCM equipment (codec).

PRACTICAL APPLICATIONS OF PCM

PCM has found widest application in expanding interoffice trunks (junctions) that have reached exhaust* or will reach exhaust in the near future. An interoffice trunk is one pair of a circuit group connecting two switching points (exchanges). Figure 12 sketches the interoffice trunk concept. Depending on the particular application, at some point where distance d is exceeded, it will be more economical to install PCM on existing VF cable plant than to rip up streets and add more VF cable pairs. For the planning engineer, the distance, d, where PCM becomes an economic alternative is called the prove-in distance. d may vary from 8 to 16 km (5 to 10 mi) depending on the location and other circumstances. For distances less than d, additional VF cable pairs should be used for expanding plant.

The general rule for measuring expansion capacity of a given VF cable is as follows:

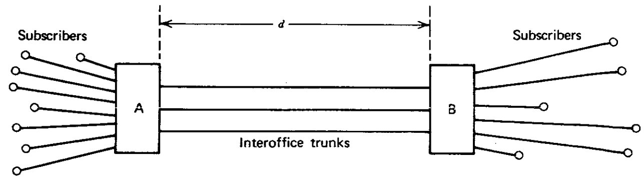


Figure 12. Simplified application diagram of PCM as applied to interoffice plant. A and B are switching centers

- For ATT D1/D2 channelizing equipment, 2 VF pairs will carry 24 PCM channels.
- For CEPT 30 + 2 system as configured by ITT, 2 VF pairs plus a phantom pair will carry 30 PCM speech channels.

All pairs in a VF cable may not necessarily be usable for PCM transmission. One restriction is brought about by the possibility of excessive crosstalk between PCM carrying pairs. The effect of high crosstalk levels is to introduce digital errors in the PCM bit stream. Error rate may be related on a statistical basis to crosstalk, which in turn is dependent on the characteristics of the cable and the number of PCM carrying pairs.

One method to reduce crosstalk and thereby increase VF pair usage is to turn to two-cable working, rather than have the "go" and "return" PCM cable pairs in the same cable.

Another item that can limit cable pair usage is the incompatibility of FDM and PCM carrier systems in the same cable. On the cable pairs that will be used for PCM, the following should be taken into consideration:

- All load coils must be removed.
- Build-out networks, bridged taps must also be removed.
- No crosses, grounds, splits, high resistance splices, nor moisture permitted.

The frequency response of the pair should be measured out to 1 MHz and taken into consideration as far out as 2.5 MHz. Insulation should be checked with a megger. A pulse reflection test using a radar test set is also recommended. Such a test will indicate opens, shorts, and high impedance mismatches. A resistance test and balance test using a Wheatstone bridge may also be in order. Some special PCM test sets are available such as the Lenkurt Electric 91100 PCM Cable Test Set using pseudo random PCM test signals and the conventional digital test eye pattern.

Practical System Block Diagram

A block diagram showing the elemental blocks of a PCM transmission link used to expand installed VF

*Exhaust is an outside plant term meaning that the useful pairs of a cable have been used up (assigned) from a planning point of view.

The Language and Techniques of Pulse Code Modulation (PCM)

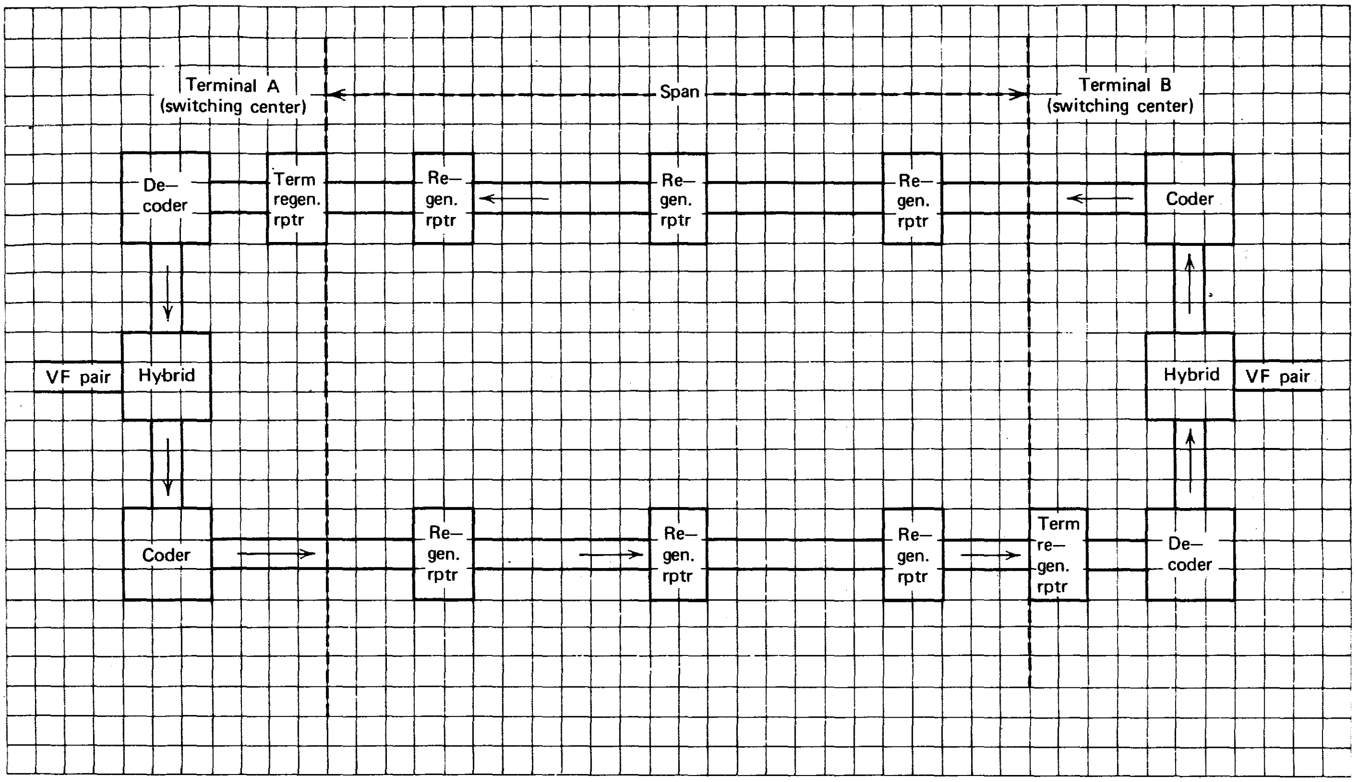


Figure 13. Simplified functional block diagram of a PCM link used to expand capacity of an existing VF cable (for simplicity, interface with only one VF pair is shown). Note the spacing between repeaters in the span line

cable capacity is shown in Figure 13. Most telephone administrations (companies) distinguish between the terminal area of a PCM system and the repeatered line. The term "span" comes into play here. A span line is composed of a number of repeater sections permanently connected in tandem at repeater apparatus cases mounted in manholes or on pole lines along the span. A "span" is defined as the group of span lines that extend between two office (switching center) repeater points.

A typical span is shown in Figure 13. The spacing between regenerative repeaters is important. The necessity of removing load coils from those trunk (junction) cable pairs that were to be used for PCM transmission was mentioned above. It is at these load points that the PCM regenerative repeaters are installed. On a VF line with H-type loading, spacing between load points is normally about 6000 ft (1830 km). The first load coil out from the exchange on a trunk repair is at half-distance or 3000 ft (915 m). This is provident for a regenerative repeater also must be installed at this point. This spacing is shown in Figure 13 (1 space = 1000 ft). The purpose of installing a repeater at this location is to increase the pulse level before entering the environment of an exchange area where the levels of impulse noise may be quite high. High levels of impulse noise induced into the system may cause significant increases in digital error rate of the incoming PCM bit streams, particularly when the bit stream is of a comparatively low level.

Commonly PCM pulse amplitude output of a regenerative repeater is on the order of 3 V. Likewise, 3 V is the voltage on the PCM line cross-connect field at the exchange (terminal area).

A guideline used by Bell Telephone Manufacturing Company (Belgium) is that the maximum distance separating regenerative repeaters is that corresponding to a cable pair attenuation of 36 dB at the maximum expected temperature at 1024 kHz. This frequency is equivalent to the half-bit rate for the CEPT systems (e.g., 2048 kb/s). Actually repeater design permits operation on lines with attenuations anywhere from 4 to 36 dB, allowing considerable leeway in placing repeater points. Table 1 gives some other practical repeater spacing parameters for the CEPT-ITT-BTM 30 + 2 system.

The maximum distance is limited by the maximum number of repeaters, which in this case is a function of power feeding and supervisory considerations. For instance, the fault location system can handle up to a maximum of 18 tandem repeaters for the BTM (ITT) configuration.

Power for the BTM system is via a constant-current feeding arrangement over a phantom pair serving both

The Language and Techniques of Pulse Code Modulation (PCM)

Pair Diameter (mm)	Loop Atten. at 1 MHz (dB/km)	Loop Resistance (Ω/km)	Voltage Drop (V/km)	Maximum Distance[a] (km)	Total Repeaters	Maximum Distance System (km)
0.9	12	60	1.5	3	18	54
0.6	16	100	2.6	2.25	16	36

[a] Between adjacent repeaters.

Table 1. Line parameters for ITT/BTM PCM

the "go" and related "return" repeaters, providing up to 150 V dc at the power feed point. The voltage drop per regenerative repeater is 5.1 V. Thus for a "go" and "return" repeater configuration the drop is 10.2 V.

As an example, let us determine the maximum number of regenerative repeaters in tandem that may be fed from one power feed point by this system using 0.8-mm diameter pairs with a 3-V voltage drop in an 1830-m spacing between adjacent repeaters:

$$\frac{150}{(10.2 + 3)} = 11$$

Assuming power fed from both ends and an 1800-m "dead" section in the middle, the maximum distance between power feed points is approximately

$$(2 \times 11 + 1)\ 1.8 \text{ km} = 41.4 \text{ km}$$

Fault tracing for the North American (ATT) T1 system is carried out by means of monitoring the framing signal, the 193rd bit. The framing signal (amplified) normally holds a relay closed when the system is operative. With loss of framing signal, the relay opens actuating alarms. By this means a faulty system is identified, isolated, and dropped from "traffic."

To locate a defective regenerator on the BTM (Belgium)-CEPT system, traffic is removed from the system, and a special pattern generator is connected to the line. The pattern generator transmits a digital pattern with the same bit rate as the 30 + 2 PCM signal, but the test pattern can be varied to contain selected low frequency spectral elements. Each regenerator on the repeatered line is equipped with a special audio filter, each with a distinctive passband. Up to 18 different filters may be provided in a system. The filter is bridged across the output of the regenerator, sampling the output pattern. The output of the filter is amplified and transformer-coupled to a fault transmission pair, which is normally common to all PCM systems on the route, span, or section.

To determine which regenerator is faulty, the special test pattern is tuned over the spectrum of interest. As the pattern is tuned through the frequency of the distinct filter of each operative repeater, a return signal will derive from the fault transmission pair at a minimum specified level. Defective repeaters will be identified by absence of return signal or a return level under specification. The distinctive spectral content of the return signal is indicative of the regenerator undergoing test.

The Line Code

PCM signals as transmitted to the cable are in the bipolar mode (biternary), as shown in Figure 1. The marks or "1"s have only a 50% duty cycle. There are several advantages to this mode of transmission:

- No dc return is required; thus transformer coupling can be used on the line.
- The power spectrum of the transmitted signal is centered at a frequency equivalent to half the bit rate.

It will be noted in bipolar transmission that the "0"s are coded as absence of pulses and the "1"s are alternately coded as positive and negative pulses with the alternation taking place at every occurrence of a "1." This mode of transmission is also called alternate mark inversion (AMI).

One drawback to straightforward AMI transmission is that when a long string of "0"s is transmitted (e.g., no transitions), a timing problem may come about because repeaters and decoders have no way of extracting timing without transitions. The problem can be alleviated for forbidding long strings of "0"s. Codes have been developed which are bipolar but with N zeros substitution; they are called BNZS codes. For instance, a B6ZS code substitutes a particular signal for a string of 6 "0"s.

Another such code is the HDB3 code (high density binary 3), where the 3 indicates that it substitutes for binary formations with more than 3 consecutive "0"s. With HDB3, the second and third zeros of the string are transmitted unchanged. The fourth "0" is transmitted to the line with the same polarity as the previous mark sent, which is a "violation" of the AMI concept. The first "0" may or may not be modified to a "1" to assure that the successive violations are of opposite polarity.

Signal-to-Gaussian-Noise Ratio on PCM Repeater Lines

As we mentioned earlier, noise accumulation on PCM systems is not an important consideration. This does not mean that Gaussian noise (nor crosstalk, impulse noise) is not important. Indeed, it may affect error performance expressed as error rate. Error rate, from

The Language and Techniques of Pulse Code Modulation (PCM)

one point of view, is cumulative. A decision in error, whether "1" or "0," made anywhere in the digital system, is not recoverable. Thus such an incorrect decision made by one regenerative repeater adds to the existing error rate on the line, and errors taking place in subsequent repeaters further down the line add in a cumulative manner tending to deteriorate the received signal.

In a purely binary transmission system, if a 20-dB signal-to-noise ratio is maintained, the system operates nearly error free. In this respect, consider Table 2.

Error Rate	S/N(dB)	Error Rate	S/N(dB)
10^{-2}	13.5	10^{-7}	20.3
10^{-3}	16	10^{-8}	21
10^{-4}	17.5	10^{-9}	21.6
10^{-5}	18.7	10^{-10}	22
10^{-6}	19.6	10^{-11}	22.2

Table 2. Error rate of a binary transmission system versus signal-to-rms noise ratio

PCM, in practice, is transmitted on-line with alternate mark inversion. The marks have a 50% duty cycle permitting energy concentration at a frequency of half the transmitted bit rate. Thus it is advisable to add 1 or 2 dB to the values shown in Table 1 to achieve a desired error rate on a practical system.

The Eye Pattern

The "eye" pattern provides a convenient method of checking the quality of a digital transmission line. A sketch of a typical eye pattern is shown in Figure 14. Any oscilloscope can produce a suitable eye pattern provided it has the proper rise time. Most quality oscilloscopes now available on the market do have. The scope should either terminate or bridge the repeatered line or output of a terminal repeater. The display on the oscilloscope contains all the incoming bipolar pulses superimposed one on the other.

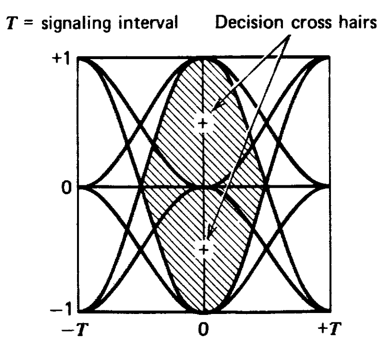


Figure 14. Sketch of an "eye" pattern

*Oscilloscopes are commonly used to measure voltage: thus we can measure the degraded opening in voltage units and compare it to the full-scale perfect opening in the same units. A ratio is developed and we take 20 log that ratio to determine S/N.

Eye patterns are indicative of decision levels. The greater the eye opens, the better defined is the decision (whether "1" or "0" in the case of PCM). The opening is often referred to as the decision area (lightly cross-hatched in the figure). Degradations reduce the area. Eye patterns are often measured off in the vertical, giving a relative measure of margin of decision.

Amplitude degradations shrink the eye in the vertical. Among amplitude degradations can be included echos, intersymbol interference, and decision threshold uncertainties.

Horizontal shrinkage of the eye pattern is indicative of timing degradations (i.e., jitter and decision time misalignment).

Noise is the other degradation to be considered. Usually noise may be expressed in terms of some improvement in signal-to-noise ratio to bring the operating system into the bounds of some desired objective; see Table 2, for example. This ratio may be expressed as 20 x log* of the ideal eye opening (in the vertical as read on the oscilloscope's vertical scale) to the degraded reading.

HIGHER-ORDER PCM MULTIPLEX SYSTEMS

Using the 24-channel D2 channel bank as a basic building block, higher order PCM systems are being developed in North America. For instance, four D-2 channel banks are multiplexed by a M1-2 multiplexer placing 6.312 Mb/s on a single wire pair (T2 digital line). Figure 15a is a simplified block diagram in the first step in the development of a higher order PCM system.

Two major elements make up a higher order system: multiplexers and repeatered line. ATT identifies its repeatered lines as follows:

T1 (deriving from D1 or D2 channel banks)	1.544 Mb/s
T2 (output of multiplexer M1-2)	6.312 Mb/s
T3 (output of multiplexer M2-3)	46.304 Mb/s

Figure 15b shows this hierarchy diagrammatically.

The higher order system will also accept data, video (TV), videotelephone, and FDM mastergroups.

To transmit data on the ATT system, sampling is made of the binary serial data transitions, using 3 PCM bits to code each expected transition of the data bit stream. The 3 PCM bits do the coding as follows: for no transition, a series of three "1"s are transmitted. If a transition took place, then

The Language and Techniques of Pulse Code Modulation (PCM)

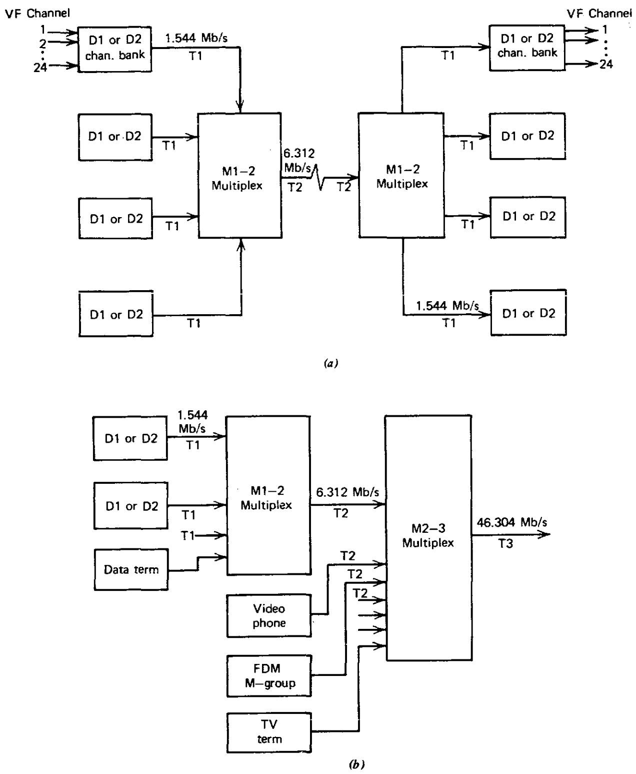


Figure 15 (a) Development of the 96-channel T2 (ATT) system by multiplexing the 24-channel D1 or D2 channel bank outputs. (b) Development of higher order PCM (ATT plan)

Bit number		Function
1	("0")	Indicates presence of transition
2	("1" or "0")	Indicates in which half of the sampling interval the transition took place
3	("1" or "0")	Indicates the direction of transition, positive or negative going (e.g., "1" to "0" or "0" to "1")

The use of 3 bits for each transition, rather than only 1 bit to indicate just the presence or absence of a transition, prevents ambiguity in case of error and reduces error rate to a degree.

Pulse "stuffing" is an important concept in multiplexers such as the M1-2 which have outputs in the megabit range. With stuffing, the multiplexer output in bits per second is always at a higher bit rate than the sum of the lower speed inputs. This is done by artificially adding bits in the multiplexer (and withdrawing them at the demultiplexer). Stuffing is this artificial addition of bits.

Pulse stuffing simplifies clock design, clocking being the heart of any time division multiplex device. It also simplifies synchronization problems and reduces necessary buffer storage. The problem is not really in the clocking itself but in the variation of propagation time of the medium interconnecting the multiplexer to a common demultiplexer (see Figure 15). The variation in propagation time causes the slowing or speed-up of the arrival, the arrival bit rate increasing and decreasing with the variation of propagation time. The receiving equipment clocks the incoming bit stream at a constant rate; therefore storage is required.

With stuffing the demulitplexer is informed of the number and location of the stuffed bits. This information is usually passed on a separate data line with the output information. With outputs operating at a speed greater than the inputs, no input information bits can be lost when the storage is exceeded or when storage units are reset.

Future systems will probably be synchronized using master clocks with each terminal area, and each terminal will have the capability of operating independently for periods of time in case of master clock failure. To compensate for slow phase variations due to variation of propagation time, elastic stores will be provided at each terminal.

LONG-DISTANCE (TOLL) TRANSMISSION BY PCM

PCM, with its capability of regeneration, essentially eliminating the accumulation of noise as a signal traverses its transmission media, would appear to be the choice for toll transmission or backbone, long-haul routes. This has not been the case. One must consider the disadvantages of PCM as well. Most important is the competition with FDM systems, the L5 system, for instance. ATT's L5 provides 10,800 VF channel capacity over long-haul coaxial cable media. The required bandwidth for this capacity on the cable is 60 MHz. To transmit the same number of channels by PCM would require on the order of 16 times the bandwidth.

Keep in mind the relationship briefly covered earlier wherein this 16-multiple concept is shown: a 4-kHz voice channel requires an equivalent PCM bandwidth of 64 kHz, assuming 1-Hz bandwidth per PCM bit transmitted.

Thus a 10,800 VF (4-kHz) channel would require, if transmitted by PCM, about 691.2 MHz. Available bandwidth is still at a premium, whether by cable or radio. Therefore, it is more economical to transmit such large packages of information by FDM techniques and suffer the consequent noise accumulation. Perhaps when waveguide, millimetric satellite communication, or optical fiber techniques become viable, then PCM will prove useful for high density, long-haul transmission.

The Language and Techniques of Pulse Code Modulation (PCM)

Jitter

There is one other important limitation of present-day technology on using PCM as a vehicle for long-haul transmission. This is jitter, more particularly, timing jitter.

A general definition of jitter is "the movement of zero crossings of a signal (digital or analog) from their expected time of occurrence." It is unwanted phase modulation or incidental FM. Such jitter or phase jitter affected the decision process of the zero crossing in a digital data modem. Much of this sort of jitter can be traced to the intervening FDM equipment between one end of a data circuit and the other.

PCM has no intervening FDM equipment, and jitter in PCM systems takes on different characteristics. However, essentially the effect is the same—uncertainty in a decision circuit as to when a zero crossing (transition) took place, or the shifting of a zero crossing from its proper location. In PCM it is more proper to refer to jitter as timing jitter.

The primary source of timing jitter is the regenerative repeater. In the repeatered line jitter may be systematic or nonsystematic. Systematic jitter may be caused by offset pulses (i.e., where the pulse peak does not coincide with regenerator timing peaks, or transitions are offset), intersymbol interference (dependent on specific pulse patterns), and local clock threshold offset. Nonsystematic jitter may be traced to timing variations from repeater to repeater and to crosstalk.

In long chains of regenerative repeaters, systematic jitter is predominant and cumulative, increasing in rms value as $N^{1/2}$, where N is the number of repeaters in the chain. Jitter is also proportional to a repeater's timing filter bandwidth. Increasing the Q of these filters tends to reduce jitter of the regenerated signal, but it also increases error rate due to sampling the incoming signal at nonoptimum times.

The principal effect of jitter on the resulting analog signal after decoding is to distort the signal. The analog signal derives from a PAM pulse train, which is then passed through a low-pass filter. Jitter displaces the PAM pulses from their proper location, showing up as undesired pulse position modulation (PPM).

Because jitter varies with the number of repeaters in tandem, it is one of the major restricting parameters of long-haul, high bit rate PCM systems. Jitter can be reduced in future systems by using elastic store at each regenerative repeater (costly) and high-Q phase locked loops.

PCM TRANSMISSION BY RADIOLINK

The transmission of PCM by radiolink is a viable alternative to PCM VF cable pair transmission under the following circumstances:

- Where physical or natural obstructions increase cable cost.
- On long spans, more than 30 mi (50 km) of relatively low VF pair capacity, up to 1200 circuits.
- As alternative routing of a cable system via radio.

Consider a situation where a large number of trunk (junction) routes are presently equipped with PCM. An FDM/FM radiolink is contemplated, and the system engineer is faced with one or several of the above circumstances. To use FM radiolinks with FDM multiplex will prove expensive. The existing PCM will have to be brought to VF (demultiplexed-demodulated) to interface with the new FDM equipment.

Use of PCM eliminates the additional multiplex equipment cost. Further, PCM channelizing equipment, if we accept groups of 24 or 30 channels at a time, is less expensive on a per-channel basis than FDM equipment.

Modulation, RF Bandwidth, and Performance

The most common modulation techniques for PCM over radio are straightforward FSK (frequency shift keying), theoretically requiring about 1 b/Hz of bandwidth, and QPSK (quadrature phase shift keying), requiring 1 Hz of bandwidth for every 2 bits transmitted at the expense of 6 dB improved signal-to-noise ratio when compared to FSK for equal error rates. As in many FM systems, it is usual to modulate the oscillator in the PCM radio transmitter, convert the oscillator output to 70 MHz, the IF, and heterodyne the IF to the output frequency in the 2-, 4-, 6—8-, 11-, and 15-GHz bands.

Care must be taken with these systems that they do not cause interference to existing FM systems. Special filtering of the RF output of the transmitter is often called for to reduce spurious and to limit sidebands. Increasing the transmission level to 8 (e.g., m = 8) or 16 increases bit rate for a given bandwidth resulting in further tightening of signal-to-noise ratio requirements.

Practical systems, however, assign bandwidths of 1 b/Hz in the best of cases. For instance, one system proposed in the United States will carry 21.5 Mb/s in a 30-MHz bandwidth at 6 GHz. Canadian Marconi describes a system as eight-level FM with a bandwidth of 0.44 x bit rate. One way to reduce "necessary" bandwidths is to reduce guard bands and permit some overlapping of power spectra. Otherwise for practical purposes eight-level systems produce 1.55 b/Hz of bandwidth.

One of the major advantages of PCM (digital) radio, as with PCM cable systems, is the ability the system has to periodically regenerate the waveform at each repeater site. Another advantage is that it is little affected by traffic loading, and any mix of voice and data traffic has no effect on system performance.

The disadvantage of PCM radio is that, at present, no more than the equivalent of 120 VF channels load an RF carrier. Again we are up against the fact that a 4-kHz channel is the input for voice information to analog multiplex, and for PCM the same voice channel has an equivalent bandwidth of 64 kHz. One proposed system, on the other hand, will transmit the equivalent of six T2 lines (6 x 96 voice channels) or 596 equivalent voice channels for a line bit rate of 39.6 MB/s requiring a 40-MHz bandwidth of RF. Equivalent FDM/FM systems operate with 1200, 1800, and up to 2700 voice channels on one RF carrier requiring often less bandwidth.

THE ADVANTAGES OF PCM SWITCHING

The implementation of PCM switching will have a profound impact on PCM transmission in the local area. PCM switching is a form of time division switching where the switching inlets and outlets carry PCM highways in the case of tandem switches. Local switches, depending on the penetration of PCM, whether to the level of subscriber line or not, may be implemented by (1) Subscriber lines to the switch analog with PCM conversion at the switch, or (2) subscriber lines PCM with direct conversion to PCM at the subscriber location. In either case the trunks (junctions) would be PCM.

A PCM switch would operate far differently from an analog switch. Crosspoint usage would be more efficient for they would be operated on a time division mode, rather than the present space division. Call routing would be kept in memory, and various crosspoints to the same circuit group would be employed on one call at different time intervals.

Signaling fits well into this arrangement. For analog, space division switches, signaling is digital, whether it is subscriber, line, or interregister. Signaling, like data transmission, must be conditioned to operate over an analog network in most cases, either in the switching equipment or in the transmission equipment. Given a network where both transmission and switching are digital (e.g., PCM), signaling can take on its own characteristic and remain digital. This reduces circuit cost and complexity.

Digital switching, namely PCM, eliminates switching losses and switching loss variation. Inside the integrated digital network circuit stability may be disregarded as a design limitation. In hybrid networks where both digital and analog systems coexist, stability must be considered at the interface, but only to meet the requirements of the analog side.

Also, by definition, the digital network is four-wire. Reference equivalent can be improved by assigning zero loss to that portion of the network that is digital (subject to overall stability considerations).

Consider a tandem switch that is PCM. The choice is to convert to the PCM mode at the switch or convert in the outlying local exchanges and use PCM as the trunk (junction) transmission media. In most cases the latter would seem desirable. Certainly the prove-in distance is reduced.

With a fully digital network, switching and transmission will be much more integrated technologies. Much more equipment will be common to both with concurrent savings of outlay in both disciplines.

RELATED READINGS

Data Handbook for Radio Engineers, 5th ed., Howard W. Sams & Co., Indianapolis, Section 21.

Transmission Systems for Communications, 4th ed., Bell Telephone Laboratories.

Lenkurt Demodulator, Lenkurt Electric Co., San Carlos, California: November 1966, March, April 1968, September 1971, December 1971, January 1972, October 1973.

Seminar on Pulse Code Modulation Transmission, Standard Telephone and Cables Limited, Basildon, England, June 1967.

PCM—System Application—30 + 2*TS*, BTM/ITT, Antwerp, Belgium.

Technical Manual-Operations and Maintenance Manual for T324 PCM Cable Carrier System, ITT Telecommunications, Raleigh, North Carolina, April 1973.

Principles of Modems (edited draft), Communication Systems, Inc., Falls Church, Virginia (limited circulation).

K. W. Catern ole, *Principles of Pulse Code Modulation*, Iliffe, London, 1969.

W. C. Sain, "Pulse Code Modulation Systems in North America," *Electrical Communication* (ITT), 48, (1 and 2), 1973.

J. V. Marten and E. Brading, 30-channel Pulse Code Modulation System, *Electrical Communication* (ITT), 48, (1 and 2), 1973.

K. E. Fultz and D. B. Penick. "The T1 Carrier System," *Bell System Tech. J.*, 44, September 1965.

The Language and Techniques of Pulse Code Modulation (PCM)

R. B. Moore, "T2 Digital Line System," *Proceedings IEEE International Conference on Communications*, June 12, 1973, Seattle, Washington.

J. R. Davis, "T2 Repeater and Equalization," *Proceedings IEEE International Conference on Communications*, June 12, 1973, Seattle, Washington.

E. Cookson and C. Volkland, "Taking the Mystery Out of Phase Jitter Measurement," *Telephony*, September 25, 1972.

"Phase Jitter and its Measurement," CCITT Study Group IV, Question 3/IV, 17-28, January 1972.

F. S. Boxal, "Digital Transmission via Microwave Radio," Parts I and II, *Telecommunications*, April and May 1972.

W. H. Smith, "PCM Microwave Links," *Telecommunications*, April 1973.

L. Katzschner, et al., "An Experimental Local PCM-Switching System," *IEEE Trans. Communications*, October 1973.

A. E. Pinet, "Telecommunication Integrated Network," *IEEE Trans. Communications*, August 1973.

G. C. Hartley, et al., *Techniques of Pulse-Code Modulation in Communication Networks*, IEE Monograph Series, Cambridge University Press, 1967.□

The Nature and Organization of Digital Transmission Channels

Problem:

An all-digital communications system consists of three essential parts: a generally acceptable data encoding format, digital-compatible transmission channels, and a control structure that takes full advantage of the switching and routing efficiencies offered by all-digital systems. The emerging data format appears to be PCM (described earlier in this section). The emerging control structure appears to be packet switching with its accompanying multilayered protocols (for example, ARPANET and IBM's SNA, explained later in this service). This report defines the transmission channels. It explains how the channels are structured to handle PCM-encoded data, and it describes how they are organized into controllable patterns. This information is essential to help you think through the transition from voice-grade telephone lines to the newer digital superhighways.

Solution:

The following new telecommunications channels operate digitally at 274 million bits per second:

1. AT&T's new coaxial cable system, the T4M system, contains 20 coaxial tubes giving 10 full duplex channels of 274 mb/s. A few segments of this are now in operation.

2. AT&T's new 18 GHz radio system, the DR18 system, transmits 7 channels of 274 mb/s in each direction. A few segments of this are now in operation.

3. AT&T's helical waveguide, the WT4 system, which modulates 60 sinewave carriers at 274 mb/s to give 60 channels of 274 mb/s each. A few segments of this are now in operation.

4. A glass fiber system powered by a semiconductor laser. Experimental repeaters have been built in Bell Laboratories to relay 274 mb/s through fibers. The repeaters on such a system could be 10 kilometers apart. Hundreds or thousands of such fibers could occupy one cable.

5. Satellite transponders could be designed to operate at a wider bandwidth than most of today's transponders and could relay bit rates such as 274 mb/s.

In the AT&T (and hence the United States) telephone network there are four main levels of digital carrier rate, of which 274 mb/s is the highest. A variety of different physical channels and multiplexers are built to conform to these four basic speeds. They are referred to as the T1, T2, T3, and T4 carriers:

- T1 carrier: 1.544 mb/s
- T2 carrier: 6.312 mb/s
- T3 carrier: 45 mb/s
- T4 carrier: 274 mb/s

The T1 carrier was designed in the 1960s to operate over twisted wire pair cables and to carry 24 voice

---

Future Developments in Telecommunications (2nd Edition) Chapter 27, pp 501-521 by James Martin.

The Nature and Organization of Digital Transmission Channels

channels. The T2 carrier came into use in 1972 and can operate over good quality wire pair trunks to carry 96 voice channels or one Picturephone channel. The T3 carrier can carry one digitized mastergroup (600 voice choice channels). It is too fast for wire pair cables and too slow for coaxial. Its bit rate could be sent over one of today's 36 MHz satellite transponders, and experimental optical fibers have been designed for it. It is possible that no major T3 transmission system will come into use. It will serve merely as a bridge between the T1 or T2 and the T4 carriers.

The bit streams of the smaller digital carriers can be fed into the larger carriers by means of multiplexers. Thus four T1 streams can be fed into a T2 stream by means of an M12 multiplexer. Figure 1 shows the digital levels and multiplexer links between them. The multiplexers are designated Mxy where x designates the lower level and y the higher level.

As indicated in Figure 1, an additional level (T1C) has been introduced above level 1 to double the capacity of T1 routes. The T1C carrier transmits 3.2 mb/s over wire pair links and is designed to be similar to the T1 carrier for ease of upgrading T1 trunk routes.

LINKS IN OTHER NATIONS

Many other nations are installing digital transmission facilities. Few, however, use the same bit rates and digital structures as the United States. A proliferation of incompatible digital systems has been installed.

International recommendations for standards are emerging. CCITT, the international organization for agreement on telecommunications standards, has recommended detailed specifications for two basic PCM systems—an international equivalent of the Bell System Level 1. One uses 1.544 mb/s to transmit 24 voice channels, and the other uses 2.048 mb/s to transmit 30 voice channels. The 1.544 mb/s recommendation is not identical to the Bell T1 carrier.

CEPT, the European authority (European Conference of Postal and Telecommunications Administrations) has recommended that the multiplexing should go up in steps of 4, giving a family of digital systems as follows:

Level	Millions of bits/second	Number of voice channels
1	2.048	30
2	8.448	120
3	34.304	480
4	139.264	1920
5	565.148	7680

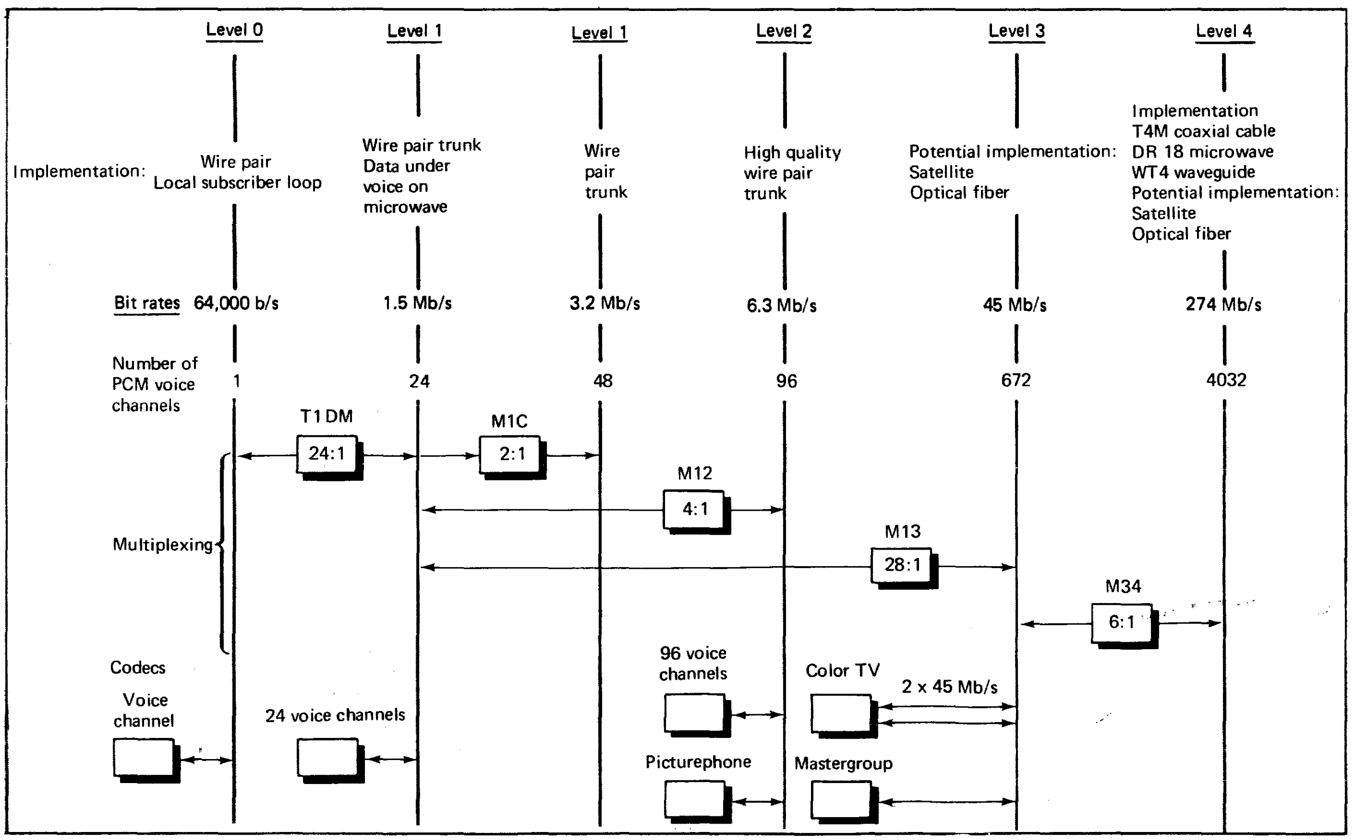


Figure 1. The hierarchy of digital channels on the Bell System

The Nature and Organization of Digital Transmission Channels

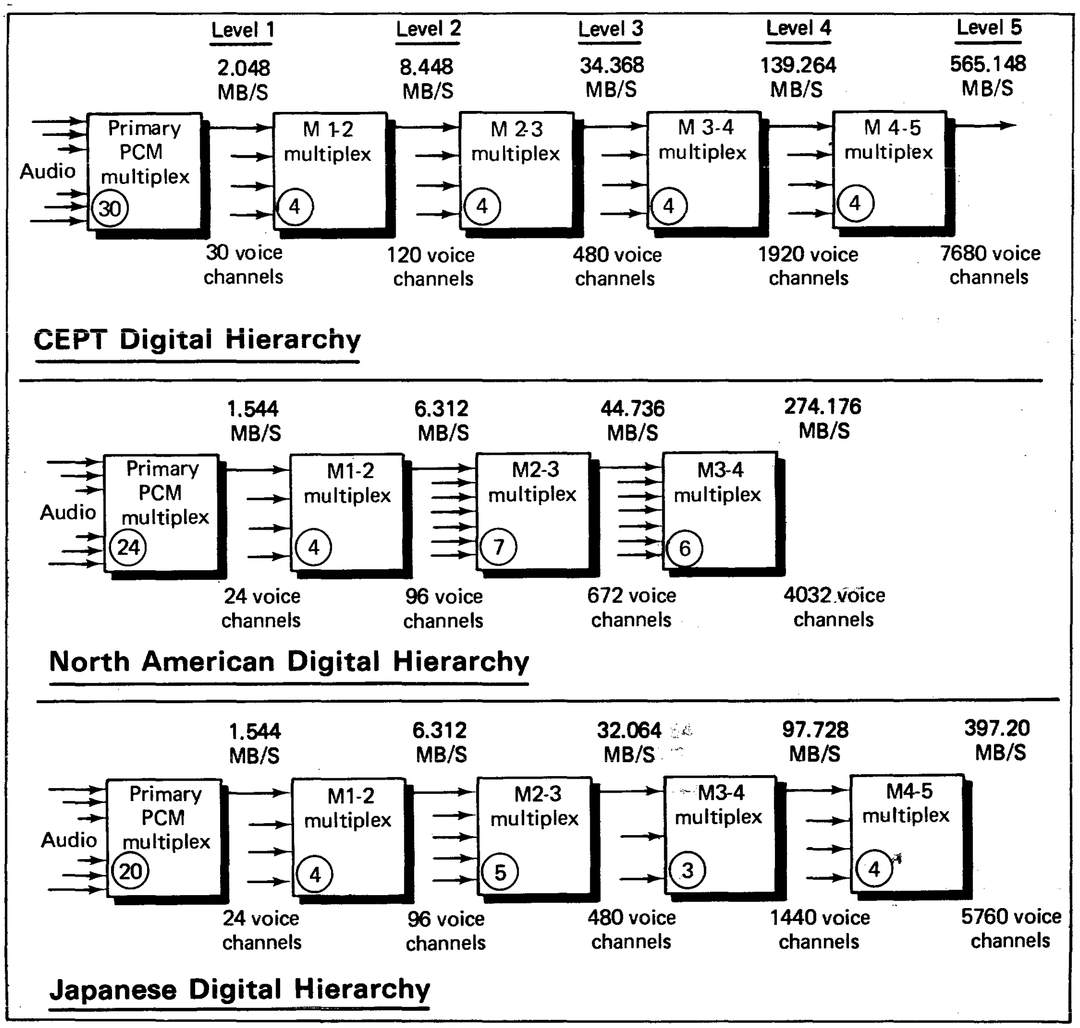


Figure 2. Summary of major international digital hierarchies

Figure 2 summarizes the digital hierarchies of North America and Japan, and that recommended by CEPT.

The higher bit rates are not exact multiples of the lower ones because extra bits are needed for synchronization, control, and identification.

Many countries outside North America now have digital wire-pair systems transmitting 1.544 and 2.048 mb/s. Some countries have introduced microwave systems operating at 13 GHz, with PSK modulation at 8.448, 34.304, and between these at 17.152 mb/s. Prototype coaxial cable systems operating at 34.304 and 139.264 mb/s have been developed. Italy uses a small coaxial cable carrying 8.448 mb/s, called micro-coaxial.

THE BELL SYSTEM T1 CARRIER

The most common use of pulse code modulation at the time of writing is in the Bell System T1 carrier. The Bell T1 PCM System multiplexes together 24 voice channels. Seven bits are used for coding each sample. The system is designed to transmit voice frequencies up to 4 KHz, and therefore 8000 samples per second are needed, 8000 frames per second travel down the line. Each frame, then, takes 125 microseconds. A frame is illustrated in Figure 3. It contains eight bits for each channel. The eighth is used for supervisory reasons and signaling, for example, to establish a connection and to terminate a call. There are a total of 193 bits in each frame, and so the T1 line operates at 193 x 8000 = 1,544,000 bits per second.

The last bit in the frame, the 193rd bit, is used for establishing and maintaining synchronization. The sequence of these 193 bits from separate frames is established by the logic of the receiving terminal. If the sequence does not follow a given coded pattern, then the terminal detects that synchronization has been lost. If synchronization slips, then the bits examined will in fact be bits from the channels—

The Nature and Organization of Digital Transmission Channels

probably speech bits—and will not exhibit the required pattern. There is a chance that these bits will form a pattern similar to the pattern being sought. The synchronization pattern must therefore be chosen so that it is unlikely that it will occur by chance. If the 193rd bit were made to be always a one or always a zero, a zero or a one could occur by chance in the voice signal. It was found that an alternating bit pattern, 0 1 0 1 0 1 . . . never occurs for long in any bit position. Such a pattern would imply a 4-KHz component in the signal, and the input filters used would not pass this. Therefore, the 193rd bit transmitted is made alternately a one and a zero. The receiving terminal inspects it to ensure that this 1 0 1 0 1 0 . . . pattern is present. If it is not, then it examines the other bit positions that are 193 bits apart until a 1 0 1 0 1 0 . . . pattern is found. It then assumes that these are the framing pulses.

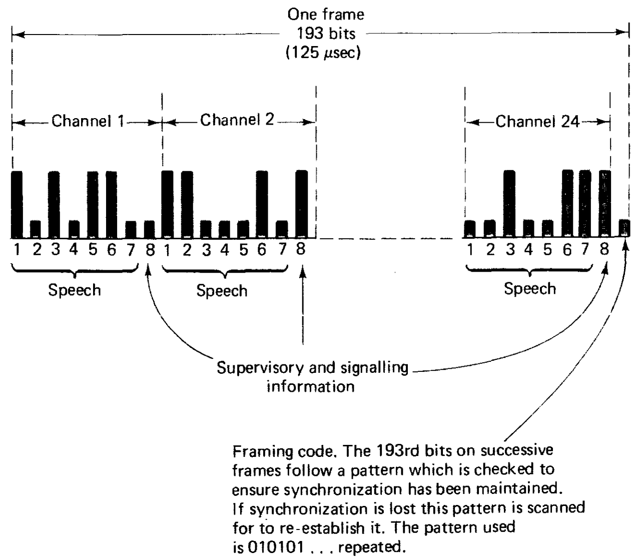


Figure 3. The bit structure of a North American PCM transmission link operating at 1.544 million bits per second (T1 carrier). The frame shown here is repeated 8000 times per second, thus giving 8000 samples per second on each of 24 channels, plus an 8000 bps bit stream for control signaling. The CCITT Recommendation for 1.544 mbps PCM is slightly different (see Figure 5)

This scheme works very well with speech transmission. If synchronization is lost, the framing circuit takes 0.4 to 6 milliseconds to detect the loss. The time required to reframe will be about 50 milliseconds at worst if all the other 192 positions are examined; but normally the time will be much less, depending on how far the transmission is out of synchronization. This is quite acceptable on a speech channel. It is more of a nuisance when data is sent over the channel and would necessitate the retransmission of blocks of data. Retransmission is required on most data transmission links, however, as a means of correcting errors that are caused by noise on the line and detected with error-detecting codes.

The permissible signal levels are not equally spaced in PCM encoding. The levels are bunched closer together at the lower signal amplitudes than at the higher ones. This gives better reproduction of low volume speech.

REGENERATIVE REPEATERS

The main reason why high bit rates can be achieved on wire-pair circuits using pulse code modulation is that repeaters are placed at frequent intervals to reconstruct the signal.

In most PCM systems working today, the repeaters are placed at intervals of between 1 and 5 kilometers. The Bell T1 System, operational since 1962, uses repeaters at intervals of 1.8 kilometers, typically, which is the spacing of loading coils employed when the wires were used for analog transmission; the repeaters replace the loading coils. These repeaters reconstruct 1,544,000 pulses per second.

A regenerative repeater has to perform three functions, sometimes referred to as the 3 "Rs": reshaping, retiming, and regeneration. When a pulse arrives at the repeater, it is attenuated and distorted. It must first pass through a preamplifier and equalizer to reshape it for the detection process. A filter removes the DC component. A timing recovery circuit provides a signal to sample the pulse at the optimum point to decide whether it is a one or a zero bit. The timing circuit controls the regeneration of the outgoing pulse and ensures that it is sent at the correct time and is of the correct width.

As shown in Figure 4 the pulses transmitted occupy half a time slot. A pulse represents a 1 bit, and absence of a pulse denotes a 0 bit. Each 1 transmitted has an opposite polarity to the previous 1. This concentrates the energy of the signal around 772 MHz rather than 1.544 MHz when a string of ones is transmitted. To avoid the DC component when a string of zeros is transmitted, a special code is substituted for each group of six zeros.

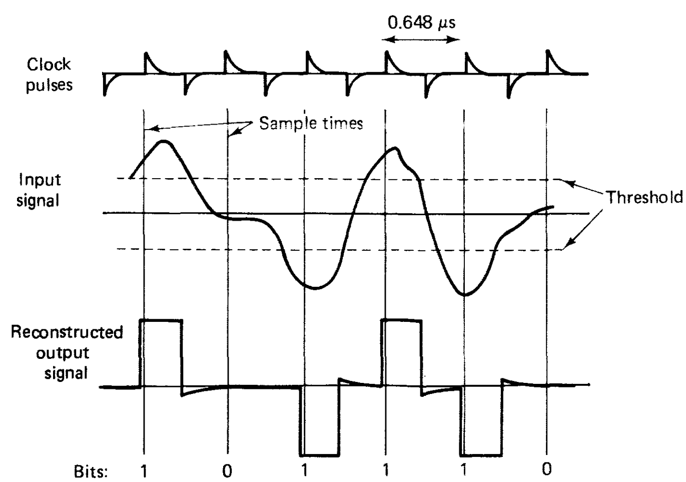


Figure 4. The repeaters, every 1.8 kilometers on a T1 circuit, reshape, retime, and regenerate the pulse stream

The Nature and Organization of Digital Transmission Channels

The T1 carrier uses a 3-level signal, as shown in Figure 4, called a bipolar signal. Higher levels of digital carrier use a binary signal—a positive pulse for a 1 and no pulse for a 0. A binary signal is more efficient; a higher bit rate can be transmitted over a given bandwidth. The T1 carrier uses bipolar signaling to reduce margin against carrier in the decision circuit.

PCM cables usually transmit the power for the repeaters down the cable itself. A filter at each repeater separates the message signal from this DC current.

CROSSTALK

City and short-haul trunk cables often contain many hundreds of wire pairs. If all of these wire pairs transmit in the same direction, they can all carry the 1.5 mb/s of the T1 carrier or the 3.2 mb/s of the T1C carrier, with the exception of a number of wire pairs reserved for fault location and order wires for communicating between manholes and central offices. A large wire-pair cable can thus be given a digital capacity of billions of bits per second. If, however, different wire pairs transmit in opposite directions, there is a danger of crosstalk. The strong signal leaving a repeater may be next to an arriving weak signal, which has been attenuated by the length of the cable. The strong signal can interfere with the weak one. This problem is dealt with in one of three ways. First the cable may contain only a small proportion of digital wire-pairs, and these are separated by normal analog telephone wire-pairs. Second the cable may be partitioned, an electrical shield separating the wires for each direction of transmission. Third, and best if it is possible, two cables may be used, one for each direction of transmission.

CCITT RECOMMENDATIONS

As mentioned, the CCITT has made recommendations for PCM systems for transmission at the T1 carrier speed of 1.544 mb/s, and for transmission at 2.045 mb/s.

Figure 5 shows the CCITT recommendations for transmission at 1.544 mb/s. As is often the case, the CCITT recommendation is slightly different to the North American standard set by AT&T. Like AT&T, it employs a 193-bit frame with 8 bits per channel as in Figure 3, but the frame alignment bit is the first bit, not the 193rd bit as in Figure 3, and it carries a different synchronization pattern. Twelve such frames are grouped together to form one multiframe.

There are two versions of it. One has a common signaling channel associated with the block of 24 voice channels. The other has signaling associated with each voice channel. The common-channel signaling scheme is shown at the top of Figure 5. The first bit of each frame serves two purposes. In odd frames it is used for maintaining synchronization, carrying a 1 0 1 0 1 0 . . . pattern in successive frames (i.e., 1 in frame number 1, 0 in frame number 3, 1 in frame number 5 . . . and so on). In even frames it gives a bit stream (4000 bits per second because there are 8000 frames per second) that carries signaling information of the type necessary for controlling a telephone network, e.g., disconnect signals.

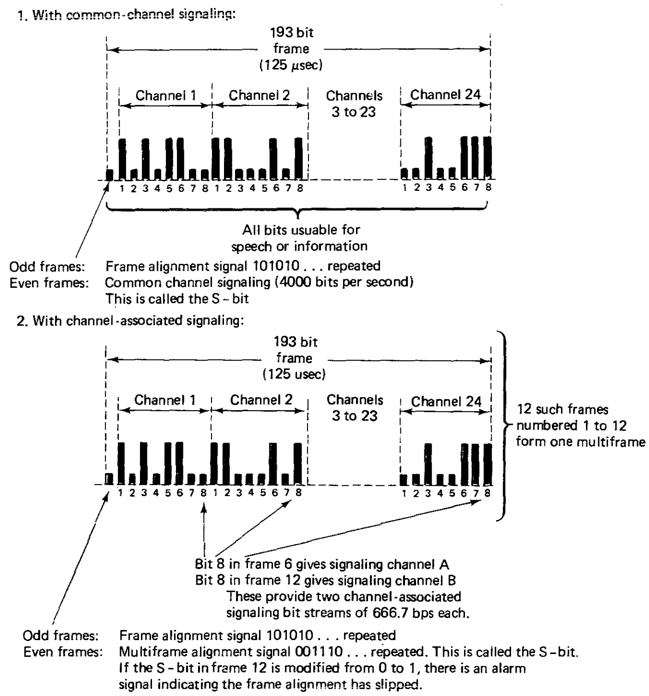


Figure 5. Signaling with PCM systems operating at 1.544 mbps employing CCITT Recommendation No. G.733

The version that allows channel-associated signaling is shown at the bottom of Figure 5. Here, two signaling bit streams are associated with each channel. The frames are arranged into groups of 12, called a multiframe, and numbered 1 through 12. Bit 8 of each channel in frame 6 is reserved for signaling channel A. Bit 8 of each channel in frame 12 is reserved for signaling channel B. These bits each occur 666.7 times per second, hence each voice channel has two 666.7 bps signaling bit streams associated with it. The first bit of each frame is used for both frame and multiframe alignment.

With the first of these schemes, the channels are composed of 8-bit words, as opposed to 7-bit words on the T1 carrier. They thus have 64,000 bits per second, as opposed to 56,000 bits per second on the T1 carrier. With the second scheme the words are also 8 bits, except that every sixth word in a channel has only 7 usable bits. The usable channel rate is therefore 62,666 bits per second.

Figure 6 shows the CCITT 2.048 mb/s recommendation, which most of the world outside North America

The Nature and Organization of Digital Transmission Channels

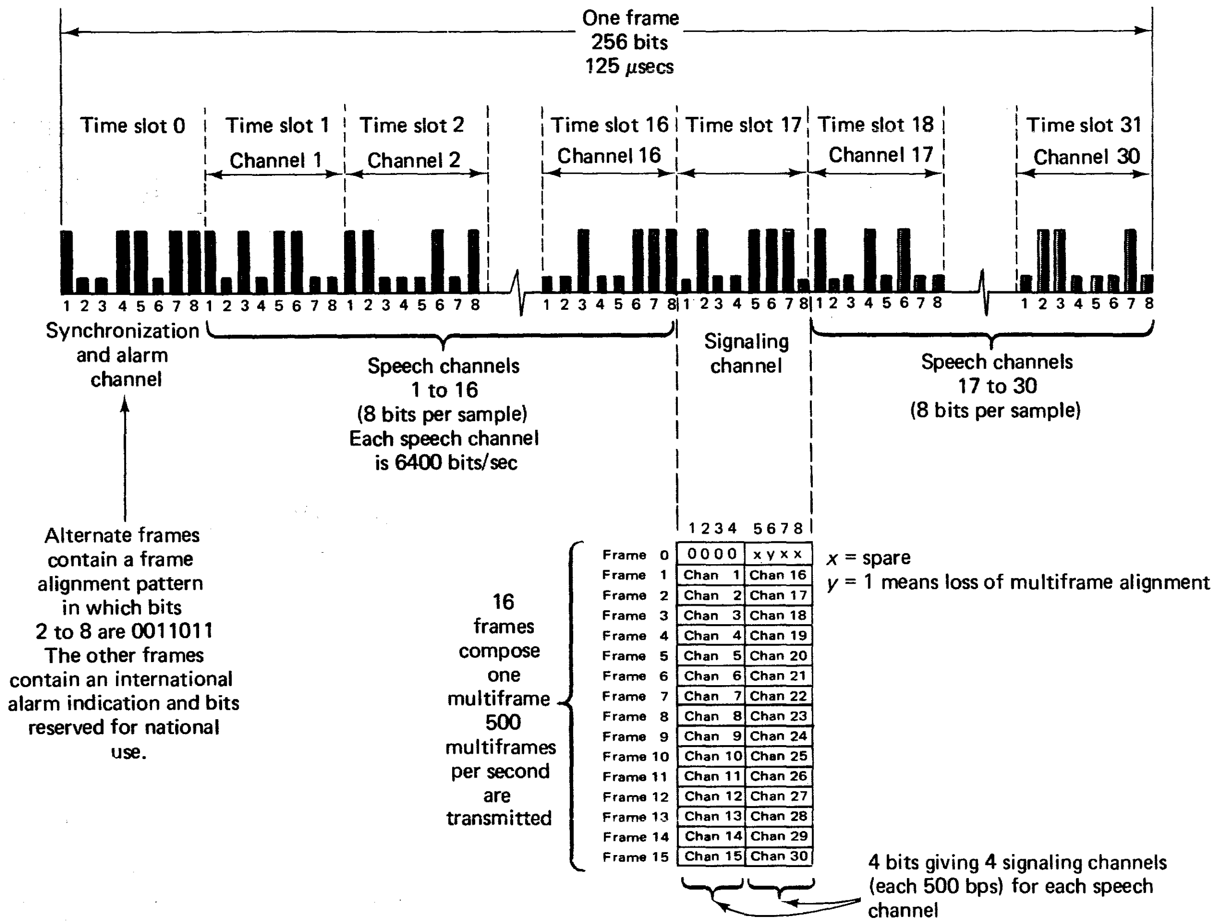


Figure 6. CCITT Recommendation for the structure of PCM channels for transmission at 2.048 million bits/second, 30 speech channels of 64,000 bps are derived, each with a signaling channel of 500 bps

and Japan is starting to use for PCM transmission. In this, 16 frames of 256 bits each form a multiframe. There are 32 8-bit time slots in each frame giving 30 speech channels of 64,000 bps each, plus one synchronization and alarm channel and one signaling channel, which is submultiplexed to give four 500 bps signaling channels for each speech channel.

The difference between the CCITT and North American standards will prevent the world from becoming linked with digital channels on satellites and other media without the need to convert from one system to another.

NATIONWIDE NETWORKS

The early PCM links were relatively short point-to-point connections between telephone offices and were somewhat experimental in nature. The Bell T1 system grew into wide acceptance until much of the U.S.A.'s short-distance wire pairs trunks were converted to T1. It is now clear that the various digital facilities must link together to form a nationwide network in the United States and eventually a worldwide network.

When the T2 carrier came into use on the Bell System in 1972, it was designed for transmission up to 800 kilometers, but much attention had been paid in its design to eventually linking it into nationwide facilities. Except for very short distances, telephone calls are handled more economically by the T2 than by the T1 carrier. The T2 carrier, however, was designed to have an additional purpose—the trunking of Picturephone signals. Distortion on a Picturephone signal is much more harmful in its effect than on a telephone signal. For this reason Picturephone signals will not be sent long distances on analog trunks; digital trunks will be used. One Picturephone signal occupies one T2 bit stream. The T2 links were restricted to 800 kilometers. Therefore, various techniques were designed to send high digital bit rates over existing analog channels—microwave and coaxial cable—to bridge the gap until nationwide digital trunking exists.

DATA TRANSMISSION

Whereas the need for nationwide Picturephone transmission is far from pressing as yet, nationwide data

The Nature and Organization of Digital Transmission Channels

transmission over the digital channel is an urgent present-day requirement. It is far more efficient to send data over the PCM telephone channels than to convert it to analog form and send it over analog channels. AT&T has a service called DDS, the Dataphone Digital Service, in which data can be transmitted from subscriber to subscriber in direct digital form—i.e., no modems.

To provide an end-to-end data service, several components are needed as well as the T1 and T2 trunks we have described. First, a data-carrying local loop must be established from the subscriber to his local central office. Second, if that central office does not yet have PCM trunks reaching it, the digital signal must be transmitted over other trunks. Third, to give a nationwide network, the segments of data-carrying PCM system must be interconnected.

An economical means to transmit data over analog trunks has come into use on the Bell System, called Data Under Voice, DUV. As we have commented, analog voice channels are frequency-division multiplexed together to form groups, which are commonly transmitted over a microwave link. A mastergroup, for example, consists of 600 voice channels, a jumbogroup of 3600. When these groups are sent over a microwave link, there is a gap underneath them in the transmitted radio band. The gap is 564 KHz wide, and nothing is transmitted in it except for a radio pilot, which is used to indicate radio continuity (Figure 7). The gap exists because the signal at the bottom edge of the band is too variable for good quality speech transmission. Data, however, can be transmitted in the gap with high accuracy.

When the T1 carrier bit stream is encoded as in Figure 4, its spectrum is too wide to fit in the gap, as shown in Figure 7. It is therefore recoded using a 7-level code. A data pilot is added, to monitor the signal, and the signal fits comfortably underneath the lowest group of telephone circuits. The radio continuity pilot has to be moved out of the way to a higher frequency.

Data Under Voice, DUV, made it possible to interconnect the digital carriers without building new physical links and to rapidly build a nationwide data network. DUV will fill the gap until nationwide T4 links come into existence. Much of its potential value lies in the fact that microwave links go to most cities, including small ones.

DUV is designed to give a low error rate. A design objective is that on a 4000 mile connection better than 99.75% of all customer channel seconds shall be free of error. A 4000 mile connection will normally go over many different multihop radio systems; 16 radio systems would typically connect in tandem.

DUV and the T1 and T2 carriers made possible the AT&T Dataphone Digital Service (DDS) offering.

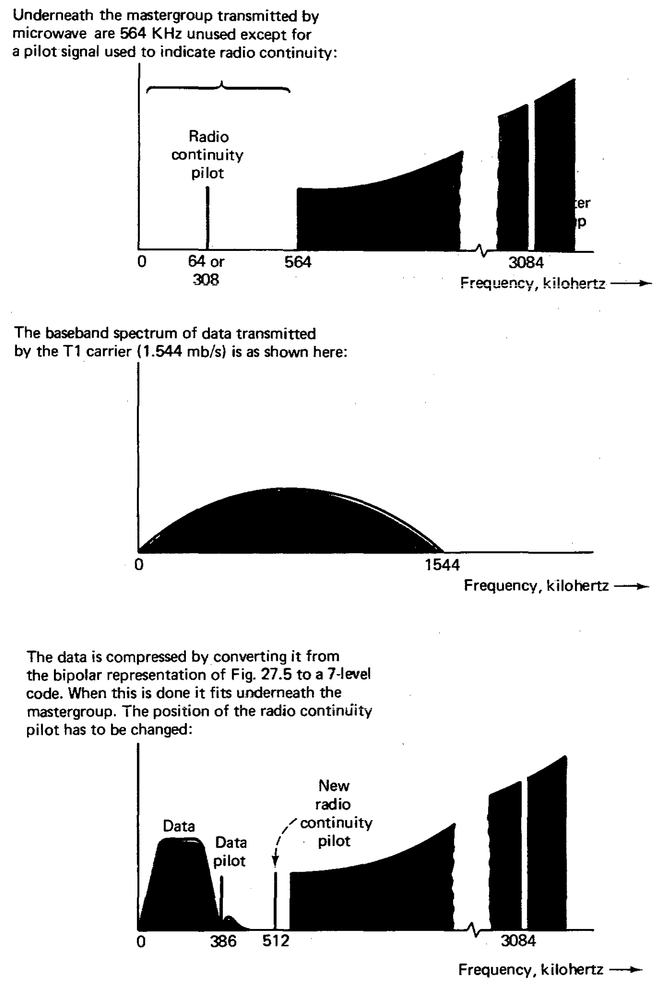


Figure 7. On the Bell System, segments of T1 carrier can be linked together nationwide by using DUV, Data Under Voice, on microwave channels

Customers leased channels at speeds up to 56,000 b/s, which are digital end-to-end and hence require no modems. DDS could presumably handle data rates much higher than 56,000 b/s were it not for the limited capacity of the local telephone loops.

By putting digital repeaters on the local loop, somewhat like the T1 carrier, digital services could be expanded up to 1.544 mb/s. It is also possible to make digital services switchable.

Western Union, like AT&T, requires nationwide digital facilities super-imposing on an existing nationwide analog network. Substantially more than half of Western Union's traffic is digital in nature. Western Union is therefore modifying existing analog microwave routes to make them hybrid microwave carrying 6.3 mb/s in the lower part of their spectrum. On other microwave routes, digital radio is being added, using the same towers and antennas and operating initially at 20 mb/s. The latter approach is called digital over-build. Western Union's WESTAR satellite transponders also transmit either analog groups or high-speed digital bit rates. Figure 8 shows the Western Union digital hierarchy.

SYNCHRONIZATION

Synchronization is vital for digital transmission. It is essential for the receiving machine to know which bit is which. This is not too much of a problem on point-to-point lines. The Bell T1 carrier solved it by adding one extra bit per frame, thereby obtaining an 8000-bit-per-second signal that carries a distinctive pattern. If synchronization slips, then this pattern is searched for, and synchronization can usually be restored in a few milliseconds.

This is fine so long as the channel bank remains intact. If, however, the bank were split up into its constituent channels and these channels were transmitted separately by pulse code modulation, then it would be advantageous to have a synchronization bit sequence for each channel rather than for the group of channels. The result would be one bit per character rather than one bit per frame, as in Figure 3. One bit per character is already used for network control signaling, and the result is 8000 bits per second in this case. This rate seems far too much for any purposes that can be foreseen at the moment. Network signaling consists mainly of sending routing addresses (the number you dial) and disconnect signals (when you replace your receiver). Therefore it has been suggested that this bit position should be shared between the network control signaling function and the synchronization function.

Synchronization becomes a much more difficult problem when a large switched network is considered. Signals are transmitted long distances over different and variable media, and their time scale must inevitably differ slightly even if the transmitting locations are synchronized. This problem must be solved before a nationwide PCM network can be established. There are two types of solutions. The first is a fully synchronous approach in which an attempt is made to synchronize the clocks of the different switching offices and to compensate for any drift in synchronization of the information transmitted. The clocks of the different offices in the network must all operate at exactly the same speed. The second is a quasi-synchronous approach in which close but not perfect clock correlation is accepted, and the multiplexing operations must be designed to cope with the imperfections. For the time being the latter approach is the more practicable one.

Two bit streams, then, that are to be multiplexed together on a T1 or higher level channel have very slightly different bit rates. To achieve the multiplexing, more pulses are available on the outgoing line than on the incoming lines. The excess time slots are filled with dummy pulses. The presence of the dummy pulses is signaled, and they are removed at the receiving end. This technique is referred to as pulse stuffing. The more accurate the clocks, the less pulse stuffing is required. As the design of the digital hierarchy evolves, the accuracy of the clocks employed may increase, and the degree of pulse stuffing needed may decrease.

AT&T'S T4M SYSTEM

In most cities the ducts beneath the streets for telecommunications channels are almost full, and more capacity is needed. Digging up the streets to lay down larger ducts is extremely expensive. A more attractive option is to replace some of today's cables with digital cables that carry a higher traffic volume.

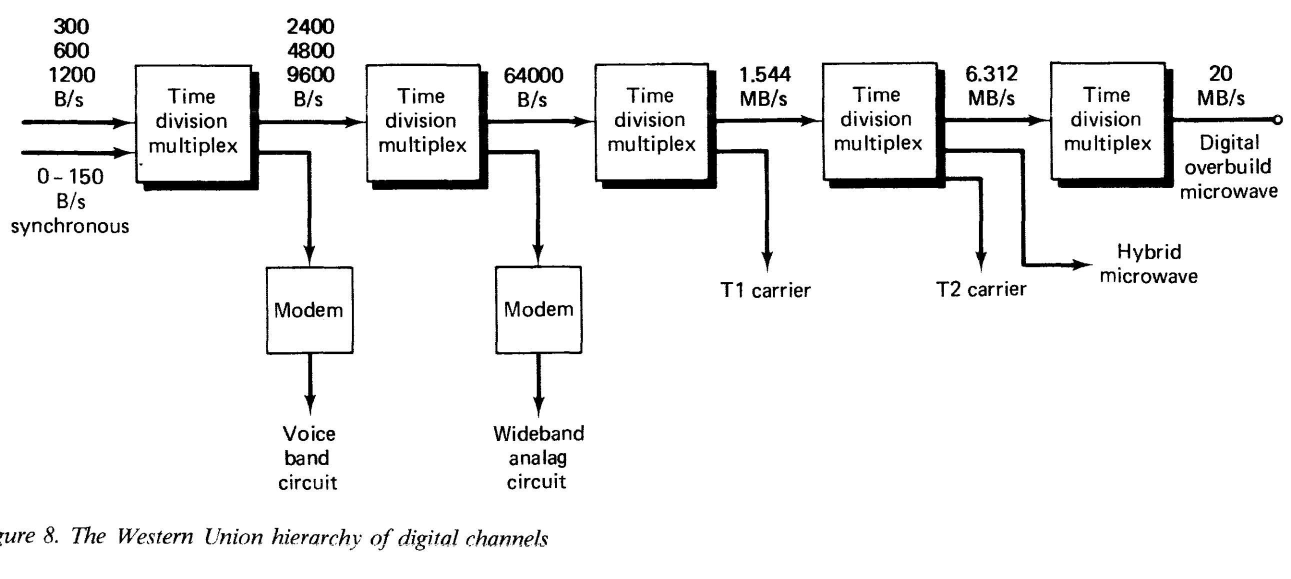


Figure 8. The Western Union hierarchy of digital channels

The Nature and Organization of Digital Transmission Channels

The Bell System T4M coaxial cable system comes in two versions, one designed to fit into existing 3½-inch ducts and the other to fit into the newer 4-inch ducts. The former has 18 coaxial tubes per cable, and the latter has 22 tubes per cable. The cables are the same as those used on today's analog long-haul systems, but the electronics are entirely different. Each tube can transmit 274 mb/s, carrying 4032 voice channels. Two tubes in a cable are spare and can be automatically switched into operation if a failure occurs on another tube. Of the remaining tubes, half transmit in one direction and half in the other. Thus, the smaller T4M cable transmits a total of 2,192 billion b/s in each direction, carrying up to 32,256 two-way telephone conversations; the larger transmits a total of 2,740 billion b/s carrying up to 40,320 two-way telephone conversations.

The central conductor of each coaxial tube carries the DC current, which is used to power the regenerators until the next maintenance office is reached. The span between such offices can be up to 180 kilometers.

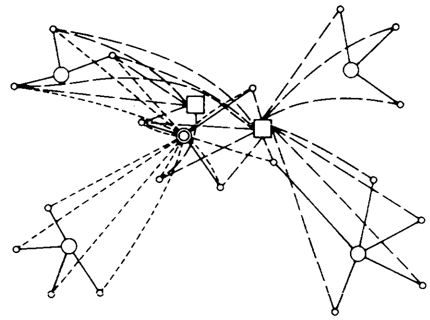
In addition to occupying a dedicated cable, the T4M bit streams are sometimes sent over certain coaxial tubes in cables that also carry other types of signals. Some composite cables, for example, have 8 coaxial tubes and 750 pairs of wires.

EVOLUTION

An enormous amount of capital is tied up in national analog telephone networks. So much money is involved that they cannot be converted to all-digital networks in a few years. The proportion of digital links will grow slowly. They will be installed first where they can be most profitable or where the pressure for extra circuits is greatest, as in crowded urban and suburban areas. The T4M system was first installed in the New York to Newark area and was pressed into service early after a catastrophic fire in 1975. It had to be tailored to the available cable ducts and traffic requirements.

Figure 9 shows how PCM lines might be used in today's typical urban environment. It presents a highly simplified picture of part of the telephone network in London. The top half of the picture shows interconnections between local, tandem, and subtandem exchanges. The bottom half shows how the network could be simplified with the use of PCM links. The simplification shown by the illustration would have appeared far more striking if it had been drawn with the 200 or so local exchanges that exist rather than with the small number in the diagram.

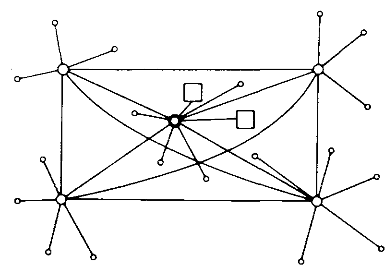
The saving would be much greater, again, if PCM concentrators were also used.



○ Local exchange
○ Subtandem exchange
◎ Tandem exchange
□ Toll/trunk outlet
——— Junction to subtandem
- - - - - Junction to tandem
— - — Junction to toll/trunk outlet

A highly simplified diagram of the present London step-by-step tandem network catering for subscriber trunk dialing (direct distance dialing in American parlance). Only a limited number of exchanges and routings are shown, in order to lessen the complexity of the figure. All direct junctions between local exchanges have been omitted.

With integrated PCM between exchanges, the routing might be simplified as follows.



If the drawing had not been so highly simplified, the saving would have appeared much greater.

Figure 9. Redrawn with permission from Techniques of Pulse Code Modulation in Modification Networks, by C. C. Hartley, P. Morret, F. Ralph, and D. J. Tarran, Cambridge University Press, Cambridge, 1967

ADVANTAGES OF PCM

PCM can give lower costs per telephone channel on short-haul lines and can greatly multiply the utilization of lines within city areas. Two trends will widen the range of economic application: the decrease in bandwidth cost due to the introduction of higher-capacity channels and the decrease in cost of logic circuitry, which is likely to be great when large-scale integration is fully developed.

PCM offer the following other advantages in addition to the overriding economic argument:

1. Because the repeaters regenerate the bit stream, PCM can accept high levels of line noise, cross talk, and distortion. A substantially worse noise-to-signal

The Nature and Organization of Digital Transmission Channels

ratio can be accepted than with frequency-division multiplexing.

2. The transmission is largely unaffected by fluctuations in the medium, provided that they do not exceed certain limits. This is termed ruggedness in a transmission system.

3. All types of different signals, such as speech, music, television, Picturephone, facsimile, telegraphy, and computer data, will be multiplexed together in a digital form. These signals can all travel together over the same facilities without interfering with one another. In analog channels the system capacity is often limited by mutual interference between different types of signals.

4. Much higher data rates can be achieved than with analog transmission. This factor will become increasingly important economically as the use of computers and terminals increases. Data transmission is increasing much faster than other forms of transmission. Unless held back by unsuitable or expensive transmission facilities, it will continue to rise probably at an increasing rate. Most common-carrier predictions of this market seem gross underestimates. There will be little relationship between present-day telegraphy and the future uses of data transmission.

5. Some future transmission media, such as optical fibers transmitting laser pulses may be inherently digital in nature.

6. Some future satellite systems may be dominated by the use of demand-assigned multiple-access. Computerized earth stations handling time-division multiplexed bit streams will probably give the most economic use of satellites.

7. Many bits are available for network control signals. The possibilities for signaling and remote control of the network are therefore much greater than with today's analog plant.

8. Encryption of signals is likely to become an important subject with the increasing concern about privacy and increasing need for security in data processing systems. Digital transmission makes effective encryption easy to achieve. Encrypting devices may perhaps be used in the private branch exchange of the future, under computer control. The analog scramblers, familiar to viewers of World War II movies, are of little use today because computer methods make deciphering of the scrambled signals easy.

9. Time-division multiplexing provides advantageous switching methods as well as transmission multiplexing. The networks of the future will integrate switching and transmission technology, both, in part, using digital techniques. Time-division switching costs will be lower if the transmission also uses time-division multiplexing.

10. Concentrators can substantially lower the cost of the local distribution network. As will be discussed, digital techniques provide a way to build inexpensive concentrators. These devices could be used in large numbers in future telecommunication networks.□

Why Protocols? Why Standards?

Problem:

The idea of a communications protocol is not unique in human affairs. Consider a simple telephone conversation:

Ring. Ring. (Station-to-station connect alert in a dial-up network.)
"Hello. Acme Trucking." (Open link. Receiver ID)
"Hello. Mr. Smith, please." (Acknowledgment. Query.)
"Who?" (Request for retransmission.)
"Mr. Smith." (Retransmission.)
"Oh, you must mean Mr. Smythe." (Error check/correction.) "Who shall I say is calling?" (Request for sender ID.)
"Mr. Zybczinski." (Sender ID.)
"Would you spell that please." (Bandwidth problem.)
"Z-B-Y-C-Z-I-N-S-K-I." (Retransmission at a slower rate.)
"Thank you. I'll see if he's in." (Acknowledgment. Implied wait for second acknowledgment before sending more.)

. . . and so on. Practically every step of a human dialogue adheres to some form of implicit or explicit protocol. And so it is with communications links except that every step must be unambiguously explicit, and the sender and receiver must speak exactly the same language when communicating, regardless of what language they speak away from the link. Unfortunately, communications designers have not yet agreed on one common, universally acceptable language.

The general subject of sender-receiver synchronization, of which protocol considerations are one component, is covered extensively throughout this section. This report reviews the concepts and important constituents of protocols to set the stage for the various protocols described in more detail throughout Section CS93. Some of the protocols, like RS-232 and ASCII, have acquired the dignity of "standards" that cut across manufacturer's equipment boundaries and are used by the world at large. Other protocols, like SDLC and UDLC, tend to be manufacturer-dependent formats whose use is generally restricted to the manufacturer's equipment or to compatible equipment produced by other manufacturers.

Solution:

Two words applying to conventional social behavior have been adopted to explain or identify certain data communications requirements. These words are protocol and handshaking. Protocol, a set of rules for diplomatic occasions, and handshaking, a greeting preliminary to a conversation between two people, are particularly appropriate to identify the set of rules that govern the orderly exchange of information between a computer and its connected terminals.

There are two basic questions that arise in discussing line protocols (IBM uses data link controls to mean the same thing.) These two questions are:

- Why is a protocol required?
- What can a protocol do?

WHY A PROTOCOL?

The first question, the why, is a little like asking which came first, the chicken or the egg. Basically, the need arises because of two things:

- Efficiencies can be achieved by reducing the amount of information transferred; this is most easily accomplished by pre-setting certain conditions—an error checking technique, for example. This means that the extra information required to identify the type of error checking need not be included in each transfer of information because the receiving end already knows what technique will be used.
- Every process requires some measure of control information exchange. This is difficult on a communications line because only one information path exists. Therefore, there must be some set of rules that allows the receiving terminal to distinguish between control information and data. (Only one path exists because of the cost of establishing separate, little-used control information paths. With a computer peripheral, a few additional connecting wires to provide separate control and data paths are of little cost concern.)

The first reason, the drive for efficiencies, further leads to the desire to reduce the necessary control information to a minimum by using arbitrary code sequences to identify certain conditions. The second reason, the need to distinguish between controls and data, means that a fixed set of rules is required so that the receiving equipment can properly interpret what is coming over the communications line.

Technically, speaking, any type of communications, even for the most simple arrangement, requires operation under a fixed set of rules. For example, transmission between two teleprinters permanently connected by a leased line still requires certain pre-established rules; parameters such as transmission speed, identification of a 'one' and a 'zero,' data code, how to verify a connection, etc., are all pre-set. The item with the most variability in this simple case is the data code. Often, in simple arrangements, the data code is the complete statement of the 'official' protocol because the other parameters (speed, etc.) are obvious. Included in the data code can be certain characters (bit patterns) reserved for control purposes, such as turning a punch on and off, etc.

Use of non-language characters for control purposes goes a long way in handling the control requirements of more complicated communications arrangements. For example, several terminals can share a communications line, but not simultaneously. Multipoint or multidrop operation requires some means of identifying which terminal is entitled to operate at any point in time. The conventional method for doing this is to have the computer address a special message, or poll, to each terminal in a predefined order. This arrangement allows the computer to control when and with whom transmission takes place on the theory that it is better to have the remote station wait a few seconds than to risk overloading the computer. Other non-communications events could take place at the same time that a terminal is transmitting, and the data could be lost if the computer cannot handle two operations at the same time. To establish polling, control information has to be exchanged. To exchange control information, it must be separated from data. To separate control information from data, a set of rules must be established. To have a more elegant way of referring to a 'set of rules' we borrow the term protocol.

Further efficiencies can be gained if the locations of certain types of control information are fixed within the transmitted information. For example, if the last data byte of a transmission is always the error checking information, then extra information need not be used to identify it. Thus, position can be effectively used to save transmission time. The set of rules for positioning information within a transmission is called message format and becomes a part of the protocol.

The two tools of the protocol designer are the use of non-language characters for control purposes and the use of position within a message to eliminate needless identifications.

Protocols stem from the need to exchange control information and the difficulty of distinguishing between control and data if they are intermixed on

Why Protocols? Why Standards?

one information path. A protocol provides a method for the orderly exchange of data by establishing rules for the proper interpretations of controls.

WHAT CAN A PROTOCOL DO?

A few examples, such as offered above in explaining why protocols are required, merely whet the appetite. A full discussion of a particular protocol, such as IBM's Bisync and SDLC or ASCII is too restrictive in that it explains in detail one set of rules without giving viable alternate approaches.

In the following paragraphs, we will set forth some overall objectives for controlling the communications task. They will serve as sort of a checklist for understanding any individual protocol.

The basic tasks required in controlling information exchange over a serial communications facility include:

- Establish and verify a connection.
- Accommodate security provisions.
- Establish and/or verify identifies.
- Establish precedence and order of transmission.
- Handle data sequencing (i.e., blocking).
- Handle error control (including retransmission).
- Permit interruptions or temporary suspension of communications with re-establishment of transmission.
- Permit control of devices and features attached to communicating stations.

Not all of these tasks need be implemented to have valid data communications. For example, a single terminal connected to a computer over a leased line does not require establishing a connection or identities; they are hard-wired into the physical arrangement of the communications system. The same single terminal connected via a dial-up (DDD) connection might require the establishment of identities to verify that the correct destination had been reached. An alternate, and quite valid procedure, is not to check who the computer reached, but merely to attempt communications. If it fails, try again. If it still fails, send a systems message that something is wrong. If the connection is dialed manually, then it is quite valid to have the verification of identities made manually. It's your choice as to how much time you save and how much you increase your budget.

These are three levels at which any of these tasks can be implemented:

- Manual.
- Fixed-logic (hard-wired).
- Programmed.

These are listed in order of increasing cost, increasing flexibility, and increasing utility. Speed of accomplishment is: manual—slowest; programmed—second slowest; and fixed logic—fastest; this represents a very large span of speeds. The frequency with which the tasks are required usually determines the most effective choice among the three. Usually, an equipment parameter determines which tasks are performed in which manner.

If the tasks are performed in response to information contained in the transmission, either directly with control codes and position or indirectly by pre-set rules, then the task is included in that particular protocol.

In addition to any control tasks accounted for by the protocol rules, the systems or applications software can be used to expand the possibilities by simply interpreting and reacting to information contained in the data portion of the transmission, which is transparent to the protocol rules. Protocol rules are difficult to modify because the key ones are frequently set up in the hardware of the remote station.

The most basic function is to establish and verify a connection. If this is not accomplished, then no communications at all can take place. Methods can range from manual initiation and coordination of a call to automatic screening of a response to determine whether it is within an acceptable set or matches the expected one. Verification of a connection is frequently linked to establishing identities in automatic procedures. If a proper response (i.e., the addressed station response includes that station's address or identifier) does not occur, then the assumption is made that a valid connection has not been established.

Security provisions are required if you wish to restrict in any manner who can reach the system and for what purpose. Most frequently, security is a system-defined function based on information contained in the text (data) portion of the transmission.

Precedence and order of transmission is a critical task. Well defined rules are required to condition all stations as to when to transmit and when to expect something back. This applies to all types of information exchanges including control exchanges as well as data exchanges. For half-duplex operation, the modems must be turned around from transmit to receive (or the reverse) for intelligible communications to take place. In full-duplex transmission, physical intelligibility is less threatened, but logical intelligibility is still dependent on each station being prepared to handle what is going on.

Why Protocols? Why Standards?

Data sequencing refers to breaking a long transmission into smaller blocks and maintaining message control among the divisions. Such divisions are usually used in conjunction with error control techniques to reduce the amount of data that must be retransmitted in case of a detected error and also to economize on buffer sizes. In addition, all error checking methods perform better (i.e., catch more errors) with shorter blocks. However, transmission efficiency decreases as the block size decreases. Block lengths of up to several thousand characters are used, but lengths of from 80 to 512 characters are the most common. Block length may or may not be a provision of a protocol.

Error control consists of a method to detect errors and a method to recover when errors are detected. Very sophisticated techniques for transmitting redundant information in special ways so that errors can be corrected from the receiving end have been known for a long time, but are seldom used. The most common method for correcting errors is simply to retransmit a block. This requires coordination of the two stations so that the flawed block is discarded by the receiving station and the block is repeated by the transmitting station. For half-duplex operation, this takes extra time to turn the line around. For full-duplex operation, there is a timing or sequencing problem to assure that the flawed block at the receiving end is the one that is retransmitted by the transmitting end. Error detection and correction techniques are actually independent. Once an error is detected, then correction takes place. (Keep in mind that no error detection technique can detect all errors. Your systems procedures must include provisions for handling errors once they are past the communications subsystem and into the data processing system. Fortunately, the incidence of undetected errors is typically low.)

Interruptions to the continuous flow of information can be disastrous if no provisions are made to accommodate planned or unplanned interruptions. Everything that has gone before may have to be discarded and the process started from scratch. In conventional data processing, this problem is solved by inserting restart points in the programs. If a program bombs (and they do on occasion), you need only go back to the last restart point rather than all the way back to the beginning. Block retransmission is the form of restart procedure most frequently employed for communications. Individual messages form natural restart points. Another procedure frequently used is a timeout if data is not received within a specified period of time (usually a few seconds) with automatic disconnection or advance to the next poll address after a timeout. Usually timeouts are an equipment-defined parameter and are seldom included in the protocol. The usual attitude is that error detection with retransmission will keep things running until the system can somehow detect that something is wrong and take corrective action such as printing a console message.

Remote stations can include many devices such as printers, card readers and punches, tape readers and punches, magnetic tape cassettes and cartridges, disk drives, plotters, CRTs, etc. If these are to be controlled from the computer site, special provisions need to be made to permit interpretation of the control sequences for the devices. Typically, such controls are included in the data portion of the transmission and are not a part of the protocol.

NO CONFERENCE CALLS

Much has been written about architecture of networks. Packet switching, networking, star network, ring network, distributed processing, multi-nodal arrangements have all been used to describe complex interconnections among multiple stations with a wide range of capabilities and responsibilities. However, the basic element of computer-oriented communications is still the simple two-way interchange between two stations. More sophisticated arrangements typically employ several of these simple links to establish a communication between the originating station and the final destination. The protocols in common use today (IBM's Bisync and SDLC and ASCII) are all oriented to two-point communications. True, each can accommodate multipoint arrangements, but only two stations are active at any one time. To establish communications beyond the simple two-point arrangement, the equipment at each point must retransmit the message, again in two-point fashion, to the next level; this is repeated until the message reaches its destination (as identified within the message). While this does not sound elegant, it works and does provide complete control over the process. One alternative is use of a network in which the communications facility supplier provides a direct path between the origination and destination points. Another alternative is to have sufficient interconnections in your network to permit any point to reach any other point; not only does this increase the cost of your network facilities drastically, it will increase the cost of your remote stations to include the extra capability for managing the additional decisions required.

However, the subject of complex network architectures and their management is beyond the scope of this report. It is only brought up to reinforce the idea that most communications today are built up from simple two-point connections.

In summary, a protocol makes possible intelligible communications, which means that origin data reaches its destination with all of its original meaning intact.□

A Tutorial on Protocols

Problem:

Armed with a definition and knowledge of the purpose of a protocol, the systems designer may want to know some of the sources of protocols and the current state of protocol development. Protocols are employed in naming the recipient of the data, in error control, flow control and synchronization of transmission, and other areas. An understanding of this becomes increasingly important if plans call for implementing a distributed system involving heterogeneous terminals, processors and transmission facilities. What are the functions of protocols in such a system and how are the various elements of protocols identified?

Solution:

Everyone has had the opportunity to overhear such cryptic conversations exchanged over the radio by taxi drivers, policemen, and aircraft pilots. Although upon hearing these conversations at first do not mean much to the layman, these abbreviated languages carry well-defined meanings and obey well-defined rules. Speakers give their name, ask correspondents if they are listening, confirm reception, etc. This form of conversation differs drastically from a face-to-face chat. The communication channel is shared by many speakers. To save bandwidth and reduce interferences, messages are short and coded. External noise and other interferences are common occurrences, hence, repetition and confirmation are normal practice. These rules are known as protocols.

The term protocol entered the computer jargon at the turn of the 70's, when the U.S. Defense Advanced Research Project Agency set out to build a network of geographically distributed heterogeneous computers.[71] Up to that time, communication between computer programs or processes was limited to processes which were located within the same machine. Inter-process communication was accomplished through the use of shared memory and special signals exchanged through the mediation of the operating system. This technique represented the analog of a face-to-face chat between processes. Inter-process communication between geographically distant systems would have left processes with the same kind of constraints that taxi drivers encounter. They would have to interact through a potentially hostile environment with limited bandwidth, delay and unreliable transmission. In addition, the processes in the different computer systems did not even speak the same native tongue, having been created by different manufacturers.

Computer veterans remember the sinuous evolution that led from binary programming to assembly code, to Fortran, Cobol, Algol, and other high-level languages. Originally viewed as a collection of tricks and hobbies, programming languages have developed into a major branch of computer science. The evolution of protocols has followed a strikingly similar path. Indeed, Protocols are common tools designed for controlling information transfer between

A Tutorial on Protocols by Louis Pouzin and Hubert Zimmerman, IEEE, from the November 1978 Proceedings of the IEEE, Volume 66, No. 11. Reprinted with permission.

A Tutorial on Protocols

computer systems. They are made up of sequences of messages with specific formats and meanings. These messages are equivalent to the instructions of a programming language, although protocol languages are still in an early stage of ad hoc development.

Actually, progress has been very rapid. Techniques for protocol formalization and verification are already under development.[25] The primitive era is passing and a body of protocol science is emerging. It is time for computer and communications specialists to acquaint themselves with the basic concepts of protocols and to make them a part of their repertoire of design and implementation tools.

Summary

The first section of this report is a short introduction to protocols. It presents a simplified and condensed view of the rest of the report, with the emphasis put on examples drawn from analogies. This should help the reader to acquire an intuitive vision of the whole subject, and to understand more easily the following sections.

The second section presents the basic elements found in protocols. Indeed, the elementary components used for building protocols are very limited in number although they appear under different names and in different forms in the literature. The first step towards understanding protocols is to uncover the simple mechanisms used in their construction.

The third section introduces the concept of a distributed system architecture. It shows how the functional system components can cooperate with one another through the use of protocols to organize an otherwise heterogeneous system into a consistent homogeneous structure. It explains the role of protocol layering in the rationalization of distributed system architectures. This is a crucial concept since, without layering, it would be impossible to allow local optimization of different parts of the heterogeneous system while preserving opportunities to establish common conventions for interprocess communication, independent of the applications to be supported.

The next section contains a discussion of the functional elements of typical layers, for example, data transmission, transport and virtual terminal protocols. These layers form the foundations of any distributed system.

The next section shifts the focus from design to engineering. It points out the advantages and disadvantages associated with the various basic protocol elements and shows how performance is affected by design choices and environmental factors.

The final section addresses standardization. As in the case of programming languages, protocols will only be effective when they are widely adopted by large communities. Due to commercial interests, this is an area replete with nontechnical complications. Major data communication standards are summarized, along with work in progress within standardization bodies.

PROTOCOL OVERVIEW

Basic Protocol Functions

In the first subsection, we introduced the intuitive notion of a protocol and provided a simple working definition. To the newcomer, it is necessary to understand intuitively the ingredients which comprise a protocol in order to acquire a general mental model of its structure and its raison d' etre. We proceed by giving an analogy with the interaction between organizations in a business context. Admittedly, this analogy will only give a macroscopic view of the nature of protocols. Terms and concepts are used and intended to be understood generally without the need for introducing very precise definitions at this point. We have made obvious approximations and omissions for the sake of achieving clarity in presentation.

Let us assume that two corporations, SIXOUNS and SNAPSE, are considering some joint business. They both produce whisky (of course). One of the two presidents will presumably invite his colleague for lunch (no martinis, please), and agree upon a business strategy. They will then turn the matter over to their marketing vice-presidents who will arrange a meeting. Later on, plant managers may get together, etc. For our model, we assume regular discussions take place between opposite numbers, i.e., between people at the same level (or layer) in the corporate hierarchy. However, these discussions are highly informal and also time-consuming during the initial stages. Since the strategy is now shaping up, SIXOUNS and SNAPSE feel that they should streamline their interactions. To that effect, they design procedures for maintaining a mutual exchange of information between the various participants, without the need for actually meeting. These procedures might include such elements as rules, message or document formats, frequency of interaction, etc., which are established by a particular layer of the corporate hierarchy. They constitute the protocol to be used by that layer.

SIXOUNS and SNAPSE have created a common distillery. Each corporation uses its own network of retail stores to market its own brand. Orders are collected separately by SIXOUNS and SNAPSE, and transmitted to the distillery. On the other hand, deliveries are carried directly from the distillery to the retailers. Occasional mixups may happen, such as delivering a SNAPSE whisky, instead of a SIXOUNS, or miscounting the number of bottles, or dropping a carton of whisky on the floor with shattering consequences. In order to prevent disputes from arising, each corporation acknowledges each order received from its retailers, and retailers have to sign delivery vouchers. Thus, discrepancies may be traced back to their origin and corrected. We may say that order and delivery protocols include error control.

The volume of orders may at times vary substantially. This fluctuation causes inefficient operation for the distillery because its production capability is limited. Sometimes, orders must be delayed because of overload, while at other times there is unused distillery capacity. Thus, it was decided that SIXOUNS and SNAPSE should place production orders by anticipation, in order to regulate the amount of whisky produced. Of course, some care must be exercised, since the distillery has also a limited space for inventory. This tricky balance is a matter of flow control.

Even though considerable care is taken in handling properly the orders, deliveries and inventory, some errors may slip through, due possibly to human factors. An occasional discrepancy is acceptable so long as it has no cumulative effect. To prevent that, the books are balanced once a month, at which time all figures must be reconciled. This accounting is not completely straightforward, as there are orders or acknowledgements in the mail, and filled trucks on the road. In fact, it is not possible for all accounts to be settled simultaneously and, as a result, there is a need for a specific set of accounting rules so that all outstanding orders, deliveries, and production be clearly accounted for each month. This protocol may be called synchronization.

Another requirement always appears whenever an exchange of information is to occur, viz, providing the address of the intended recipient. In the whisky business, an order may be sent by mail to the following address:

SIXOUNS Corp.
Order Dept.
4/5 Bourbon St.
DOWN HATCH 70127 YU

Or the retailer may dial 435-2217, ask for extension 652, and say: Hello Sally . . . This is Louis. I would like . . . etc. Or he may send a TWX message, or something else. In these examples, the composite address is actually a concatenation of several address elements. Some elements are meaningful to the Post Office or the public telephone system. Other elements are only meaningful within the SIXOUNS corporation. The retailer must learn them beforehand.

A different example illustrates another addressing possibility. A friend of mine happens to be in town and has previously advised me that he will be staying at the DALTON hotel. (Of course, he had to know me and my address). How can I send him a message? I may call the DALTON, and leave a message for Mr. Jim DOLITTLE with the operator. I could also ask the hotel operator for my friend's room number, and the next time I may just leave a message for room 625. Now, there could be more than one Jim DOLITTLE in the DALTON, or he may have moved to a different room, or checked out. If I really want to be sure that my message gets to the right person, I have to check further.

Things would be simpler if every person in the whole world had one portable telephone (or TWX) number which only he can use. When it answers, then we know the right person is at the other end.

The previous examples illustrate three typical methods used for conveying information to its destination. The first example corresponds to a hierarchical addressing strategy, the second example denotes a method based on dynamic allocation of temporary names, and the third example illustrates mapping. In the protocol jargon we often use the term naming to designate methods of composing addresses and selecting a proper correspondent. Each of the above three naming approaches, namely hierarchy, allocation and mapping will be discussed in the second section along with the other key elements of protocols, namely, error control, flow control and synchronization.

Distributed Systems

So far, our examples have been taken from a common business setting without any reference to computer systems. In fact, there are definite similarities between the functional organization of a computer system and an organization. A piece of software, often called an executive, allocates system resources (memory, files, channels, processor time, etc.) to subsystems (or programs) in charge of specific services. In turn, these sub-systems allocate their own resources, and control the execution of their tasks, and so on. There is some difference, however. When drawing a corporation chart, the president is normally placed at the top. When drawing a computer system functional diagram, the executive is normally at the bottom, but this is immaterial for our purposes.

A Tutorial on Protocols

When two or more computer systems are interconnected for a common purpose, they are said to make up a distributed system. As we have seen already by analogy, cooperation between these entities will normally be formalized into a number of protocols established at each layer of the system structure.

In the example of the whisky business, we assume that SIXOUNS and SNAPSE are similar organizations, with the same functional levels (or layers). Thus, it is rather straightforward to structure their interaction protocols by functional layers (marketing, production, etc.). Establishing a business relationship between a whisky corporation and a corn farm would require a different model, because these two organizations exhibit totally different structures. Perhaps the most sensible solution would be a contract between the farmer and the purchasing department of the whisky corporation. As we can infer, each case of business connection requires a specific solution, along with a specific set of protocols.

Even though most computer systems are structured along similar lines, there are a multiplicity of differences between systems from various manufacturers. It is also common that a single manufacturer produces several families of computer systems, with distinctive structural differences. In this case, computer systems are termed heterogeneous. Nevertheless, it is usually possible to work out some sensible interconnection scheme between any two computer systems.[17] One must decide how layers of each system are to be paired, and what protocols are to be used.

The development of an interconnection scheme between computers is often a significant investment, as it requires skilled computer professionals who know the intricacies of both systems. On the other hand, if most computer systems would contain very similar, or even identical layers, their interconnection could become a simple routine task. This is the case with computer systems of the same family. In this case they are termed homogeneous.

Many technical problems encountered in building distributed systems would disappear with homogeneous computer systems. But homogeneity cannot be a general solution. We can expect that there will always be a variety of suppliers. Even with a single supplier computer systems evolve, due to changes in technology and in user needs. On the other hand, as we have mentioned earlier, heterogeneity may introduce technical incompatibilities and require ad hoc adaptations. Is this dilemma irreconcilable? Experimental developments of distributed systems have led to a middle of the road approach, which aims to take full account of heterogeneity in the real world, while making interconnection more or less straightforward. This approach is presented below. We call it a reference architecture.

Reference Architecture

Let us assume we want to connect a cassette recorder and a radio set of arbitrary makes. We may study the electrical diagrams of both devices, and try to determine appropriate points where we might plan an electrical coupling, so that sounds received by the radio set become recorded on a cassette tape. It may be a little difficult to locate these appropriate coupling points, if the diagrams are not well-documented. Electrical signals may also require some adjustment to be performed by an intermediate black box, which we would have to build. Then, we may connect external wires with a soldering iron. If things do not work satisfactorily, we will need metering instruments and possibly some professional advice.

Interconnecting a cassette recorder and a radio set becomes much easier if both manufacturers have anticipated the need, and installed some plugging socket carrying well-defined electrical signals. In this case, we do not have to study the inner workings of a whole device. We only have to understand how to use the pins of the socket. These pins are the only peepholes into the device. We call them visibility points.

Computer systems are more complex than cassette recorders. Thus, interconnecting them requires more than the definition of a simple socket. As we mentioned earlier, computer systems are layered structures. Interconnecting layers requires the definition of protocols, which are the set of rules followed within each layer by the interactions between interconnected systems. Furthermore, protocol definitions assume some interactions between layers within each computer system. For example, in the SIXOUNS-SNAPSE agreement described earlier, protocols established for the marketing layer and the production layer assume some interactions between marketing and production within SIXOUNS and SNAPSE. We call an interface the set of rules followed by the interactions between two layers within a single computer system.

Once interfaces and protocols are defined between interconnected systems, they may work in cooperation to achieve a common purpose. They become a distributed system. If there exists a common set of definitions of all interfaces and protocols used in distributed systems, it is no longer necessary to study the inner workings of each system. Interfaces and protocols are analogous to sockets in the connection

A Tutorial on Protocols

of a cassette recorder and a radio set. They make interconnection much easier.

If the construction of distributed systems required that all interconnected systems follow rigidly a precise model defining systems in their entirety, then distributed systems would only comprise computer systems of the same family. Indeed, not many manufacturers, if any, would accept being constrained to such an extent in the design of their own products. But if the constraints are limited to interfaces and protocols necessary for building a model distributed system, it may become commercially attractive for many manufacturers to accept these constraints as the counterpart of a wider market. For easy reference to the set of interfaces and protocols of a distributed system model, we introduce the term reference architecture. Then, we may say that a system conforms to the reference architecture when it contains the visibility points of the distributed system model, i.e., the interfaces and protocols necessary for interconnection.

Basic Protocols Layers

Most business organizations are structured along similar lines. Below the president, one can find a number of departments: finance, marketing, production, purchasing, etc. There is no proof that this structure is the best, but in practice it seems to be satisfactory. The same rationale applies to layers of a distributed system. Even though the choice of appropriate protocol layers is not yet supported by a rigorous demonstration, a certain consensus has emerged from experience. The basic protocol layers consist of transmission, end to end transport, and application oriented functions. Among the latter, the virtual terminal and file transfer protocols are the most commonly used.

The tranmission layer includes functions pertaining to the transfer of data between geographically distinct locations. In the example of the whisky business, transmission is analogous to the transportation of whisky bottles. Whether transportation is carried out by truck, train, or helicopter, is immaterial, as long as whisky is delivered as ordered. Similarly, whether data are carried through wires, microwaves, or satellites is immaterial, as long as they are delivered without alteration at the proper destinations. New transmission techniques such as packet switching require some specific packaging of the data. This packaging is analogous to the practice of putting goods into containers for shipping. The transmission protocol consists of using properly any available transmission system.

The end to end transport layer (or transport for short) is in charge of checking the integrity of the transmission layer. Indeed, no transmission system is totally reliable, and some are far less than perfect. When data must travel through more than one transmission system (e.g., for international traffic), it is not unusual that each public carrier will blame the other in case of trouble. Thus, the transport protocol includes error control for assuring that data are delivered correctly, or else for having them corrected or retransmitted. In addition, the transport layer acts as a transportation bureau for the benefit of application layers.

It receives transport requests from other layers and makes the best use of available transmission systems. A similar function is carried out in the distillery. Packages for the retailers are prepared for shipment, but they are not delivered separately. Indeed, several packages may be destined to the same retailer, and delivery trucks follow established routes. Thus packages must be assigned to some trucks, depending on their route and their available capacity. The set of mechanisms carrying out the transport protocol in a distributed system is called a transport station.

The virtual terminal layer performs various adaptations between the characteristics of physical terminals and application programs. Indeed, real terminals available on the market may use different codes, keyboards, formats, etc. and new varieties of terminals are popping up constantly. Thus it is practically impossible for every application program to work properly with any terminal. For example, let us assume that an application program outputs series of lines having a length of 120 characters. Each character is binary coded in the specific code used by the computer, say EBCDIC. How can we use a display terminal working with ASCII coded characters and having a width of 80 character positions? We may define some adaptation rules:

- Translation of EBCDIC characters into ASCII characters
- Reformatting 120-character lines into 80-character lines.

There is no universal solution for shrinking lines. A first option is to select a subset of 80 positions out of 120. Some fields will be truncated, but this may be acceptable if they contain redundant information. A second option is to fold lines, i.e. each output line will be displayed as an 80-character line followed by a 40-character line. This presentation may be acceptable for program listings. A third option is to use 3 lines on the display for 2 output lines. This is acceptable for plain text.

A Tutorial on Protocols

Assuming that we can select the most suitable adaptation the application program works as if it were in communication with a 120-character wide EBCDIC terminal, which is not really there. This is why we call it a virtual terminal.

Actually, a virtual terminal does not have to be identical with an existing real terminal. Rather, it is a model of an ideal terminal with enough adjustable parameters so that it can look like a large number of real terminals. The virtual terminal protocol is the set of rules for setting the parameters and exchanging data with a virtual terminal.

The file-transfer layer is in charge of moving or copying files across computer systems. Ideally, in a distributed system, the location of a file should be immaterial. In practice, it is not so, for a number of reasons such as the following. Access is faster when file and application program are located in the same computer system. Storage costs may vary with systems. Security may not be good enough on certain systems. Thus, file transfer is frequently used in distributed systems. The file transfer protocol includes rules for opening and closing sender and receiver files, for data transfer, and for error recovery. If data translation is required, it may be performed by a specific user-provided routine, or it may be invoked as a protocol option.

In the above, we have introduced an intuitive presentation of the basic protocol layers of a distributed system. At this point, we would like to warn the reader that there is so far no universal agreement about the names and the number of layers. There is some analogy with geology. Where some people would identify only a sand layer, others would find a sand layer, a gravel layer, and a thin clay layer. No one is wrong, it is just a matter of defining sand.

Concepts in Distributed Systems

Some basic protocol concepts, such as error control, synchronization, etc., have been introduced informally earlier in this section. We shall have occasion to draw upon a few more conceptual terms in the body of this article. Thus, a short presentation, without rigorous definition, is included below to convey a sufficient understanding of these additional concepts, before proceeding to the next sections.

Process: A program running on a computer. It could be microcode running on a microprocessor or conventional software running on a mainframe processor. The term is most frequently used to distinguish among several running programs in time-sharing systems.

Resource: When a program is executed on a computer, it needs some area of memory, some processor time, and usually access to some input-output device, such as a teleprinter, a disk file, or a digital plotter. These equipments, or parts thereof, are called resources. By extension, any machinery made available to users is also called a resource. For example a data base, a compiler, a text editor, a graphics system.

Activity: One or several processes, along with appropriate resources, which work towards a common goal; e.g., a seat reservation application handling all the terminals of an airline company may be called an activity.

Entity: When information is exchanged within a distributed system, it is often immaterial whether we consider sources and destinations as processes, activities, files or terminals. Due to the heterogeneous nature of computer systems, a source of information might be called a process on one system, and a terminal on another system. In order to avoid misnomers, we use the generic term of entity to designate anything capable of sending or receiving information.

Correspondent: When two entities exchange information, each one is said to be the other's correspondent.

Port: Interconnected pieces of equipment are usually linked by a cable plugging onto a socket on each equipment. Each socket comprises a number of pins, which are linked to the other sockets pins through a wire in the cable. A pin has no function of its own. It is only a convenient device for speeding up interconnection, and for identifying signal channels. A port is analogous to a pin. It is a convenient device for establishing and identifying information channels between entities.

Association: An information channel between two ports or two entities. Also used in a general sense to mean any kind of relationship.

Liaison: A reliable association. Reliability is obtained by using a transport protocol providing for error and flow control.

To Access: To establish an association and exchange information with an entity.

Context: Some kinds of information transfers are composed of totally unrelated messages; e.g., when a point-of-sale terminal sends a message containing item number, quantity, credit card number, and salesman identification, there is no correlation with

A Tutorial on Protocols

the previous or the following message. Each message is self-contained and may be interpreted and processed independently. In other cases, messages are interdependent; e.g., when accessing a time sharing system from a terminal, a user might follow a question and answer procedure, giving successively his name, password, account number, etc. Clearly, messages sent from the terminal are no longer independent. They are interpreted in taking account of the previous ones. This means that the user logging process remembers something about the ongoing conversation. The information set aside during the logging in of a user is called a context. Usually, the context contents evolve with the arrival of new messages, or other events related to the user activity, such as occurrence of an error, or exhaustion of allocated processor time. In a dialogue, the turn (i.e., whose turn is it to send the next message) is part of the dialogue context.

Context Identifier: In the previous example of a time-sharing system, the user logging process may be unique and in charge of logging in all incoming users. Since a logging operation may take a significant time due to slow user reactions, it would be irritating for other incoming users to have to wait until the logging process is free. Therefore a typical logging process is constructed for handling several users in parallel, thence, several contexts in parallel. Obviously, messages arriving from the incoming users terminals must be sorted out and interpreted in the proper context. Since a user cannot be completely identified and validated until the logging procedure is completed, sorting out messages on the basis of the originating terminal appears to be the most practical method. Therefore, the terminal (or its physical line) number is used as a search key to find the context of an incoming message. The key relating a message to its context is called a context identifier.

Context Initialization. Necessarily, a context has to be brought to existence. Such an operation is called context initialization. The initialization may be static (i.e., it is built in the environment, e.g., at system generation time), or it may be dynamic. In this latter case, some specific messages are accepted out of context, and the result of their interpretation is to build an initial context. Normally, other specific messages are intended for the reinitialization of an existing context, and for terminating (i.e., destroying) an existing context. Typically, statically initialized contexts are protected from destruction, but they may be reinitialized.

Context Synchronization: When two entities communicate, they are related by an association, which they have to remember. Thus there must be an association context. On a single computer system an association context could be an area of memory shared by the two entities using it. In a distributed system memory sharing is rather cumbersome, because there is no memory directly addressable from geographically distant entities. In practice, either entity maintains its own view of the association context. We may say that an association context is distributed. Ideally, the two context views should always be identical, but this is impossible in practice due to message transit delays, not to mention occasional failures; e.g., let us assume that two entities called PING and PONG are engaged in a dialogue, and that it is PING's turn to send a message. In either context view some variable, say TURN, should read PING. As soon as PING has sent its next message, it updates its own context view so that TURN now reads PONG. However, the TURN of PONG's context will read PING until PONG has successfully received PING's message and updated its own context view. In the meantime the two views are inconsistent. Should PING's message get lost, the inconsistency becomes permanent (eventually, some external action is necessary to break the deadlock). Bringing both context's views into a consistent identical state is called context synchronization.

C-Name, and Local Name: Putting together a collection of heterogeneous computer systems is similar to an assembly of football teams, aircraft crews, church choirs, and police squads, coming from various countries. Probably only the French would understand when AMBROISE is called, and know who he is. These are several DAVID's and PAUL's. There is also someone from Netherland, but only he can pronounce his name. Clearly, there is a need for some way to call people that should be unambiguous, and understood by everyone. A possible solution could be to give each person a badge with a unique number. In a distributed system, such global identifiers are called C-names (common names) as opposed to local names used locally within each computer system.

Cycle: Let us assume a life insurance company produces 10,000 new policies every year (only one policy for one customer). Policies are numbered sequentially for 1 to 999,999. When all policy numbers are exhausted the numbering starts again with 1. Thus it takes one hundred years to go through one cycle of the policy numbering scheme. The number 125,786 is sufficient to identify uniquely a policy as long as it never happens that some customer keeps a life insurance for more than one hundred years. Communications protocols frequently use sequential numbering schemes to identify messages and detect losses and duplicates. The message numbering cycle must be long enough so as to outlive any numbered message. Otherwise it could happen that two different messages carry simultaneously the same identifier.

Evolution Trends

Techniques used in distributed systems are now submitted to seemingly conflicting forces. Stability and standardization are crucial for achieving world-wide interworking between all kinds of computer systems, terminals, data bases, etc. On the other hand computer and communication technologies are evolving at a swift pace. The advent of composite services which tightly integrate communications and data processing is tipping off the historical balance between the communications domain, mainly regulated and monopolistic, and the data processing domain, vigorously competitive and innovative.

Traditionally, standardization has progressed at a snail's pace, especially in the data processing field, where competing manufacturers never showed much enthusiasm for intersystem compatibility. A new spirit is now distinctly spreading. The demand for interconnection and world-wide access to computerized services has reached the point where standardization has become unavoidable. Standardization bodies are now engaged in a race with the requirements of distributed systems.

Two major standardization bodies are referred to in this report.

1) Comité Consultatif International Télégraphique et Téléphonique (CCITT) is practically a common carriers club. In most countries common carriers are government controlled monopolies often termed PTT's (Post, Telegraph, and Telephone).

2) International Standard Organization (ISO) is the forum of national standardization bodies, and a few international standardization associations.

Over the past fifteen years CCITT has produced interface standards intended for data transmission over analog telephone circuits. These standards are known as the V-series. Since 1972, new standards intended for digital networks are being worked out. They are known as the X-series. They include standards for packet switched services, even though some public packet networks use mostly analog circuits.

ISO has produced some standards for the control of data transmission over circuits. These standards are known as high-level data link control (HDLC).[43, 44] Objectives for the next few years are the definition of a standard architecture and basic protocols applicable to heterogeneous distributed systems.

Presumably, some new developments will appear that may obsolete the technical premises on which the existing standards are founded. Indeed, protocol performance and even the choice of protocols depends on environmental factors, which may vary substantially; e.g., the sizeable increase in bandwidth brought about by satellites or optical fibers may lead to a redistribution of data storage functions. The layered approach systematically adopted in distributed systems is essential for minimizing the impact of technical evolutions, since it attempts to concentrate and isolate within a specific layer all functions related to a particular system objective. Should this layer need to be redefined, the impact on other layers should be negligible. The difficulty resides in a careful definition of functionally independent system layers.

The experience acquired by a number of experimental and industrial distributed systems has uncovered some basic layers that are good candidates for a first cut at distributed systems standards. They are as follows: 1) data transmission, 2) end-to-end transport, 3) virtual terminal, 4) and file transfer.

For the Reader

This section was intended to be a primer to the rest of this report. We would like to bring again the attention of the reader to the approximate and intuitive nature of the material presented in this section. The following sections give more technical coverage, and at times redefine concepts from a different point of view, with another set of justifications.

ELEMENTS OF PROTOCOLS

The design of an automobile must take into account a minimum set of basic requirements: wheels, engine, body, steering, brakes, etc. The very same approach applies to protocols. They would be incomplete, or inappropriate unless they properly handled a number of basic functions. This section explains these basic functions, but does not describe any specific protocol.

Naming

Protocols involve communications and thus transfer of information. When a piece of information must be transferred from one domain into another, some unambiguous indication of the destination must be specified. In most circumstances, the sender does not perform the physical transfer of information by itself. Typically, the sender invokes some common underlying mechanism along with a set of parameters, including names or addresses, e.g., MOVE (data) FROM (A) TO (B).

A Tutorial on Protocols

In homogeneous systems, names such as A or B designate well-defined entities, e.g, core locations, files, processes physical or logical devices. Computer systems make use of various naming schemes, depending on the type of resource to which names are attached. Names may be of fixed or variable length, and follow specific conventions: alphabetic, numeric, special characters, and so on. Changing naming conventions in existing systems is an insuperable task because the conventions are usually so ingrained in the design that any change causes a major upheaval.

Thus networks of heterogeneous systems bring about new problems, as there is typically no commonly agreed naming scheme. A usual approach in such a situation is to superimpose a new network-wide scheme, while keeping existing ones inside each system. For example, telephone numbers are superimposed on persons or companies names. We shall see now how these common names can be created.

For convenience, we shall refer to names which exist in each computer system as local names. Network-wide names will be called C-names (for common names). There are a number of ways to construct C-names and relate them to local names. Three most frequent methods are discussed below: hierarchical concatenation, allocation, and mapping.

Hierarchical Concatenation: We imagine that each computer system has a set of local names $\{L_j\}$. To each set is assigned a unique C-name, $\{C_i\}$. The C-names, obviously must be assigned by global agreement to preserve their uniqueness. It is not ruled out that more than one C-name is assigned to a set $\{L_i\}$of local names, but the same C-name would not be assigned to two different local name sets because of the potential ambiguities this would introduce. The network name space is obtained by concatenating local names with a C-name, i.e., $\{< C_i \times L_i >\}$.

For example:

Local names in system *X*:

JOHN.FILE3.TEXT
TOM.LETTERS.MARY

Local names in system *Y*:

SYS COBOL BIN LINK.3
FACTORY PARTS

Local names in system *Z*:

53409
42121

The corresponding *C*-names might be

X TOM.LETTERS.MARY
X JOHN.FILE3.TEXT
Y SYS COBOL BIN LINK.3
Y FACTORY PARTS
Z 53409
Z 42121

This scheme is similar to the international telephone numbering plan in which phone numbers are obtained by concatenation of a country number with an internally assigned national number.

Allocation: A second method is to permanently allocate C-names to only a few processes, for example, the "logger" process in time-sharing systems which validates remote user access to the system.

Another set of C-names is allocated to each local system for dynamic association with local processes. At least one process in each local computer system is assigned a permanent C-name and the responsibility for dynamically assigning the others allocated to the system.

For example, suppose that C-names are simply integers in the range of 1 to 9999. Let us further assume that system A has been assigned C-names 6100-6199 for its use and system B has been assigned 4300-4399. Let us further suppose that there is a network access process at system B whose permanent C-name is 4301. Finally, let a process in system A, with C-name 6192 attempt to access a process in system B whose local name in system B is ZOOM (see Figure 1).

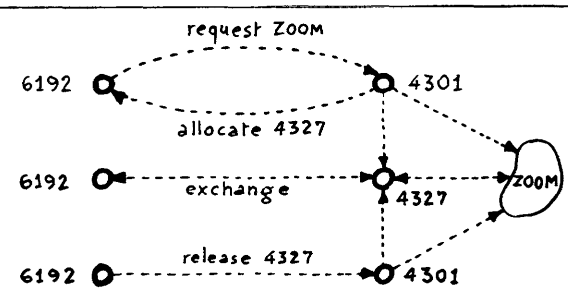


Figure 1. Allocation

Process 6192 would send a request to access local process ZOOM to the network access process in system B whose C-name is 4301. The network access process locates or creates process ZOOM within system B and allocates an unused C-name for it, say, 4327, and sends a reply to process 6192 which associates C-name 4327 with local process ZOOM.

Thereafter, process 6192 exchanges information with process 4327 until such exchanges are no longer necessary, at which time process 6192 reports to the network access process at system B that the C-name

A Tutorial on Protocols

4327 can be freed again. Alternatively, the ZOOM process could make this last report.

Plainly, the originating process (6192) needed a considerable amount of external knowledge to conduct this exchange. It needed to know that there was a network access process at system B and that its C-name was 4301. It also needed to know the local name, ZOOM, at the process it wanted to communicate with.

This strategy is closely related to the method used in the ARPANET initial connection protocol [31, 61] associated with the network control protocol (together they constitute the ARPANET transport protocol layer).

Mapping: An alternative to concatenation is to create a set of C-names $\{C_i\}$ and statically assign them among all the processes, files, and other local entities which must be accessed on a network-wide basis (see Figure 2). The C-names assigned are mapped by each local system into the associated local name. The relationship between local and global naming mechanisms is discussed further in this section. This strategy is used in the Cyclades network.[65]

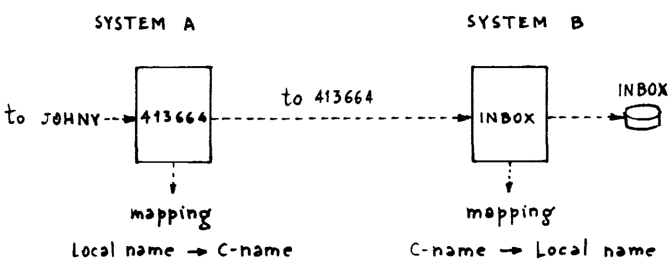


Figure 2. Mapping

Other strategies besides concatenation, allocation, and mapping are possible. Indeed, network designers have tended to invent ad hoc strategies that suit their implementation constraints rather than developing network naming techniques for every general use.

Discussion: Among the three methods described above, the hierarchical concatenation method is apparently the simplest one because a C-name may be parsed easily into a system name and local name, but it is only practical when local names are rather homogeneous. Otherwise, C-names take so many different formats that protocols become unwieldy and inefficient. In particular, the introduction of a new set of local names may require modifications in a number of existing network access protocols when the characteristics of the new set have not been anticipated.

The allocation method is favored by a number of operating system-minded people. It has the advantage of fitting within conventional single computer structures. Most operating systems in use over the past ten years were designed as geocentric objects centralizing all critical functions. Accessing entities is usually a multistep procedure beginning with a well-known log-in procedure. This vision of a rigidly partitioned universe is still taken for granted by a number of computer professionals.

An advantage of the allocation method is that users can access other computer systems almost as if there were no network. It is also contended that using a small number of C-names, for active entities only, saves overhead in network access machinery.

The deficiencies of this method are apparent from its advantages. Entities must be rigidly associated with specific computer systems and this puts a straitjacket on users who might prefer instead homogeneous access to all resources, with computer system boundaries fading out. In other words, the proper vision is a network resources rather than a network of computers. In addition, the allocation method is more complex in implementation as it requires the management of changing associations between C-names and local entities. Indeed, it implies a dynamic modification of the structure of the naming mechanism. Due to transit delays and fuzzy states associated with any distributed system, there appear transient conditions which require specific safeguards to prevent errors; e.g., a C-name may be accidentally released on one end and reassigned without the other correspondent being aware of this event. The elimination of inconsistent states takes additional mechanisms and delays the allocation of C-names.

The mapping method provides for a homogeneous name space for accessing any network entity. It is similar to a national telephone numbering plan. Mapping C-names onto local names is a matter of local implementation. This allows both permanent and temporary associations between entities. This method is therefore more general than the allocation method, which provides only temporary associations. Any desirable access control procedure can be triggered as part of the mapping machinery, depending on accessed entities or their correspondents.

Such a facility makes it possible to offer homogeneous network-wide access protocols for specific services, which may be available on different computers, e.g., compilers, text editors, help, distributed data bases. This flexibility results from the ability to design and implement specific access protocols independently from the others, because they are attached to well-defined entities. In principle, the allocation

A Tutorial on Protocols

method could offer an equivalent facility, but in practice the logger machinery is too sensitive software to be frequently adapted to new classes of users.

A criticism of the mapping method is that it takes overhead in scanning large tables of C-names when they are permanently associated with local names. However, this objection is not well substantiated. Indeed, a search is always necessary whether the key is a C-name or a local name, and there is no reason why searching by C-names should be less efficient. Space occupied by the C-name table is not critical, as it can be on secondary storage, like any file directory.

A few examples of the flexibility inherent in the mapping method are as follows.

- Users may choose local names according to their own symbolism, e.g., in their native tongue.
- Several local names may map into a single C-name, for the convenience of using at will abbreviated or full local names.
- Since local names do not have to be known remotely, there cannot occur any ambiguity when identical local names are used by several computer systems.
- A resource may be moved to a different computer system without its users knowing it. Actually, this capability is truly effective only when the transmission system can use C-names as message destinations. Examples of such networks are CIGALE, DCS, and ETHERNET.[66, 29, 58]

Group Names: When the total space of network names becomes very large, C-names may become very long and may generate overhead or inconvenience. A countermeasure is to partition the name-space into a hierarchy of subsets designated by group names.[66] Communications taking place between correspondents belonging to the same group require only short names. Communications crossing group boundaries require full length names. Thus there are two or more C-name formats, depending on the number of partition levels crossed by communications. This is similar to telephone dialing for local and long distance calls. However, the analogy is only partial, because a group does not carry any geographical connotation, as opposed to an area in the telephone system. Entities of a single group may be geographically scattered throughout. On the other hand, it is obviously possible to define certain groups as having geographical boundaries, but not necessarily restricted to a single piece of land.

For this technique to be effective, there must exist natural clusters of correspondents establishing a significant proportion of mutual exchanges. This is almost always the case when data communication is directly related to human users who are naturally clustered geographically (urban areas) or within a business structure (corporation, institution, affiliation).

Some refined techniques may be used for partitioning the network name space so that short names be also usable to communicate with neighbor groups.[69] When groups are not geographical, the notion of neighborhood has to be defined within some space, e.g., the space of integers representing C-names.

Port Names: When two entities communicate, they exchange information in the form of messages, following the rules of a certain protocol. It might appear sufficient that messages contain some destination field designating unambiguously the receiving entity. Thence, the name of the receiving entity could be used in the destination field of a message. However, this technique is not generally applied for the following reasons.

- Since computer systems are mainly heterogeneous, protocols are designed so as to make a minimum of assumptions about the characteristics of the communicating entities and of their environment. In particular, they tend to use names which are independent of any entity.
- Information exchanged between two entities may pertain to several independent channels (or associations).

  Consequently, the notion of port name is commonly used as a substitute for entity names. Port names present the following characteristics:

- They constitute a homogeneous network-wide name space. Therefore, they are C-names by definition.
- They are defined independently of any entity, but they may be assigned to any one.

In the naming techniques presented earlier (allocation and mapping) C-names are in effect port names. A port name may be allocated temporarily to an entity in the allocation method, or permanently in the mapping method.

Communications protocols are then defined as operating from port to port. Each port to port association constitutes an independent information channel. An entity may use several port names in order to maintain simultaneously several independent channels with another entity.

A Tutorial on Protocols

Port names may be created and distributed randomly to entities of a distributed system, but a possibly time-consuming table search would then be necessary to locate them. Customarily, some simple algorithm is used to relate port name and computer system, e.g., contiguous series of integers are allocated to each system, or a port name is formed by concatenation of a computer system C-name and a local integer. Many variants are acceptable, as long as resulting port names are homogeneous.

Association Names: When two entities exchange information, they make up an association. In some systems, any two entities may exchange information at any time without prior arrangement, e.g., the public mail, or a message switching system. Actually, if traffic is observed from the point of view of an end user, information exchanged is usually associated with an existing relationship which may be temporary or permanent, e.g., a business connection, a complaint, an acquaintance. In order to interpret correctly the arriving information, it is necessary to relate it to some other information kept into the receiver's memory. This information already in store is called a context. It may contain such items as the number of the last message correctly received, the number of errors encountered, the address of a buffer for the next message, etc. The context contents are specific to the communication protocol in effect between the corresponding entities.

Relating an arriving message to the appropriate context requires that some item of information contained in the message designate this context unambiguously. Such an information is called a context identifier. We will examine in the following some typical methods for choosing context identifiers. There are two cases.

1. Entities cannot establish and operate more than one association at a time (Figure 3). In this case, there is not really a need for a context identifier, because the entity can easily locate its only existing context, if any. However, in the absence of context identifier, any arriving message would be deemed related to the only existing association. Spurious messages might cause disruption. In order to tell valid messages from others, the protocol used with the association should include error control. Schemes such as carrying a password within each message, or encrypting messages may be used as a form of error control.

2. Entities may establish and operate multiple concurrent associations (Figure 4). In this case, messages must carry contexts identifier. Identification can be exchanged between entities when they set up an association. This is similar to the business practice of using file numbers to reference letters. It does not matter whether an entity uses its own or its correspondent's identifier when it sends a message, as long as this convention is defined by the protocol used with the association (Figure 5). An entity engaging simultaneously in several associations is in effect multiplexed. Logically this entity becomes visible as a set of independent instances. Thus context identifiers are extensions to the name of an entity for multiplexing purposes. For example, Figure 5, entities 3612 and 4230 have set up an association. The association context identifier in entity 3612 is 11. Thus messages sent by entity 4230 should carry 11 as context identifier. To that effect, the association context in entity 4230 contains 3612-11, i.e., the context identifier of the same association as seen by entity 3612. At this point it is clear that entity 4230 holds all the information it needs for labeling messages properly. The same holds true for the reverse direction. Contexts are built when an association is set up by exchanging context identifiers assigned by either entity.

Context identifiers can be assigned in the same way as C-names. For simplicity, such identifiers may be integers which are assigned in monotonically increasing order. If the maximum value of the context

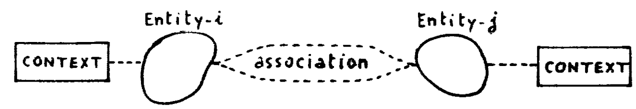


Figure 3. Single association

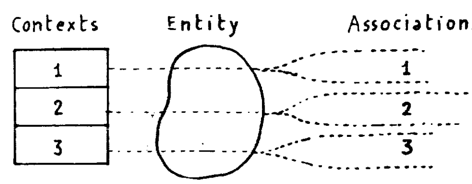


Figure 4. Multiple association

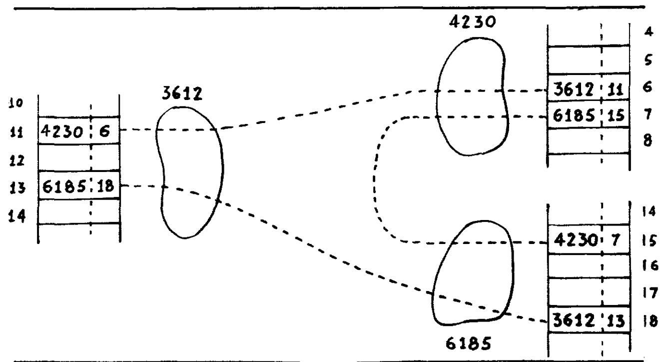


Figure 5. Context identifiers

A Tutorial on Protocols

identifier is large enough compared to the lifetime of associations and the rate at which associations are created, it should never happen that a new identifier is assigned while the same value is still assigned to an existing association context. However, this cannot always be guaranteed.

On the other hand, context identifiers might be assigned, released, and reused dynamically, but since these identifiers are shared by two processes, there may be a risk of transient conditions in which the interpretation of identifiers might become ambiguous (e.g., the crash of one computer system may cause it to lose all knowledge of previously assigned identifiers). The same problem can occur with the dynamic allocation of C-names.

In a distributed system, any dynamic allocation scheme comes with additional complexity and potential for trouble. The complexity results from the need to synchronize allocation or release. The potential for trouble derives from the vulnerability to crashes or erratic working in which the current allocation status may be corrupted. Due to the technical realities, static allocation schemes offer simpler and more robust solutions.

Assuming that entities are allocated C-names in a reliable way (e.g., by mapping), one can capitalize on this by using C-names as context identifiers (Figure 6). Since there is no additional information, there can only be one association between two entities, however, a single entity may still set up multiple associations with other distinct entities. This restriction is not as effective as it might appear. Indeed, the real limitation is only one association for a pair of C-names. Since port names act as substitutes for entities in communication protocols, one may allocate to an entity as many ports as it needs for setting up parallel associations with another entity. The reader may wonder at this point why this subsection is titled association names, while only context identifiers have been discussed. Indeed a context is the only embodiment of an association, and either entity sees one facet. An association is similar to a coin, only one side is visible. Either entity is only concerned with its own side. Thus context identifiers are sufficient to name associations for the participating entities.

It may be necessary to identify associations for external entities, e.g, a monitoring system. The network-wide C-name of an association is simply the pair of its context identifiers C-names.

Error Control

When information is moved over some distance, it must use transmission media, such as wires, optical

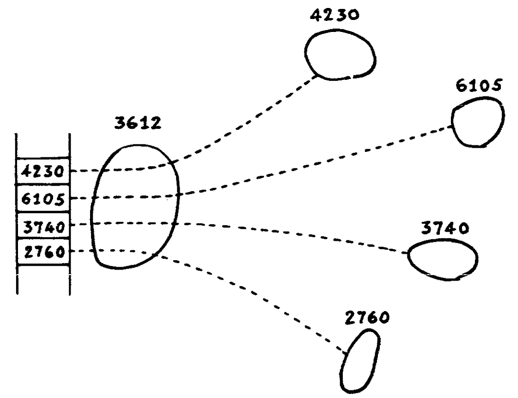


Figure 6. C-names as context identifiers

fibers, microwaves, satellites. Occasionally, noise, improper tuning, physical damage, or human interference, corrupt or shut off signals used for conveying messages. Thus mechanisms are necessary to detect and hopefully, recover from transmission errors. The same concern applies when transmission is effected through more complex systems, e.g., packet networks, which purport to carry out their own internal error control. End-to-end service quality may be considerably improved over that of a nonerror-controlled system, but residual errors are often encountered that can be traced to interference problems, installation changes, operation, management, hardware and software failures, etc.

Information exchanged between two entities appears as bit strings, sent and received in byte size blocks. Bit integrity may be assumed with a probability of one bit in error in 10^{10} to 10^{11} bits, which is customarily taken as satisfactory. This results from the properties of error detecting codes carried along with data.[79] Actually, these figures could be made as low as required by using any of several error detecting and correcting codes. However, there may still occur losses or duplication of blocks, which are not trapped by error detecting codes.

In order to assure block integrity, they are labeled with unique identifiers. As long as a receiver keeps track of correctly received identifiers, it can discard duplicates. Blocks received are acknowledged by returning their identifiers. After an agreed delay, called a time out, the sender retransmits unacknowledged blocks. This simple scheme, called Automatic Repeat Request (ARQ) is one of the most commonly used in communication protocols.[8]

For simplification and efficiency in handling identifiers, blocks are numbered sequentially by the sender. This gives the receiver the ability to detect missing blocks, and possibly speed up transmission by signalling the sender, rather than waiting for time out. Multiple blocks may be acknowledged with a single acknowledgment, if the assignment of identifiers is sequential.

Cycle: Identifiers cannot be unique indefinitely, however, since they would require an ever increasing number of bits. Practically an upper limit is set after which the old identifier numbers are reused. Two different blocks or acknowledgements carry the same identifier if they are separated by one cycle, or any integral number of cycles. Obviously, an undetected error would arise if a duplicate block created within cycle N-1 were received in lieu of the proper block transmitted within cycle N. The lifetime T·ID of an identifier is the sum of: maximum network transit delay for a block; maximum response time for the receiver to generate an acknowledgment; maximum network transit delay for its acknowledgment.

The cycle period must be longer than T·ID at the maximum block sending rate.[64] As a safeguard the minimum cycle period might be chosen to be 2T·ID.

Example: Let us assume that the maximum transit delay of a transmission network is 1·s, and that the receiver response time is negligible. This would yield T·ID = 2 s, thus the minimum cycle period is 4·s. If blocks may contain as little as 200 bits at a sending rate of 50 kbits/s, the cycle may contain up to 1000 identifiers. Thus the identifier field must be at least 10 bits. Communication paths involving satellites and high bit rates require either large identifier fields or larger block sizes.[77]

Recovery: A protocol may rely on an underlying service for automatic error recovery. As will be seen in the following, the virtual terminal protocol assumes that a transport protocol at the lower level provides for virtually error free transmission. However, some errors cannot be recovered in case of major system failure. Therefore, protocols must provide for checkpoints; i.e., the possibility to backtrack up to a point in the past from which activities may be restarted safely. Taking a checkpoint involves some context saving. This must be done at each end of the communication when the two corresponding activities are in synchronized states. This aspect will be covered in a later part of the report.

Flow Control

The amount of information transmitted by a particular source might exceed the capacity of the receiver or the capacity of an intermediate transmission system. A crude solution could be to discard traffic when it cannot be absorbed. This form of flow control may be used as a last resort to prevent deadlocks, but it cannot be considered as an efficient tool.

An objective of flow control is to maintain traffic within limits compatible with the amount of available resources.[68] This objective could be achieved simply by putting drastic limits on senders rate, but throughput degradation would be significant. In practice, flow control is a set of policies attempting to optimize the use of the system resources, while keeping data rates close to their nominal values. The whole subject is rather complex and is still an important research topic.[15,24,68,70]

This report focuses on two mechanisms commonly used as flow control building blocks. When two entities exchange information, they must be able to regulate each other's flow. Two basic mechanisms are used: stop and go, and credit.

Stop and Go: The receiver either accepts or rejects all traffic, except some short signalling messages. This technique is common with data link control procedures, such as BSC or HDLC.[44] The receiver has to take into account the transit delay necessary to carry stop and go signals to the sender. When this delay is large compared to the transmission delay of a block, a certain amount of traffic may still be flowing for a while. Also, when the receiver can resume traffic, the transit delay of a go signal translates into idle time. In short, this technique is simple in implementation, and works satisfactorily on terrestrial links. It becomes ineffectual for end-to-end flow control on satellite links, or across store-and-forward systems.[77].

Credits: We assume the sender is not allowed to transmit unless he has received from the receiver an indication of the amount of traffic that it can accept. These are known as credits. The use of credits completely protects the receiver to the extent that it can set aside enough resources for the credited traffic. However, the receiver may decide to give more credits than its available resources, in the case where transit delays or sender response times are long enough to schedule additional resources in the meantime.

By mutual agreement, an error control scheme may be used as a simple flow control scheme, if acknowledgments are taken to mean credit for one or several more blocks. This is the case in HDLC.[44]

Integrity: A flow control scheme might fail if signalling messages were not protected against loss and duplication: e.g., if a go signal were lost, the sender might be programmed to remain silent forever. If credits were duplicated, excessive traffic might be sent. Several techniques are used to protect the signalling messages.

A Tutorial on Protocols

- Individual signalling messages may be acknowledged.
- Signalling information may be embedded (piggybacked) in the control field of blocks sent in the reverse direction, when traffic is bidirectional.
- Redundant signalling may be used: losses are automatically recovered by a subsequent message and duplicates are harmless.

The credit scheme introduced initially in the CYCLADES[65] transport protocol[82] is of the latter type. It is now commonly used in other protocols.

Interrupts: A flow control share of concern is that it may work too well. If, on occasion a receiver stops accepting traffic for an exceptional period of time, the sender may want to send a command calling for some reinitialization or cancellation of the receiving activity. This could become impossible due to a blockage in the normal input channel at the receiver. A way around this impediment is a mechanism allowing special "interrupt" messages to bypass the flow control system in order to reach the receiver under such circumstances. This is equivalent to a secondary channel or out-of-band channel being available.

Interrupt messages may in turn generate excessive or unwarranted traffic and must be submitted to some form of flow control. Since the use of these control messages is assumed to contribute a small percentage of the total traffic, the flow control schemes adopted for interrupt messages may be rather simplistic, as throughput efficiency is not at a premium (e.g., the protocol may allow only one interrupt in transit between any two entities).

Synchronization

As is apparent from the previous sections, entities involved in communications must remember a certain number of parameters, or state variables, e.g., associations, message numbers, credits, delays, etc. As a whole, this information is termed protocol context. Context information is created at system generation, or as a result of explicit context setting commands, or on an incremental basis during traffic exchange. Due to transmission delays and response time of the machinery at each end of a communication, contexts evolve asynchronously. However, it is occasionally necessary to make sure that both sides of a protocol context are simultaneously in a well-defined state, e.g., at initialization, resetting checkpointing, closing, etc. This is termed synchronization. Indeed, one should keep in mind that contexts maintained at each end of an association are actually local views of the same machinery. Discrepancies between both views are due to message transit delays, and are normally transient. Putting a protocol into a quiescent state should automatically leave consistent contexts after a certain delay. Any persisting discrepancy between contexts is an error condition. Synchronization is intended to guarantee context consistency, or else to report an error.

The challenge in distributed systems is that there is no ideal observer capable of freezing each entity in a well-defined state. Only received messages can give clues about the correspondent state as it were at some point in the past. When error control context is not yet set up or has been destroyed, messages exchanged are not protected against loss and duplication. Therefore, context building must work even in adverse environment. Some Paradoxes:

1. No state can be held totally certain. Indeed, the knowledge of the states of a distributed system is acquired via messages, which require some transit delay. There is no guarantee that states are reported correctly or remain stable. One can only assume that components keep working properly with a certain probability.

2. Certainty is not a Boolean variable. What is certainty when it is distributed? E.g., two entities A and B must start working together. Each of them steps through preliminary tests, such as:

> Am I ready?
> Is the other ready?
> Does the other know that I am ready?

If no other entity is monitoring the situation, the decision to get started relies entirely on both A and B, but they may reach different conclusions (yes, no, perhaps), depending on their preliminary tests. Thus there may not be clear cut definition of readiness. However, it may be statistically satisfactory that either entity start working as soon as it finds both of them ready.

3. The last message cannot be critical. If it were, it had better be acknowledged, and it would no longer be the last one. But, why should one transmit a nonessential message? Some possible reasons are to improve efficiency, to introduce symmetry, to decrease uncertainty. Indeed, uncertainty may be decreased in the sense that certainty is only reached through incremental steps. Thus one more message may bring the entity into a desired state, instead of going to a recovery procedure.

This intriguing problem of uncertainty is ubiquitous in distributed systems. Not everything can ever be ascertained, but general solutions have been proposed

A Tutorial on Protocols

in the literature[63] for securing some critical states.

Example: Let us take a seemingly simple example. How can two corresponding activities determine that they both have terminated? A practical situation could be mutual data base updating. Let us call the two activities A and B. One might think of the state diagram in Figure 7.

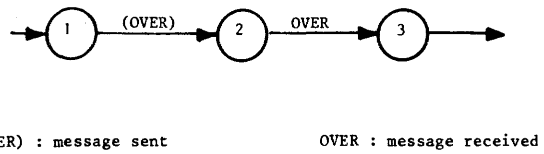


Figure 7. Partial synchronization

In state 1, A is active. When finished, it sends OVER to B and goes to state 2, where it waits for OVER from B. At this point A is certain that they both are finished, and goes to state 3; i.e., vanishes.

Unfortunately, this does not always work because B may be finished before A and send OVER first. The diagram in Figure 8 appears to be better, as it covers the two cases. It also coves a third case, when both A and B are finished first! Or so they think. Then both A and B send OVER and go to state 3. Both receive OVER from the other, and they both are finished. As long as every message sent is correctly delivered, the diagram in Figure 8 is foolproof; but this assumption can only be made if we can rely on an underlying transport service that is 100 percent reliable.

On the other hand, it is instructive to assume the underlying transport service is not 100 percent reliable and that, on occasion, messages may be lost, duplicated, or delivered out of sequence. In that case, the diagram in Figure 8 is not guaranteed to work either because A may get stalled in state 3, if the OVER message from B has been lost. What could A do? Going to state 4 and vanish is unsafe, since B might have not yet reached state 1 after all and it might need A's cooperation to complete its task. On the other hand, going back to state 1 and repeating OVER would not help, if B had already gone to state 4 and had vanished.

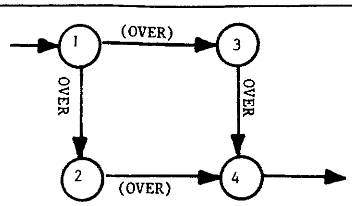


Figure 8. Simple synchronization

A possible solution is in Figure 9. No activity can vanish until the other has reached state 2. In case one activity gets hung up in state 2, it knows that the other has at least reached state 1, and possibly state 5. If any activity is hung up in state 1 or 2, it initializes a diagnostic procedure, which is specific to the environment. If there is no remedy, they may be uncertain about the other's state, but both of them have achieved their task completely.

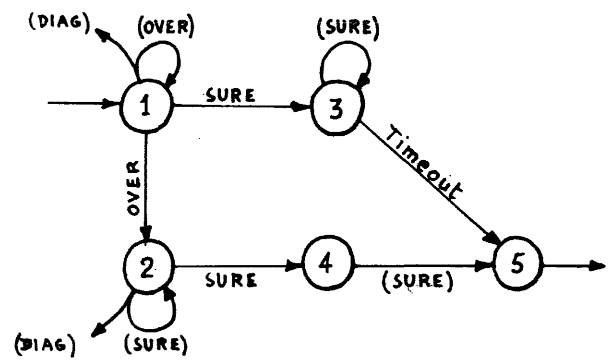


Figure 9. Safe synchronization

This example should not be taken as a paragon for all synchronization schemes. It only intends to illustrate the nature of the difficulties encountered and of possible solutions.

Guidelines: The complexity of synchronization introduces added logic and delays (message transit delays and timeouts). One should not introduce more complexity into a system than is strictly necessary. Before designing a scheme, one should state explicitly:

What is to be achieved?
What is to be prevented?
What interferences are expected or allowed from the environment?

When the same set of operations is repeated, e.g., context initialization, it may occur that messages of several successive instances get intermixed due to duplicates or variable transit delays. Additional protection is required for discriminating messages pertaining to each instance. Practical schemes introduce quiescent states or numbering cycles long enough for all stranded messages to be discarded naturally.[32] Nontrivial solutions have been proposed in the literature.[20,76] They in turn are sufficiently complex and constraining to create reliability problems. Simple and provable solutions are yet to be discovered.

The design of safe synchronization schemes is far from trivial. A typical approach consists of representing the logic at each end with state diagrams,

A Tutorial on Protocols

and then analyzing all possible conditions leading to stable states. Usual pitfalls are loops, deadlocks, or inconsistent stable states. This is illustrated in the literature.[4] It turns out that even simple diagrams are difficult to analyze exhaustively by hand due to the multiplicity of abnormal conditions. Automated validation techniques are now becoming available.[25]

DISTRIBUTED SYSTEM ARCHITECTURE

Motivation

Within human organizations as well as industrial processes, information is naturally distributed. To date, cost effectiveness considerations have generally led to centralized automatic data processing in order to take advantage of economies of scale and to minimize communications costs. The picture is drastically changing with the advent of microelectronics and packet switching. Technology now makes it possible to fragment data storage and processing capability according to the natural structure of human or industrial organizations and for information to flow along its natural paths within the structure (Figure 10).

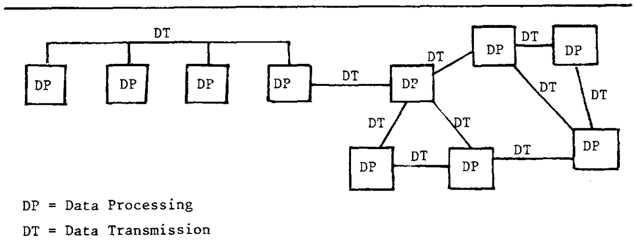


Figure 10. Physical structure of distributed systems

But technology alone is not enough, and a distributed system must be organized and structured in such a way that its different parts can cooperate efficiently and the whole system can evolve according to the evolution of the organization. This is the purpose of the system architecture, the organizational chart of a distributed system, which often reflects the organization of the parent enterprise itself (Figure 11).

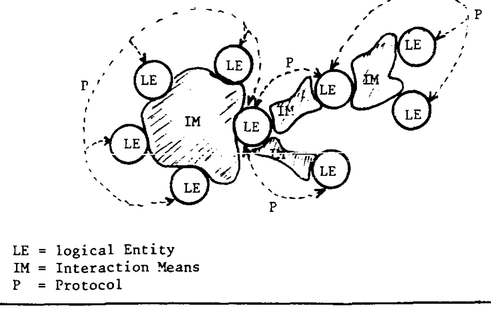


Figure 11. Logical structure of distributed systems

Up to this point, we have described a distributed system in terms of the relations (protocols and interfaces) between its various parts. There are various ways to structure a distributed system to achieve these relations. In this section, we emphasize the relational aspects of a distributed system and review the basic structuring techniques. By the term reference architecture we refer to the set of relations (and their definitions) to which the various parts of a distributed system must conform. Each element of the system can only see an external manifestation of these relations. It cannot see their internal implementation or structure within the other elements. We refer to this filtered view of the system as its visibility relative to that part, or simply as its visibility. In this section, we introduce the concept of visibility in a reference architecture and its application.

A central aspect of distributed systems is the structure of interactions between logical entities, just as the structure of relations between divisions, departments, and individuals is central to an organization.

Interactions between logical entities make sense only if they contribute to a common objective. A protocol defines for each entity how it must interact with the other entities in order to achieve its objective. Commonly, the name of a protocol is matched to the function carried out by interacting entities (e.g., transport protocol, file transfer protocol).

Structuring Techniques

Almost any distributed system architecture is based on a small set of simple structuring techniques, viz, multiplexing, switching, wrapping, cascading, and assembling, which are described in this portion of the report.

Multiplexing: In most cases, resources in distributed systems must be shared between different activities. For instance, in a packet switching network, lines, node processors, and memory are shared between all users. Rather than requesting all activities to coordinate themselves directly when sharing resources, the network can provide statistical multiplexing mechanisms which take care of allocating resources to activities dynamically when they need them. (The allocation itself can be organized as a distributed activity.) This technique, already used in any multiprogrammed system permits activities to proceed as if they were independent from each other (see Figure 12). Each activity interacts only with the network and its multiplexing mechanisms.

Switching: Interaction is usually restricted to pairs of entities. More generally, however, the action of one entity may be directed to a subset of the entities with which it may interact. Therefore, each action must be

A Tutorial on Protocols

associated with a switching parameter which can be interpreted by the underlying mediating system (interaction means), in order to route the action to the proper destination(s) (see Figure 13). In most cases the action is directed to one entity only, i.e., each entity must be associated with one address. Other types of switching may also be used such as "all entities" (broadcast) or "any entities with such and such characteristics" (associative switching). Switching permits us to define independently interactions within subsets of entities.[54]

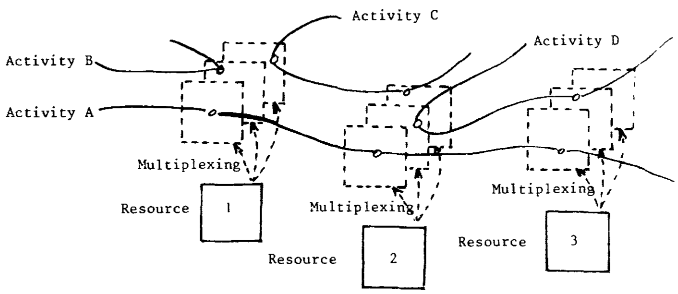


Figure 12. Multiplexing resources between activities

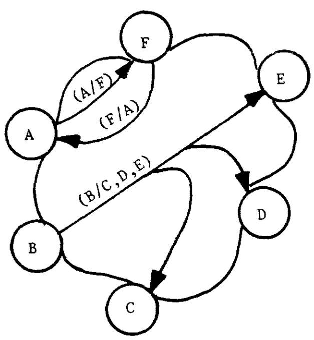


Figure 13. Addressing actions

Wrapping: A pair of entities interacting according to a specific protocol may be used to provide other entities with a more powerful interaction means (see Figure 14). When applied recurringly, a hierarchical structure is obtained and we use the term wrapping to indicate how it was created (see Figure 15). The essential feature of wrapping is that each protocol layer communicates with its counterpart at the other end of the interaction means. Of course switching and multiplexing techniques can be used in conjunction with wrapping in order to provide interaction means between any pair of higher level entities, as well as alternate paths between them.

Cascading: Cascading consists of forming a linear string of entities, each one interacting only with its neighbors. The resulting cascade can in turn be used as an interaction means (see Figure 16). A common form of cascade is found between successive nodes in a packet switching network where all protocols between neighbors are identical in the recurring cascade. Again switching and multiplexing techniques can be

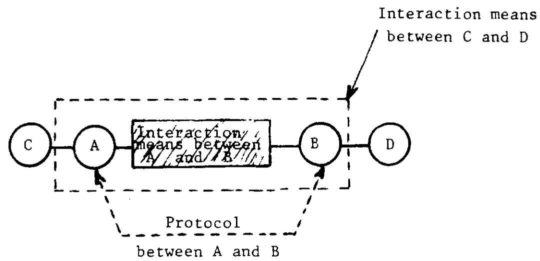


Figure 14. Wrapping assembly

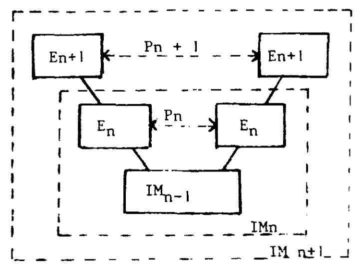


Figure 15. Hierarchical assembly

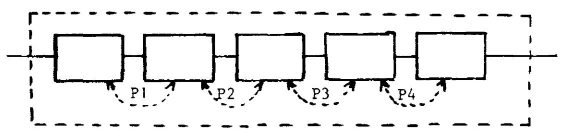


Figure 16. Cascading assembly

used in conjunction with cascading in order to form multipath networks.

Assembling-Star assembly: Entities can be assembled into a star configuration where a central entity interacts individually with satellite entities (see Figure 17). There is no direct interaction between satellite entities. The protocol of the star is decomposed into protocols (usually identical) between each satellite and the central entity, which also performs the coupling of all these protocols.

Meshed assembly: Entities can be assembled into a meshed network where each entity interacts with its neighbors (see Figure 18). The protocol of the meshed network is defined by protocols (usually identical)

A Tutorial on Protocols

between neighbors and by their coupling inside each entity. This assembling technique is frequently used to organize route adaptation as a distributed activity in packet-switching networks.

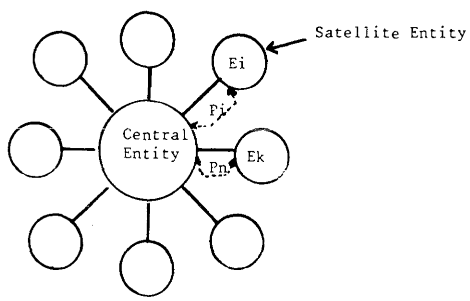


Figure 17. Star assembly

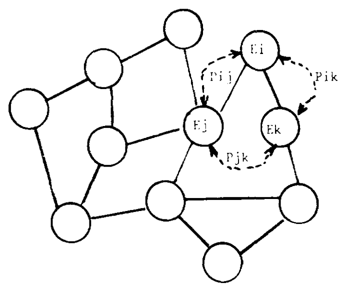


Figure 18. Meshed assembly

Reference Architecture and Visibility

Defining the architecture of a distributed system consists essentially of specifying interactions between logical entities which compose the system, i.e., their protocols. The design and implementation of any specific entity is only constrained to behave externally according to the system protocols. Any other functional or technical aspects of its design or implementation are independent choices which will not be explicitly visible to the other entities, except perhaps as performance variations. This essential characteristic of a reference architecture is its visibility to the other entities. The architecture serves as a reference for relations between entities.

Visibility is essential to the architecture of heterogeneous networks.[84,85] It allows systems from different manufacturers to interwork according to common rules, even though they may have different internal structures.

Protocols between systems are exercised only at physical or logical interfaces between them. The only points at which the physical embodiment of the protocols are visible are at these interfaces. The architecture must include the definition of relations through such interfaces.

These interfaces (or visibility points) must be properly located within each equipment for them to serve their original purpose. For example, if one manufacturer's equipment is to be replaced by another's, the physical interface to that equipment should be identical with a visibility point. The interfaces at which a system conforms with the architecture are known as reference interfaces (see Figure 19). The reference interfaces occur only at the boundaries of protocol layers and, when used in conjunction with the distributed system protocols, form the visibility points of the reference architecture.

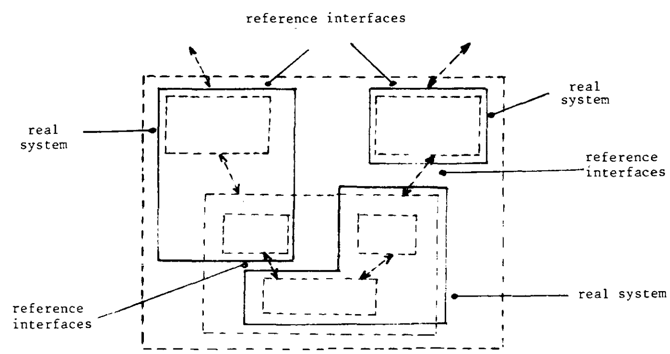


Figure 19. Reference architecture and reference interfaces

Layered Architecture

The concept of visibility can in turn be applied recursively to the reference architecture itself in order to provide maximum flexibility in its own evolution; when a set of entities are assembled using wrapping, cascading, or another technique. The resulting entity can then be considered as a whole; i.e., without any assumption on its internal structure. This entity will in turn be assembled with other entities into a higher level entity, etc.

The resulting architecture is often referred to as "layered architecture;" at each step of its assembly process a new higher layer is added which views the former (lower) layers as a single entity without any assumption about its internal structure. Under this assumption, successive layers are said to be independent. This technique is very similar to the one used in structured programming where each module is defined functionally, as seen from outside, without regard to its internal structure.

Computer Network Architecture

The concepts and techniques developed in the previous sections are applicable to any distributed

A Tutorial on Protocols

system (e.g., small network of identical microprocessors, an organization, as well as a large heterogeneous resource sharing computer network). In the specific case of computer networks, the constraint of preexisting computers and transmission systems must be taken into account. Most computer networks make use of the same basic reference architecture with three major protocol layers: data-transmission, end-to-end transport control, and application control.[2, 7, 16, 19, 26, 53, 55, 58, 65, 71]

Data Transmission: Data transmission facilities are often provided by common carriers or PTT's. These are physical systems which will interact with data processing systems in order to let them communicate. The reference architecture for computer networks must therefore identify a data-transmission entity used as an interaction means between data-processing entities (Figure 20).

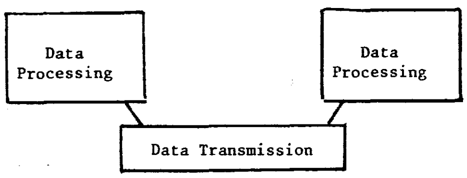


Figure 20. Data processing wraps data transmission

The data-transmission part of the network may also use computers internally, as in packet switching, but these computers act as intermediate nodes (i.e., they store and forward packets to the next node).

In the reference architecture, the data-transmission entity ends where the end data-processing systems begin. In case a computer is used both as an end data-processing system and as an intermediate node for other data-processing systems, the border between data-transmission and data-processing is located inside the computer itself (see Figure 21).

End-to-End Transport Control: In most computer systems, several independent processes run simultaneously. In that environment, they happen to share local resources. Cooperation within a distributed network takes place between individual processes rather than between systems as a whole. The network conventions must cater to interprocess communication (i.e., define how processes access each other and how information exchanged between processes is passed from one computer to another through the data-transmission system, how errors are detected and corrected, etc.).

The reference architecture must therefore include a transport layer, on top of the data-transmission entity. Thus each end data-processing system will include a transport control entity, often referred to as a transport station. Transport stations interact through the data-transmission entity, according to an end-to-end transport protocol, (see Figure 22). This assembling is of the "wrapping" type.

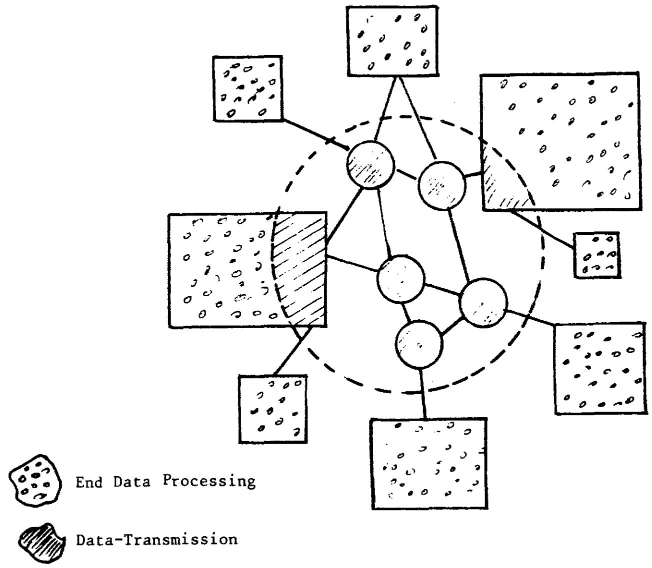


Figure 21. Border between data transmission and end data processing

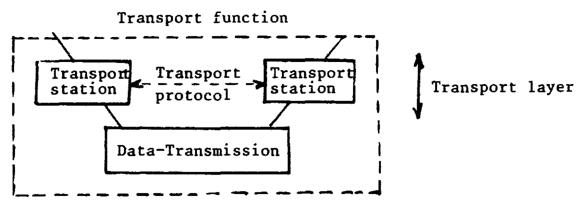


Figure 22. Transport layer

The rationale for identifying this transport component is that its function is not specific to one application but can be used by any of them. In other words, all kinds of applications can share the transport components. They add their own sets of conventions (or protocols) to define the meaning as well as the format of data they exchange through the transport function.

As long as the transport layer has not been defined and built, a network cannot exist, since processes cannot communicate, but the transport layer in itself is not sufficient. It just allows resources to get connected and must be augmented with conventions on the use of resources. This type of convention might be referred to as application control.

Application Control: It is clear that conventions for

A Tutorial on Protocols

the use of all resources within a network need not, and cannot, be defined and implemented at one time. It is also obvious that some resources are not widely shared, at least at the beginning of the network life. It can be observed that the pressure from users for an application control protocol to be defined is strictly related to the actual amount of sharing of the corresponding resource.

It seems to be a general rule today that the initial requirement on any large network is for a set of conventions allowing people to use the same terminal to access a variety of services. Thus the next layer in the basic reference architecture (above the transport layer) is a terminal layer, which provides higher layers (application programs and human users) with a terminal oriented communication facility (Figure 23). Terminal-like entities interacting through the transport entity are often referred to as virtual terminals while the protocol between them is termed a virtual terminal protocol.

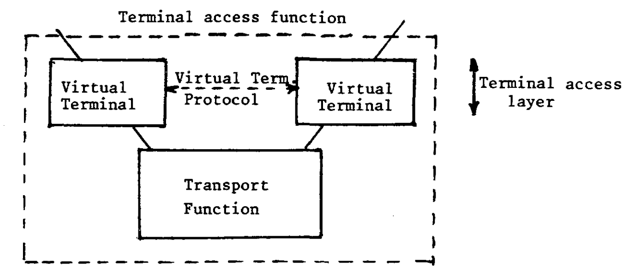


Figure 23. Terminal access layer

Later on, when the users start using the network, they are gradually led to compare equivalent services provided by different computer systems and tempted to use the best of each. The pressure then builds up for further compatibility:

1) Users want to move their data from one computer system to another (e.g., they may want to use a text editor on one system to prepare a program and ship it to another system for execution). This leads to define extensions of the basic architecture to include file transfer or more generally file management.

2) Users are frustrated with meaningless differences between control languages of each system (e.g., different log-in procedures) and this leads to standardizing the most widely shared procedures between human users and data processing systems.

As sharing develops, extensions to the reference architecture and corresponding high-level protocols will be defined (e.g., for distributed data-base management, etc.). High-level protocols are addressed in the paper by Sproul and Cohen in this issue.[72]

THE DESIGN OF PROTOCOLS

This section of the report attempts to characterize protocols in the various layers of the basic reference architecture. We develop one particular model of network operation for ease in exposition.

Data Transmission

Many kinds of data-transmission facilities are used for computer networks (telephone lines, satellite links, etc.) and the most widely used computer networks nowadays are based on packet switching. A packet switching network is basically a meshed assembly of minicomputers, called nodes, interacting through transmission lines. The major activity of a packet switching network is to transmit packets from a source to a destination. Routing tables in each node indicate which line is the next step to each destination. In other words, routing tables determine for each source-destination pair which cascade of nodes will be involved (see Figure 24). Each node is shared (multiprogrammed) between a number of entities

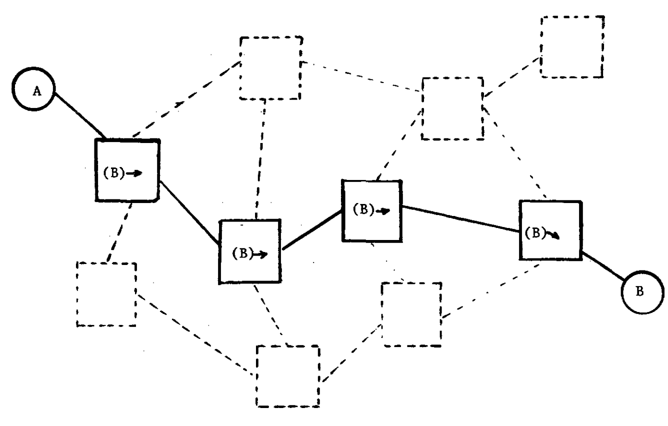


Figure 24. Cascade of nodes built by routing tables

(processes); therefore, the various cascades may be considered to be independent.

Along a cascade, packets are transmitted from node to node (store and forward), using some line control procedure. Route adaptation is usually organized as a meshed-assembly with one entity in each node (see Figure 25). Neighboring entities exchange information they have on the state of the network components (lines and nodes). Each entity integrates the information from its neighbors and passes the result to all neighbors.

A Tutorial on Protocols

This meshed distributed activity is responsible for determining best routes and updating routing tables accordingly; thus, reducing the problem of reliable packet switching to "packet forwarding" along dynamically predetermined cascades. A similar propagation mechanism may also be used for congestion control as in CIGALE.[66,68] If a node or a line fails, route adaptation will immediately determine a new route, but some packets may have been lost. Error detection and recovery is usually performed end to end on top of the cascade (see Figure 26), either within the data-transmission layer or in the transport layer.[2,26,28,39,58,65,75,10,82]

Broadcast packet networks exhibit another kind of distributed organization.[1,29,58]

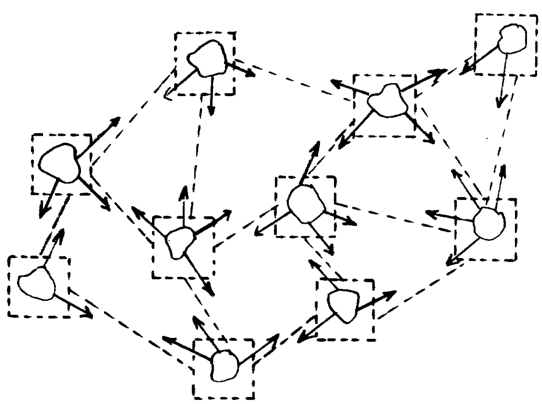


Figure 25. Meshed assembly of route adaptation entities

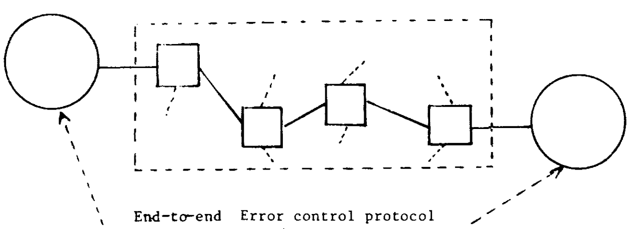


Figure 26. End-to-end error control performed on top of a cascade

Transport Protocol

The transport layer must provide the higher level protocols with a reliable interprocess communication facility.

The transport protocol comprises a set of tools, along with operating procedures, which allow processes to set up associations, and transfer blocks of information from one process domain to another's.

Naming—Ports and Liaisons: In order to communicate, processes must be given C-names by the transport layer. A most common scheme is to use hierarchical port names composed of a transport station number and a local port number.

{Port Name} : : ={Transport Station #}{Local Port #}.

In each system, processes/resources are mapped locally into port names (see Figure 27).

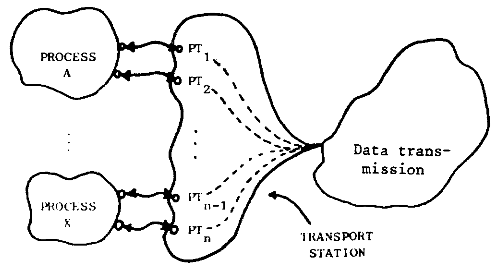


Figure 27. Port names and process names

When two processes engage in a conversation, an association between ports called liaison must be identified. The simplest way to identify a liaison is to use a pair of port names.[13,26,28,82] This allows a process to participate simultaneously in several associations (see Figure 28). Another method[9] consists of restricting ports to a single association and thus uses a single-port name to identify the association. A contention problem arises when several processes want to access the same (sharable) resource. Additional protocol machinery[4] must then be provided.[31,36,61]

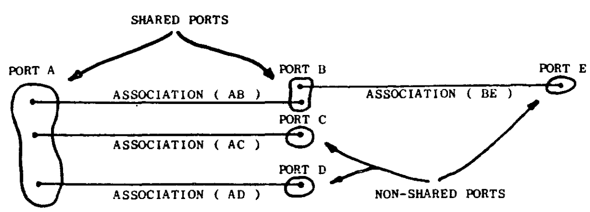


Figure 28. Ports and association

It is generally agreed that switching should form the basic sublayer in the transport layer, with error control, flow control, etc., on top (see Figure 29). This permits the transport protocol to be decomposed into protocols between pairs of ports and reduces its complexity. In this case, each packet is related to an association and handled at each end in

A Tutorial on Protocols

the context of this association. For transaction-oriented applications, the association between ports may be restricted to the exchange of one message only. It is then simply identified with the pair of port names. This type of association has been included in. [13,28,42,82] It is often referred to as Lettergram or Datagram facility.

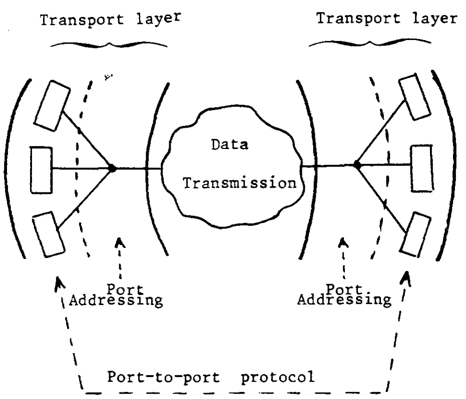


Figure 29. Port-to-port protocol on top of addressing

Messages—Letters and Telegrams: The transport layer usually provides processes with transportation of two kinds of messages. [13,14,28,42,82] 1) variable length blocks, often called letters, intended for regular exchange of data.[82] 2) Small pieces of data (e.g., 8 bits), often called telegrams intended to transfer interrupt-like signals, and thus not submitted to flow control on letters.

Letters may be longer than packets accepted by the data-transmission facility, and thus, transport protocols can include fragmentation of letters into packets and their subsequent reassembly upon arrival in the transport station of the destination data-processing system. Fragments are usually sequentially numbered within the letter[82] (i.e., fragments refer to letters), or within the liaison (i.e., letters refer to fragments) as in[14] (see Figure 30).

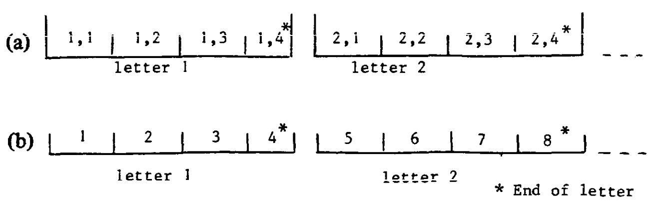


Figure 30. Fragments numbering. (a) Fragments numbering within letters (b) Fragments numbering within a liaison

Error Control: End-to-end error control on letter-grams and telegrams is usually based on individual acknowledgments, while liaisons make use of more refined schemes. In any case, pieces of information must be uniquely identified with cyclicly reused names. It is often considered that the (0, 1) numbering is sufficient for Telegrams, while letters or fragments require a longer cycle, e.g., 16 bits![13,14]

Flow Control: Most transport protocols control flow with a credit scheme. Credits indicate the number of letters[42,82], or fragments[14] which the receiver is prepared to accept beyond the last acknowledged letter or fragment (see Figure 31). This scheme is preferred to a stop and go scheme which is ineffectual across packet switching nets because of the round-trip delay. Another advantage of the credit scheme is that the corresponding control information (ACK #, Credit) is self-contained; i.e., it does not refer to previous control information. The information in each new (ACK #, Credit) message replaces any preceding information, thus providing automatic error control in case of loss or duplication.[68]

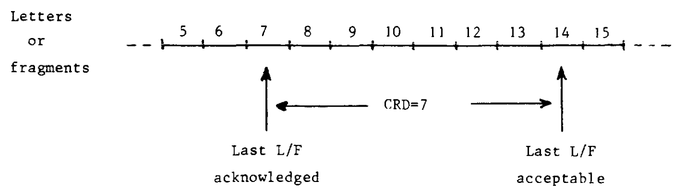


Figure 31. Flow control with credit

Another scheme used in[9], where each flow control message refers to the preceding ones (incremental credits) has proved to cause deadlocks in case of loss.

Synchronization: Liaisons must be initialized, possibly re-initialized, and terminated. To this effect, symmetric protocols are often used (see Figure 32) in order to avoid contention when both ends happen to initiate synchronization at the same time.

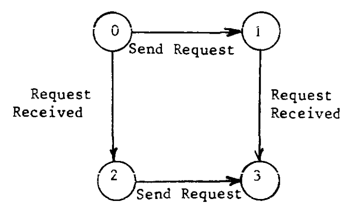


Figure 32. Symmetric rendez-vouz synchronization

Virtual Terminal Control

The virtual terminal layer provides higher layers in

A Tutorial on Protocols

the basic reference architecture, with a terminal oriented communication facility.[23,27,41,59,81] The virtual terminal protocol defines a set of procedures and messages formats which may be interpreted for controlling a large variety of real terminals.

Naming: Terminal oriented communications take place between pairs of correspondents (terminal-program, terminal-terminal, or program-program). It can thus be nested in the protocol of an association, e.g., a liaison. Hence, the virtual terminal protocol can use ports and liaisons names provided by the transport layer. This permits the virtual terminal protocol to be defined as a protocol between a pair of virtual terminals (see Figure 33).

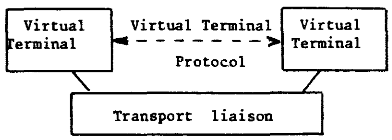


Figure 33. Protocol between a pair of virtual terminals

Error Control: Virtual terminal belongs to end data-processing systems, where transport stations are also located and the virtual terminal layer relies upon end-to-end transport control performed by the transport layer.

Data Representation: Terminal-oriented communications are based on text representation, i.e., a data structure made of pages, lines, and character positions which may be filled with characters. The data structure may be more complex and be associated with access control functions (e.g., protected zones on data-entry terminals). The virtual terminal protocol must incorporate conventions for data representation and access to the data structure, such as character code, addressing, etc.

Dialogue Control: The data structure, e.g. for display on the terminal, can be accessed by both ends. The virtual terminal protocol must therefore comprise mechanisms to coordinate accesses from both correspondents. A widely used scenario consists of using some sort of semaphore representing the right to write in the data-structure, usually called the "turn," which is passed to the other end when the owner of the turn has finished writing. This type of dialogue is often referred to as "alternating dialogue."

Access to the data-structure is normally done on the basis of messages[27,41,59,81], rather than characters.[23] This reflects the evolution of terminals from a character at a time to message mode, the former mode being totally inadequate on store and forward networks.

Synchronization: Most virtual terminal protocols include some kind of purge mechanism which permits the data-structure to be reinitialized without reinitializing the whole protocol (equivalent to a clear display operation).[41,59] This type of mechanism uses telegrams to initiate the operation even if letters are blocked due to flow control in the transport layer.

Negotiation of Options: A characteristic of terminals is the variety of functions they may provide. For this reason, any virtual terminal protocol includes a negotiation mechanism which permits both correspondents to agree on the virtual terminal characteristics which they are going to use.[41,59] In most virtual terminal protocols conventions on default characteristics permit negotiation to be avoided in simple cases.

File Transfer/Management Protocol

The file transfer protocol should provide a set of tools intended for handling complete files across computer systems. Example of functions to be performed are: finding, copying, deleting, renaming, or creating.

The potential complexity of file transfer/management is due to heterogeneity of file names, structure, access procedures, and data representation. Examples of such protocols can be found in[34,60]

Another characteristic of file transfer protocols is that they must often include their own error recovery procedures on top of the transport layer. The reason for this is that file transfer is an automatic process which cannot rely on the intervention of a human operator. In addition, the duration of a file transfer may be such that the probability of failure of one computer system is not negligible. The recovery procedure is based on periodical check-pointing involving end-to-end synchronization. In case of failure, the transfer is restarted from the last check-point in the transfer.

PROTOCOL AND PERFORMANCE INTERPLAY

Good systems, like good meals, do not come out simply by putting together good ingredients. Choosing a proper set of tools and parameters depends on a number of factors.

- Efficiency—What is to be optimized? Bandwidth, throughput, transit delay?
- Complexity—Do we want to reduce implementation costs, core size, execution time?

A Tutorial on Protocols

- Service quality—What error rates, delay variations, manual interventions, are tolerable?
- Service parameters—What are the requirements for message length, traffic density, response time, number of terminals?
- Transmission parameters—What are line speeds, error rates, propagation delays?
- Overall architecture—What is the proper distribution of functions within system layers?

These are just a sampling of considerations which must be taken into account in the design of a network architecture. In practice, a number of parameters or system components are given as constraints. In a sense, this is usually useful, as it narrows the set of options, and often simplifies design decisions.

The reader should not expect to find in this report off-the-shelf recipes for building protocols. Rather, we shall attempt to point out how typical protocol ingredients interplay with one another, and how they relate to some of the factors mentioned earlier.

Acknowledgements

Simple transmission protocols use an alternation scheme:

Sender		Receiver
Send block	⟶	Receive block
Receive ACK	⟵	Send ACK
etc.		etc.

During the acknowledgment round-trip time, the sender is blocked. If this time is relatively very short, throughput degradation may be negligible. This is the case with unidirectional data flows on terrestrial circuits. However, data flows are also bidirectional. ACK's compete with data blocks in the reverse direction. Even if ACK's are given priority, they must wait for the end of transmission of the current block, Hence, delayed ACK's translate into lost bandwidth. How much is lost depends on the characteristics of the reverse traffic, but reverse traffic also needs ACK's which are again delayed by competing data blocks. Finally, simple analysis shows that both traffics become paced with one another[30], as shown in Figure 34. Block Bi is acknowledged by Ai, while block B'i, is acknowledged by A'i.

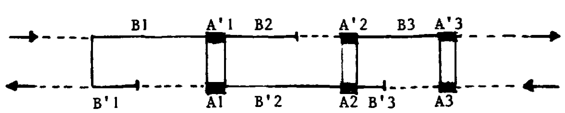


Figure 34. Bidirectional traffic.

A remedy could be separate acknowledgment channels. This would decouple the two opposite data flows, at the cost of added hardware, software, and bandwidth. Some communication systems used this scheme in the form of bidirectional modems, which provide a low bandwidth return channel.

A more frequently used technique for improving throughput is to use a protocol allowing transmission of a number of blocks before waiting for ACK. This is termed anticipation, and is used in the more refined error control schemes.

Sender		Receiver
Send block B1	⟶	Receive block B1
Send block B2	⟶	Receive block B2
Send block B3	⟶	Receive block B3
Send block B4	⟶	Receive block B4
Receive ACK A4	⟵	Send ACK A4.

Blocks are numbered, and acknowledging block B4 implies acknowledgments of all previously sent blocks. There is a significant gain in bandwidth at the cost of larger buffering on the sender side. Indeed the sender must keep copies of all unacknowledged blocks in order to be able to send them again if they have not been received correctly.

There is a limit to the number of blocks that can be queued, waiting for ACK's. Sender and receiver may agree on a maximum, but in any case it must be lower than the numbering cycle to avoid ambiguities.

Long blocks transmitted in one direction may still delay ACK's to the point of exhausting the maximum amount of block numbers in the other direction when the blocks are short.[22] A countermeasure could be to increase the numbering cycle, but more buffers would be required for unacknowledged blocks.

Piggybacking ACK's (i.e., embedding ACK's in the control field of data blocks in the reverse direction) yields additional throughput, as it saves the overhead of sending distinct ACK messages. However, the disadvantage is increased ACK delay, since the ACK field cannot be processed until a complete block has been received and validated. This effect is more pronounced when block sizes on each direction are heavily unbalanced[33, 37] It has also been proposed to place ACK's in a trailer rather than in the header of a block[37]

Retransmission

Transmission errors are usually corrected by retransmission, strategies may be selective or sequential.[18] Selective retransmission is executed on

A Tutorial on Protocols

receiver's requests for individual blocks. The advantage is to keep retransmission to a strict minimum. On the other hand, the receiver must manage blocks pending predecessors and deliver them to processing in a correct sequence. This strategy is suitable with random errors when sequential retransmission is too costly. Satellite transmission paths fall into this category.[77]

Sequential retransmission may be executed on either side with requests starting back from the first erroneous block and proceeding sequentially. Due to the delay introduced in triggering retransmission, a number of blocks following an erroneous block may have been correctly received, but will be retransmitted. An advantage of sequential retransmission is that it is straightforward to implement. It is suited to bursty errors, which may effect several blocks in a row, when ACK round-trip delay is small. If round-trip delay and numbering cycles are large, a substantial amount of correctly received traffic may have to be retransmitted. This would be acceptable with low block error rates say less than 10^{-5} . Otherwise, lost bandwidth is important. This case applies to terrestrial lines (very short propagation delay) and also to letters exchanged over a liaison across a packet network (very low error rate).

Broadcast channel (radio or satellite) utilization depends critically on retransmission strategies. The simplest one is the ALOHA[1,3] scheme, in which colliding blocks are retransmitted after a random delay. The maximum utilization of the channel is 18 percent. A number of more sophisticated schemes have been analyzed, and new ones are generated every year.[7] Under a number of assumptions, and at the cost of increasing protocol complexity, maximum channel utilization may reach or exceed 80 percent.[48,49]

Flow Control

Traffic will be lost if no buffer is available to accept incoming blocks. When the data transmission system includes intermediate storage, a receiver may defer accepting traffic for a certain period of time. However, it is vulnerable to network countermeasure, such as the discard of traffic. Thus the question is: How many buffers must be set aside for the expected traffic? This is a classical problem in real-time systems design. A number of books cover this subject.[57] The system designer has access to analytic models based on queuing theory, simulation, and measurement of live traffic. Assuming that hypotheses about real traffic are valid, results are customarily histograms relating load factors and probability of traffic overflow. Unless buffers, table entries, CPU cycles, are overallocated, chances are not totally negligible that occasionally traffic will have to be delayed or discarded. This is acceptable if retransmission is possible, as long as the percentage of bandwidth used up by retransmission remains low enough. Here the tradeoff is between buffer and bandwidth.

A second question is: how to control the sender's flow? As seen earlier, the use of stop and go mechanism is simple, because it requires little or no strategy. However, stop-and-go messages have to compete with data blocks on two-way channels, hence they may be delayed. Store and forward transmission systems add even more delay between sender and receiver. An evaluation should be made of the probability of excess arrival, or idle time. This is related to block rate, block size, and transit delay characteristics. Occasional loss of bandwidth may be acceptable, up to a point.

Better control is offered with credit schemes in the sense that it is possible to restrict senders to the exact amount of buffer space available at the receivers end. However, in networks comprising a very large number of low activity terminals, the probability of a terminal transmitting a block is small, hence buffers would be mostly squandered. In such conditions it is customary to overbook buffers, as seats on airplanes. As a result, overflow may occur with certain probability. Flow control is still useful, but only as a second line of defense.

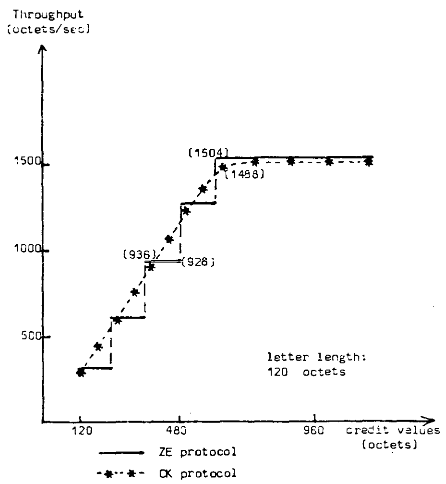
Another situation is high volume traffic between computers, such as in the case of file transfer for batch processing. Traffic loss or idle time would degrade service performance. There the question is: What credit policy would assure a certain throughput rate? Experience is still scarce in this area. Modeling and simulation provide interesting insights, and may help tuning flow control parameters in a specific network. It is shown that throughput increases almost linearly with credit values, up to a maximum determined by the block size (see Fig. 35, excerpted from [50]). In this figure, the curve ZE refers to the protocol by Zimmerman and Elie in [82] and CK refers to the protocol by Cerf and Kahn described in.[14]

From these results it might be tempting to increase the block size by multiplexing blocks of several parallel data streams into larger blocks. Results from simulation show that the effect is a decrease of throughput of each data stream up to a cross point corresponding to large credit values (see Fig. 36 excerpted from [50]).

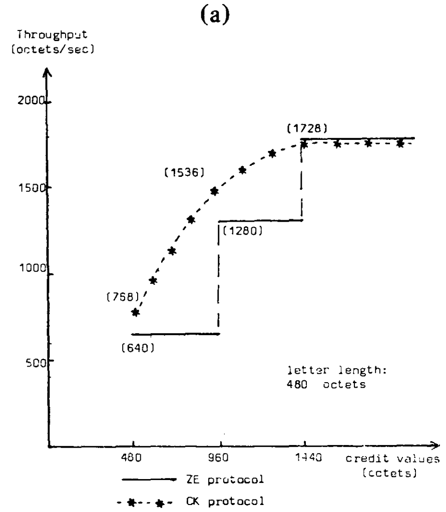
Protocol Overhead

The influence of the major protocol functions has been covered in the previous sections. There remains an odd lot of considerations that do not come easily under

A Tutorial on Protocols

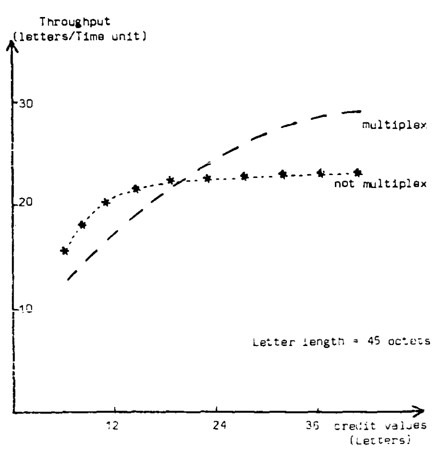


(a)

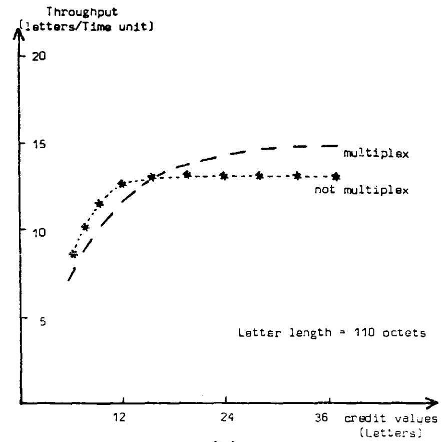


(b)

Figure 35. Flow control and throughput



(a)



(b)

Figure 36. Multiplexing and throughput

meaningful headings. We lump them together under the term overhead.

When defining message formats one is faced with the choice of fixed versus variable fields. Fixed formats mean economic parsing, but also unused bits. There is also the tradeoff between the number of different message headers, and the number of control fields within each header. Choices should be based on assumptions about the percentage and average size of each type of message in the total traffic. Measurements of ARPANET pointed out that the user bits going end to end above the transport protocol (NCP)[9] represented 24 percent of all the bits carried through the packet net[47] not counting internal service traffic. The remaining 76 percent were headers, trailers, ACK's, etc. Indeed, a large proportion of real traffic is made of one-character packets, which results from a characteristic of some of the hosts on the network rather than the network itself.

The latter example illustrates perhaps undesirable aspect of layered systems. When layers are carefully designed as independent system components, each one introduces its own layer of overhead without regard to what happens elsewhere. The overall combined overhead can be quite large.

Protocol Sensitivity

In various places we have mentioned interplays between protocols and environmental factors. A summary of the most common interplays is introduced here for easier reference.

Error Type: A protocol can only cope with errors for

A Tutorial on Protocols

which it has been designed. For example, X.25, a standard interface for public packet networks,[11,12] contains no mechanism for the detection of errors generated by the packet network. It relies entirely on error signalling from the network.

Error Rate: Performance may change only slightly over large variations in error rate. However, there are thresholds past which degradation is drastic. Quite often the threshold may be adjusted in varying some protocol parameters: block length, anticipation window, retransmission policy, acknowledgment policy. If this is not sufficient, it may be necessary to use a different protocol, e.g., forward error correction instead of or in addition to retransmission if the ratio of retransmitted packets exceeds say 10 percent of all transmitted packets.

Transit Delay: Delay is a major factor in tuning properly error and flow control schemes for getting maximum throughput. In particular, large transient variations may be more detrimental than a long but stable delay.

Protocols require that some assumptions be made about maximum transit delay. If these assumptions are not valid, performance may be degraded, or worse, undetected errors may occur if the transit delay exceeds the cycle over which the error control identifiers apply.

Blocks Out of Sequence: A block getting out of sequence by one or a few positions is rarely a burden, when this is anticipated in the protocol. It would be a different matter if some blocks could be out of sequence by, say, 100 or more positions, as this would require considerable buffering to be available at the final destination.

Duplicates: Duplicates arise due to retransmission and create no hardship when they are anticipated and the number of them remains below a few percent. However, the coupling of duplicates with excessive transit delays may be disastrous, since duplicates could appear to be valid messages.

Traffic Patterns: Although one of the least known aspects, it has been clearly recognized that performance is strongly dependent on traffic characteristics: message length, distribution, message rates, competing traffic.

Resource Management: This is the dual of traffic patterns, since any message in transit ties up some resources. It is certainly the most critical area in protocol peformance.[68]

STANDARDIZATION

The Rationale for a Standard

Standardization means agreement on reducing the number of ways of doing the same thing. Diversity may be justified when it stems from substantially different constraints or objectives. Nevertheless, it is fair to recognize that in the computing field diversity is often the result of adherence to local or individual standards. A considerable proportion of human skill and financial resources are devoted to reinventing the wheel and converting from one standard to another. Indeed, not everybody is in favor of standardization, unless the standards are theirs. Hence, we observe a standards war: PL/1 against Cobol, EBCDIC against ASCII, X.25 against HDLC, etc. Thus the term standard should not be taken at face value. In practical terms, standardization does not eliminate diversity and never will. It is only a patient trimming process weeding out excess diversity.

With the policy of limited diversity comes a number of advantages.

- Techniques may be better understood and improvements can be concentrated on smaller set.
- Engineering and maintenance costs may be reduced.
- Individual suppliers and products are likely to support a wide range of these techniques and thus the choice of available systems may be greatly increased for many applications.

There are also counterarguments.

- Individual techniques are often tailored to an application and too much generality may be more expensive.
- Suppliers may not favor increased compatibility when they control a certain market.
- Past investment may rule out major changes.
- Innovation could be stifled.

Thus standardization is a subtle game; what techniques should be standardized and when?

Major Existing Standards

A complete review of existing standards may be found in the literature.[18] Here we shall only focus on a few major standards that are most relevant from a data processing and data communication standpoint.

A Tutorial on Protocols

CCITT-V24 (alias RS-232) Modem Interface for telephone circuits up to 20 kbits/s. This is the most widely used interface for data transmission. It is data and procedure transparent, but without capability for switched circuit signalling.

CCITT-X.21 Interface for switched digital circuits. It covers the whole range of speeds, from 600 bits/s up to megabits per second range, although there are few switched digital networks which operate at higher speeds. It is completely transparent to data and procedures. The dialing phase (connection setup) is based on electrical signalling, rather than control messages. This is a definite shortcoming due to an early design, at a time when telephone administrations had not yet anticipated packet switching.

CCITT-X.25 Interface for public packet networks.[11,12] It works on the principle of virtual circuits carrying packets in sequence, and provides for cascading flow control between source and destination. It will presumably be gradually adopted by equipment manufacturers as an alternative to switched circuit or multipoint line interfaces.

There is, however, some dissatisfaction about the complexity of this interface, which was drafted up in a remarkable hurry. Various design flaws and ways to interpret the standard are still matters to be resolved[4]. A major deficiency, especially for the level of complexity of X.25, is the lack of mechanism assuring end-to-end integrity of data flows. This will likely prevent X.25 from becoming the workhorse of data transmission, except where an end-to-end protocol is also provided.

CCITT-X.3, X.28, X.29 (alias PAD).[12] These standards are a package intended for the remote handling of asynchronous ASCII terminals across public packet networks. X.28 is the interface with the terminal and its human user, X.29 is the interface between the remote handler and an application program (or an access method) located in a computer, X.29 requires the use of X.25.

The Packet Assembler Disassembler (PAD) does not fit within the reference architecture introduced earlier. Rather, it appears to be derived from a centralized network structure (Fig. 37), in which a physical circuit has been replaced with a virtual one. In this context, the PAD is equivalent to a remote terminal handler in a star network (Fig. 38).

ISO-HDLC: High-level Data Link Control (HDLC)[43 44], is a data link control procedure worked out within the International Standard Organization (ISO).[43,44] It was designed primarily for multipoint circuits, at a time when packet switching was not anticipated. Since then it was revamped for handling quasi-symmetric point-to-point circuits.

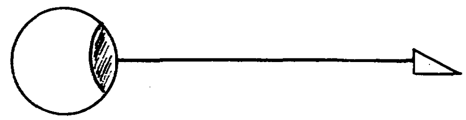


Figure 37. Central handler

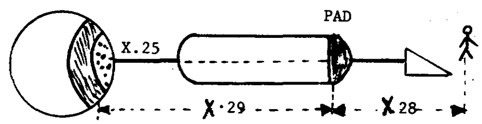


Figure 38. Remote handler

Although HDLC contained most of the capabilities desirable for virtual circuit interface, it was not chosen by CCITT. It only appears at the link level (physical line) of X.25, while the packet level (virtual circuit) is controlled by a similar but different procedure.

Presently, studies are underway for the definition of a public data network interface (Frame Mode DTE) using HDLC.

Future Standard Development

On the computer network scene, early initiatives have been taken by groups of experimenters for reaching some consensus on internetwork protocols. Gradually, these pre-standards seep into the industrial world, and may become a basis for future standards. Most of this groundwork is carried out within research communities developing around experimental computer networks (ARPANET[71], CYCLADES[65], EIN[2]). Their international forum is the Working Group WG 6.1 of the International Federation of Information Processing (IFIP). This group has produced proposals for several basic layers of the reference architecture: datagram format[40], transport protocol[42], virtual terminal protocol.[41] These protocols (or equivalent variants) have been validated through experiments, simulation, and analysis. Higher level protocols, such as file transfer, remote job entry, are still being investigated, and not enough experience is available for reaching a large consensus.

A new subcommittee entitled "Open Systems Interconnection" has been set up by ISO in 1977 (ISO/

A Tutorial on Protocols

TC97/SC16) for the purpose of standardizing architecture and protocols required by heterogeneous networks. Hopefully, this may give a serious boost to computer network standards.

The following is a brief summary of recent trends in protocol evolution.

Protocols of the transmission layer are dependent on the technology; circuits, radio, satellite. Circuit characteristics do not change very much, thus line procedures tend to stabilize, however, optical fibers will bring a quantum jump in transmission bandwidth. The utilization of radio and satellite transmission is undergoing evolution.[6,45,46,48,49]

So far, packet-switching standards are limited to virtual circuits (X.25), which may be appropriate for terminal-to-host communication, but datagrams are more effective for broadcast type applications. The controversy[67] between datagrams and virtual circuits should settle down if public networks offer both services. A datagram standard is under study within CCITT and could possibly be agreed upon by 1980.

Transport protocols are now reasonably well understood functionally, but performance tradeoffs still require further investigation.[73,74] New transmission media using broadcast techniques[29,45,46,58], will likely foster more sophisticated types of associations 1-to-N or M-to-N, which present protocols do not handle. This may bring elegant solutions to some problems of distributed data base, for which point-to-point protocols are rather awkward.

Virtual terminal protocols of the first generation handled Teletype-like terminals. Some early commercial developments are now under way with a second generation designed for handling programmable displays. A period of maturation will be necessary for better understanding the concepts and practical requirements brought about by the rapid expansion of programmable terminals.

All other protocols are still in the domain of research and experiment.

The convergence of personal computing and CB ratio should not take much time to spawn a new sociological phenomenon. The evolution of protocols in this mass communication context is a fascinating and open question.

It certainly would appear rational to define standards for network architecture, protocols, and interfaces. Existing networks will remain alive much longer than their individual components. What makes networks obsolete are neither their components, which can be replaced, nor their protocols, which can be improved, but the services they offer. It is only when pressure builds for new services which cannot be accommodated on an existing network, that a decision is made to build a new network offering both existing and new services. The life cycle of a network is about 10 years.

The advent of cheap microsystems is pushing industry into a reassessment off its computing resources. A new life cycle is starting with distributed architectures. This opportunity may be auspicious for new standards to succeed. The market scene is a cornucopia of new products from a host of suppliers. Common protocols are the only viable solution towards heterogeneous systems.[83,85] There is presently a definite concern among users and suppliers for a basic set of standard protocols within data processing systems, not just for transmission interfaces. If this trend is not thwarted by powerful and conflicting interests, standardization should become active in those areas identified previously as the most vital layers of a distributed resource sharing system, viz. transport protocol, virtual terminal, resource management (tasks, files, peripherals), and data base access.

CONCLUSIONS

There is some similarity between the evolution of programming languages and the evolution of protocols, except that they are separated by about twenty years. Initially, programming languages were invented to meet practical needs. Compilation was a matter of brute force. Thereafter, came models, grammars, theories, and theorems. Then structured programming and lastly proofs of correctness.

Like programming languages, protocols realize abstract machines. Messages and their headers are operands and instruction sets of low level languages designed to be used by machines rather than people. The first generation of protocols, basically those which were developed for the ARPANET, was rather difficult. The second generation triggered by CYCLADES attempted a more precise definition in ALGOL like procedures which were simulated before implementation.[84]

Thereafter came models and formal representations.[7,21,36,38,51,62,75] Recently some verification techniques have been proposed.[5,25] Unfortunately, the body of knowledge accumulated with programming languages does not transfer very well to protocols. It seems that the reason is mainly the introduction of asynchrony and noise in the network environment. These protocols must perform in an environment of uncontrolled parallelism and mutual suspicion, which is a far cry from the relative neatness of modern programming languages.

A Tutorial on Protocols

Occasionally, one attempts to name a few great inventions in computer history, the transistor, the core memory, the microprocessors. Objects have always been better candidates than concepts. However, protocols will rank among the major innovations of the 1970's. Not because they introduce new techniques or new applications, but for the sake of their conceptual potential.

For the first time the design of computer systems puts the emphasis not on the internal management of resources, but on communication between resources of different systems. The very idea of a network architecture defined by an international standard is a novel concept in itself.

What we are witnessing is the emergence of an abstract construction as a standard, which in time will become the model from which commercial systems are derived. Again, the similarity with programming languages is striking. But the implications of a common network architecture are far significant. Within such a framework, any set of processes around the world could start exchanging information. The final stage is a distributed world machine. Naturally, there are other than technical implications worth investigating.[52]

We may think this goal is not achievable, because technology is changing so fast that concepts will be obsolete almost as fast as they are formulated. It is precisely this problem that a layered architecture can solve. Layers and protocols that are dependent on specific technologies may have to change, but the architecture should remain and evolve.

REFERENCES

[1]N. Abramson, "The ALOHA system—Another alternative for computer communications," *AFIPS-FJCC*, vol. 37, pp. 281-285, Nov. 1970.
[2]D.L. A. Barber, "A European informatics network: Achievement and prospects," in *Proc. ICCC* (Toronto, Ont., Canada), pp. 44-50, Aug. 1976.
[3]R. Binder, N. Abramson, F. Kuo, A. Okinaka, and D. Wax, ALOHA packet broadcasting: A retrospect," *AFIPS-NCC*, vol. 44, pp. 203-215, 1975.
[4]G.V. Bochmann, "Notes on the X.25 procedures for virtual call establishment and clearing," *ACM Computer Commun, Rev.*, vol. 7, no. 4, pp. 53-59, Oct. 1977.
[5]G.V. Bochmann and J. Geosei. "A unified method for the specification and the verification of protocols," in *Proc.* IFIP Congress, pp. 229-234, Aug. 1977.
[6]F. Borgonova and L. Fratta, "A new technique for satellite broadcast channel communications," presented at the 5th ACM-*IEEE Data Commun, Symp.*, pp. 21-24, Sept. 1977.
[7]J. Bremer and A. Danthine, "Communication protocols in a network context," *ACM Sigcom*, in *Proc. Sigops Interface Workshop on Interprocess Communication* (Santa Monica, CA), p. 6, Mar. 1975.
[8]H.Q. Burton and D.D. Sullivan, "Errors and error control," *Proc. IEEE*, vol. 60, no. 11, pp. 1293-1301, Nov. 1972.
[9]S. Carr, V. Cerf, and S. Crocker, "Host-to-host protocol in the ARPA computer network," *AFIPS SJCC*, pp. 589-597, May 1970.
[10]P.M. Cashin, "DATAPAC network protocols," in *Proc. ICCC* (Toronto, Ont., Canada), pp. 150-155, Aug. 1976.
[11]CCITT-X.25-*Orange Book*, vol. VIII-2, pp. 70-108, 1977.
[12]CCITT-Provisional recommendations X.3, X.25, X.28, and X.29 on packet switched data transmission services, p. 100, Oct. 1977.
[13]V. Cerf, A. McKenzie, R. Scantlebury, and H. Zimmermann, "Proposal for an international end-to-end protocol," *ACM Sigcom Comp. Commun. Rev.*, vol. 6, no. 1, pp. 68-89, Jan. 1974.
[14]V. Cerf and R. Kahn, "A protocol for packet network intercommunication," *IEEE Trans. on Commun.*, vol. 22, no. 5, pp. 637-648, May 1974.
[15]A. Chatterjee, N.G. Georganas, and P.K. Verma, "Analysis of a packet switching network with end-to-end congestion control and random routing," in *IEEE Proc* (Toronto, Ont., Canada), pp. 488-494, Aug. 1976.
[16]G.E. Conant and S. Wecker, "DNA: An Architecture for heterogeneous computer networks," in *Proc. ICCC* (Toronto, Ont., (Canada), pp. 618-685, Aug. 1976.
[17]I.W. Cotton, "Computer network interconnection: Problems and prospects," Nat. Bur. Stand., Compu. Sci. Special Publication, 500-6 Apr. 1977.
[18]I.W. Cotton and H.C. Folts, "International standards for data communications: A Status report," in *Proc. 5th ACM-IEEE Data Commun. Symp.*, pp. 4.26-4.36, Sept. 1977.
[19]S.D. Crocker, J.F. Heafner, R.M. Metcale, and J.B. Postel, "Function oriented protocols for the ARPA computer network," *AFIPS-SJCC*, vol. 40, pp. 271-279, May 1972.
[20]Y. Dalal, "More on selecting sequence numbers," *IFIP-WG* 6.1, protocol note no. 4, Oct. 1974.
[21]A. Danthine and J. Bremer, "An axiomatic description of the transport protocol of CYCLADES," presented at the Conf. Computer Networks and Teleprocessing, Aachen, Germany, 1976.
[22]A. Danthine and E. Eschenauer, "Influence on packet node behavior of the internode protocol," *IEEE Trans. Commun.*, vol. COM-24, pp. 606-614, June 1976.
[23]J. Davidson, N. Hathaway, J. Postel, N. Mimno, R. Thomas, and D. Walden, "The ARPANET TELNET protocol: Its purpose, principles, implementation, and impact on host operating system design," in *5th ACM-IEEE Data Commun, Symp.*, pp. 4.10-4.18, Sept. 1977.
[24]D.W. Davies, "Flow control and congestion control," presented at COMNET '77, John Von Neumann Soc., Budapest, Hungary, pp. 17-36, Oct. 1977.
[25]J.D. Day, "A bibliography on the formal specification and verification of computer network protocols," in *Proc. Computer, Network Protocols Symp.* (Liege, Belgium), p. 3. Feb. 1978.
[26]Digital Equipment Corp., NSP specifications, version 2, p. 87, Nov. 1977.
[27]A. Duenki and P. Schicker, "Virtual terminal definition and protocol," *ACM Sigcom Comput. Commun. Rev.*, vol. 6, no. 4, pp. 1-18, Oct. 1976.
[28]EIN-An end-to-end protocol for EIN, EIN Rep. 76 002, p. 22, Jan. 1976.
[29]D.J. Farber and K.C. Larson, "The system architecture of the distributed computer system," in *Proc. Communications Syst. Symp. Computer-Communications Networks and Teletraffic*, NY, pp. 21-27, Apr. 1972.
[30]G. Fayolle, E. Gelenbe, and G. Pujolle, "An analytic evaluation of the performance of the "Send and Wait" protocol," *IEEE Trans. Commun.*, vol. COM-26, 3, pp. 313-319, Mar. 1978.
[31]E. Feinler and J. Postel, eds., *ARPANET Protocol Handbook*, Network Info. Center, SRI International, for the U.S. Defense Communication Agency, Jan. 1978.
[32]J.G. Fletcher and R.W. Watson, "Mechanisms for a reliable timer-based protocol, in *Proc. Symp. Computer Network Protocol* (Liege, Belgium), p. 17, Feb. 1978.
[33]E. Glenbe, J. Labetoulle and G. Pujolle, "Performance evaluation

A Tutorial on Protocols

of the HDLC," in *Proc. Computer Network Protocols Symp.* (Univ. Liese, Liese, France), p. 7, Feb. 1978.

[34]M. Gien, "A file transfer protocol (FTP)" in *Proc. Computer Network Protocols Symp.* (Univ. Liege, Liege, Belgium), p. 7, Feb. 1978.

[35]J. Glories, The SITA story, SITA Monogr. 4, p. 26, May 1973.

[36]M.G. Gouda and E.G. Manning, "Protocol Machines: A Concise formal model and its automatic implementation," presented at ICCC, Toronto, Ont., Canada, pp. 346-350, Aug. 1976.

[37]J.L. Grange and P. Mussard, "Performance measurement of line control protocols in the CIGALE network," in *Proc. Computer Network Protocols Symp.* (Univ. Liege, Liege, Belgium), p. 13, Feb. 1978.

[38]J. Harangozo, "An approach to describing a data link level protocol with a formal language," in *Proc. 5th ACM-IEEE Data Communication Symp.*, pp. 4.37-4.49, Sept. 1977.

[39]W.S. Hobgood, "The role of the network control program in systems network architecture," *IBM Syst. J.*, vol. 15, no. 1, pp. 39-52, 1976.

[40]IFIP-Basic Message Format for Inter-Network Communication, ISO/TC97/SC6 N1281, p. 7, Apr. 1976.

[41]IFIP-WG 6.1 Proposal for Standard Virtual Terminal Protocol, Doc. INWG Protocol # gi or ISO/TC97/SC16 N23, p. 56, Feb. 1978.

[42]IFIP-WG 6.1-Proposal for an Internetwork End-to-End Transport Protocol, ISO/TC97/SC16 N24, p. 46, Mar. 1978.

[43]*ISO-High Level Data Link Control*—Frame Structure, IS 3309, 1976.

[44]*ISO-High Level Data Link Control*—Elements of Procedure, IS 4335, 1977.

[45]I.M. Jacobs, R. Binder, and E.V. Hoversten, "General purpose packet satellite networks," Nov. 1978 *Proceedings of the IEEE* Volume 66, No. 11, pp. 1448-1467.

[46]R.E. Kahn, S.A. Gronemeyer, J. Burchfiel, R.C. Kunzelman, "Advances in packet radio technology," IBID. pp. 468-1496.

[47]L. Kleinrock, W.E. Naylor, and H. Opderbeck, "A study of line overhead in the ARPANET," *Commun. Ass. Comput Mach.*, vol. 19, no. 1, pp. 3-13, Jan. 1976.

[48]L. Kleinrock and F. Tobagi, "Random access techniques for data transmission over packet switched radio channels," *AFIPS-NCC*, vol. 44, pp. 187-201, May 1975.

[49]S. Lam and L. Kleinrock, "Packet switching in a multi-access broadcast channel: Dynamic control procedures," *IEEE Trans. Commun.*, vol. COM-23, no. 9, Sept. 1975.

[50]G. LeLann and H. LeGoff, "Advances in performance evaluation of communication protocols," in *Proc. ICCC* (Toronto, Ont., Canada), pp. 361-366, Aug. 1976.

[51]G. LeLann, "Distributed system—Toward a formal approach," in *Proc. IFIP Congr.*, pp. 155-160, Aug. 1977.

[52]J.C.R. Licklider and A. Vezza, "Applications of information networks," this issue, pp. 1330-1346.

[53]J.H. McFadyen, "Systems network architecture: An overview," *IBM Syst. J.*, vol. 15, no. 1, pp. 4-23, 1976.

[54]J.M. McQuillan, "Enhanced message addressing capabilities for computer networks," this issue, pp. 1517-1527.

[55]J.M. McQuillan and D.C. Walden, "The ARPA network design decisions," *Comput. Networks*, vol. 1, no. 5, pp. 243-291, Aug. 1977.

[56]E.G. Manning, E. Gelenbe, and R. Mahl, "A Homogeneous Network for Data Sharing—Modeling and Evaluation. New York: Amer. Elsevier, Aug. 1974, pp. 345-353.

[57]J.A. Martin, *Systems Analysis for Data Transmission.* Englewood Cliffs, NJ: Prentice-Hall, 1970.

[58]R. Metcalfe and D. Boggs, "ETHERNET: Distributed packet switching for local computer networks," *Commun. Ass. Comput. Mach.*, vol. 19, no. 7, pp. 395-404, July 1976.

[59]N. Naffah, "High level protocol for alphanumeric data-entry terminals," to be published in *Comput. Networks.*

[60]N. Neigus, "File transfer protocol," in *Proc. ARPA-NIC*-17759, p. 50, July 1973.

[61]J. Postel, "Official initial connection protocol," in *Proc. ARPA-NIC*-7101, p. 5, June 1971.

[62]____, "A graph model analysis of communication protocols," UCLA-ENG-7410, 1974.

[63]L. Pouzin, "Network architectures and components," in *Proc. 1st European Workshop on Comput. Networks* (Arles, France), IRIA ed., pp. 227-265, Apr. 1973; also, IFIP-WG 6.1, INWG no. 49.

[64]____, "Network protocols," NATO International Advanced Study Institute on Computer Communication Networks, University of Sussex, Brighton, Nordhoff ed., pp. 231-255, Sept. 1973.

[65]____, "Presentation and major design aspects of the CYCLADES computer network," in *Proc. 3rd ACM-IEEE Commun. Symp.*, Tampa, FL., pp. 80-87, Nov. 1973.

[66]____, "CIGALE, the packet switching machine of the CYCLADES computer network," in *Proc. IFIP Congress* (Stockholm, Sweden), pp. 155-159, Aug. 1974.

[67]____, "Virtual circuits vs. Datagrams-technical and political issues," Presented at the *AFIPS National Comput. Conf.*, NY, pp. 483-494, June 1976.

[68]____, "Flow control in data networks: Methods and tools," in *Proc. ICCC* (Toronto, Ont., Canada), pp. 467-474, Aug. 1976.

[69]____, "Names and objects in heterogeneous computer networks," in *Proc. 1st Conf. European Cooperation in Informatics* (Amsterdam, The Netherlands), pp. 1-11, Aug. 1976.

[70]E. Raubold and J. Haenle, "A method of deadlock-free resource allocation and flow control in packet networks," in *Proc. ICCC* (Toronto, Ont., Canada), pp. 483-487, Aug. 1976.

[71]L.G. Roberts and B.D. Wessler, "Computer network development to achieve resource sharing," *AFIPS-SJCC*, vol. 36, pp. 543-549, May 1970.

[72]R.F. Sproull and D. Cohen, "High level protocols," this issue, pp. 1371-1386.

[73]C. Sunshine, "Interprocess communication protocols for computer networks," Stanford Electron. Lab., TR 105, 258 pp., Dec. 1975.

[74]____, "Factors in interprocess communications protocol efficiency for computer network," *AFIPS-NCC*, pp. 571-576, June 1976.

[75]*Proc. Symposium on Network Protocols* (Liege, Belgium), Feb. 1978; This is a collection of 36 pages containing a considerable amount of significant information on recent work.

[76]R. Tomlinson, "Selecting sequence numbers," IFIP WG 6.1, protocol note, no. 2, Sept. 1974.

[77]K.C. Traynham and R.F. Steen, "SDLC and BSC on satellite links: A performance comparison," *ACM Sigcom Comput. Commun, Rev.*, vol. 7, no. 4, pp. 3-9, Oct. 1977.

[78]L. Tymes, "TYMNET, a terminal oriented communications network," *AFIPS-SJCC*, pp. 211-216, June 1971.

[79]A.J. Viterbi, "Error control for data communication," ACM Sigcom Comput. Commun. Rev., vol. 6, no. 1, pp. 27-37, Jan. 1976.

[80]D. Walden, "Host-to-host protocols," INFOTECH State of the Art Rep., Network Systems and Software, Maidenhead, England, pp. 287-316, 1975.

[81]H. Zimmermann, "Terminal and access in the CYCLADES computer network," in *Proc. International Computing Symp.* (Juan les-Pins), North-Holland/American, Elsevier edit., pp. 97-99, June 1975.

[82]____, "The CYCLADES end-to-end protocol," in *Proc. 4th ACM IEEE Data Commun. Symp.* (Quebec, Canada), pp. 7.21-7.26, Oct. 1975.

[83]____, "High level protocols standardization: Technical and political issues," in *Proc. ICCC* (Toronto, Ont., Canada), pp. 373-376, Aug. 1976.

[84]____, "The CYCLADES experience: Results and impact," in *Proc. IFIP Congr.* (Toronto, Ont., Canada), pp. 465-469, Aug. 1977.

[85]H. Zimmermann and N. Naffah, "On open systems architecture" presented at the ICCC, Kyoto, Japan, pp. 669-674, Sept. 1978.□

A Concise Summary of Public, Leased, and Private Transmission Facilities

Problem:

This report is offered primarily to clarify the sometimes conflicting terminology used to define the various classes of currently available transmission facilities and to separate the facility definitions from transmission media definitions. Facilities are generally defined by their accessibility and by the way they are switched. The most accessible service is the public dial-up facility (telephone system). Next are the leased facilities, which can consist of lines leased from common carrier or noncommon carrier sources. Last are the strictly private facilities in which a company builds, maintains, and completely owns dedicated communications lines and equipment. The switching classifications, at present, are line, message, and packet switching. Line switching implements a continuous point-to-point connection, from dial-up to hang-up (as in the telephone network), and is not concerned with the message content of the connection. Message switching is, as its name implies, concerned with messages. In a message-switched network, a point-to-point connection is made only for the duration of a discretely identified message, but the message length can be indeterminate. Packet switching is similar to message switching except that the "message" length is fixed, thus the term "packet"; and the original message may be physically divided among many packets. Packet switching, like line switching, is not concerned with the length of the original message. Also, packets are switched from point-to-point completely under computer control, so a physical connection is never actually made between sender and receiver. The computer-controlled connection is called a virtual connection.

The facility media are not necessarily related to the type of service and switching. The most common medium consists of millions of miles of plain copper wire that interconnects telephones in the AT&T network. The media are graded by their degree of widebandedness, which translates into transmission speed. Copper wire is the narrowest bandwidth media. In order of increasing widebandedness, the other significant media are radio, conditioned wire lines, coaxial cable, microwave, satellite (really a form of microwave), and optical fiber.

A Concise Summary of Public, Leased, and Private Transmission Facilities

Solution:

In designing a teleprocessing network the systems analyst has a number of types of communication lines that he may use in the construction of his network. This report summarizes these types.

TRANSMISSION RATES

Perhaps the most important parameter in comparing communication lines is their transmission speed. The communication lines in use vary in transmission rate from about 50 bits per second up to more than a million. Speeds higher than those of a voice line, however, are not widely available on a dial-up basis, although switched wideband facilities now exist in some cities.

Figure 1 gives an estimate of the relative popularity of different data transmission time requirements and message lengths. The digits from 1 to 9 indicate the relative usage of different areas on the chart, the most popular being real-time dialogue systems with a response time of 1 to 3 seconds. Many areas on the chart not marked by a digit also have data transmission in use, but it is less common than the marked areas. As will be seen, the 50,000-bits-per-second line does not intersect any of the marked areas. Consequently, it was not surprising, perhaps, that when AT&T introduced their 50,000-bits-per-second switched service to the first few cities in 1970, there were few subscribers for it. Given some years of development, however, this situation will almost certainly change. If higher-speed lines are available at reasonable cost, the digits on this figure will migrate in a southeasterly direction.

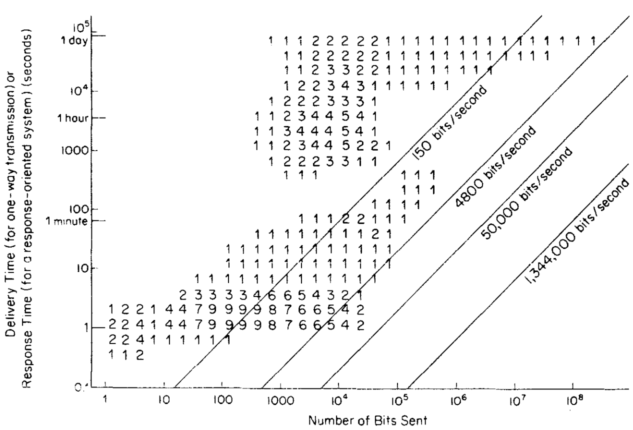


Figure 1. An estimate of the relative popularity of different data transmission time requirements and message lengths. The figures from 1 to 9 indicate the relative usage of different areas on the chart, the most popular being real-time dialogue systems with a response time of 1 to 3 seconds. Many areas on the chart that are not marked by a digit also have data transmission in use, but it is less common than in the marked areas. The lack of digits around the line speed of 50,000 bits per second may have been caused by lack of line facilities rather than by a genuine absence of requirement

Figure 1 was given to fifty or so systems analysts, all highly experienced in diverse areas, and asked how they thought the digits might change in the next five years or so. Their estimates differed over a surprisingly wide range, and it became clear that they were imagining entirely different uses of data transmission. Figure 2 shows a composite diagram made from some of the more plausible replies. The moral seems to be that, given good line facilities, the data transmission industry will grow in many new directions, including some that are barely anticipated today.

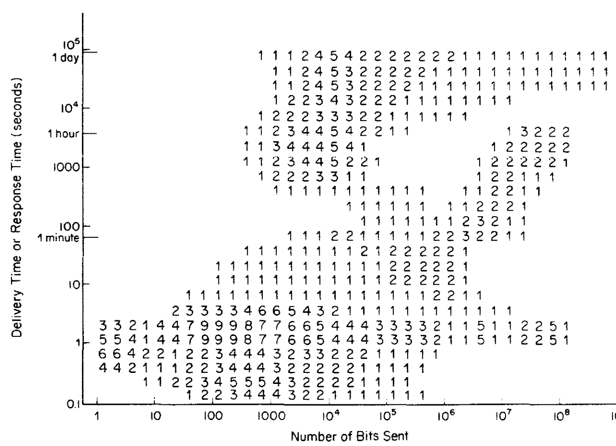


Figure 2. A guess at the relative popularity of different data transmission uses five years hence. A group of experienced systems analysts were asked to make estimates of how Figure 1 would change. Their estimates differed over a surprisingly wide range, suggesting that, given appropriate transmission facilities, a wide variety of new uses for data transmission will emerge (see text)

CATEGORIES OF LINE

Table 1 lists the main types of leased and public communication links in order of increasing speed. The speeds have been listed in terms of the number of data bits per second that may be sent over the line. Communication lines fall into one of three categories of speed:

1. Subvoice grade. Lines designed for telegraph and similar machines transmitting at speeds ranging, in the United States, from 45 to 150 bits per second. Some

Systems Analysis for Data Transmission by James Martin, Chapter 17, pp. 229-254 © 1972 Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.

A Concise Summary of Public, Leased, and Private Transmission Facilities

1. Public (Dial-Up) Lines

	Bit Rate (bits per second)		Bandwidth (kHz)	Type of Line			Half Duplex or Full Duplex
				United States		United Kingdom	
	Fixed	Dependent on Modem		AT & T	Western Union		
Subvoice grade	45 (6 Murray characters per second)	Up to 45			TTY-TWX and CPT-TWX CE-TWX		HDX HDX or FDX
	50				Telex	Telex	HDX
	110 (10 DIC characters per second)	Up to 150 Up to 200			TTY-TWX and CPT-TWX CE-TWX	Datel 200	HDX HDX or FDX FDX
The public telephone network		Up to 600 Up to 1200				Datel 600 Datel 600	FDX FDX
		600 to 4800 (Certain modems achieve higher speeds.)	3 (Not all freely usable because of network signaling.)	Public network		Public telephone network	HDX or FDX
Switched wideband networks	600 1200 & 2400 4800* 9600* 38,400*	(Other speeds will be achievable with other modems.)	2 4 8* 16* 48*		BEX (Certain cities only)		FDX
	50,000	Up to 50,000		Dataphone 50 (few cities only)			FDX

*Planned but not yet available.

2. Leased Lines

	Bit Rate (bits per second)		Bandwidth (kHz)	Type of Line			Half Duplex or Full Duplex
				United States		United Kingdom	
	Fixed	Dependent on Modem		AT & T	Western Union		
Subvoice grade		Up to 45		1004			HDX/FDX
	50					Tariff H	FDX
		Up to 55 Up to 75		1002 1005			HDX/FDX HDX/FDX
	100					Datel 100	FDX
		Up to 150 Up to 180		1006	1006		HDX/FDX HDX/FDX
	200					Datel 200	FDX
Voice-grade lines		Up to 600 Up to 1200	3 3			Datel 600 Datel 600	FDX FDX
		600 to 10,500 (For the higher rates, conditioning is needed.)	3	3002 (C1, C2, C4 and C5 conditioning)		(Datel 2000 refers to conditioned or high-quality voice lines.)	HDX/FDX
Wideband	19,200 40,800 50,000 230,400		24 48 48 240 1000 (approx.)	8803 8801 8801 5700 (originally TELPAK C) 5800 (originally TELPAK D) (5700 & 5800 tariffs all provide "bundles" of smaller bandwidth lines.)		Special quotation	FDX FDX FDX

Table 1. Types of communication lines available

A Concise Summary of Public, Leased, and Private Transmission Facilities

countries have lines of higher speed than their telegraph facilities, but still much slower than the capacity of voice lines. Britain, for example, has its Datel 200 service operating at 200 bits per second. Most industrialized countries outside North America have a similar 200-bit-per-second service (a CCITT standard). All these lines are today commonly obtained by subdividing telephone channels.

2. Voice grade. At present, telephone channels normally transmit at speeds from 600 to 4800 bits per second. Speeds of 9600 and possibly higher may become common in the near future. Dial-up telephone lines are commonly used for speeds of 1200 or 2400 bits per second today. Speeds up to 3600, however, may soon become common on the public network.

A speed of 4800 is already possible with a reasonable error rate but requires elaborate modem design. Telephone organizations in some other countries have not yet permitted the use of such high speeds over their telephone lines. In many countries, 600 or 1200 bits per second is still the maximum.

3. Wideband. Wideband lines give speeds much higher than voice channels, using facilities that carry many simultaneous telephone calls. Speeds up to about 500,000 bits per second are in use, and higher bit rates are possible if required.

All these line types may be channeled over a variety of different physical facilities. This report and indeed the tariffs themselves normally say nothing about the medium used for transmission. It might very well be wire, coaxial cable, microwave radio, or even satellite. The transmission over different media is organized in such a way that the channels obtained have largely the same properties—same capacity, same noise level, and same error rate. The user generally cannot tell whether he is using a microwave link, coaxial cable, or pairs of open wires stretched between telephone poles. Only satellite transmission requires different data-handling equipment, and here only because a delay of one-third of a second or so occurs in travelling the great distance.

SWITCHED VS. LEASED LINES

The next most important parameter about the lines is whether they are public switched lines or not.

Voice lines and telegraph lines can either be switched through public exchanges (central offices) or permanently connected. Facilities for switching broadband channels are in operation in some countries, although most broadband channels today are permanent connections.

When you dial a friend and talk to him on the telephone, you speak over a line connected by means of the public exchanges. This line, referred to as a "public" or "switched" line, could be used for the transmission of data. Alternatively, a "private" or "leased" line could be connected permanently or semipermanently between the transmitting machines. The private line may be connected via the local switching office, but it would not be connected to the switchgear and signaling devices of that office. An interoffice private connection would use the same physical links as the switched circuits. It would not, however, have to carry the signaling that is needed on a switched line. This is one reason why it is possible to achieve a higher rate of data transmission over a private line. Another reason is that private lines can be carefully balanced to provide the high quality that makes higher-speed data transmission possible.

Just as you can either dial a telephone connection or have it permanently wired, so it is with other types of lines. Telegraph lines, for example, which have a much lower speed of transmission than is possible over voice lines, may be permanently connected or may be dialed like a telephone line via a switched public network. Telex is such a network; it exists throughout most of the world, permitting transmission at 50 bits per second. Telex users can set up international connections to other countries. Some countries have a switched public network, operating at a somewhat higher speed than Telex but at less speed than telephone lines. In the United States, the TWX network gives speeds up to 150 bits per second. TWX lines can be connected to Telex lines for overseas calls. Also, certain countries are building up a switched network for very high-speed (wideband) connections. In the United States, Western Union has installed the first sections of a system in which a user can indicate in his dialing what capacity link he needs.

ADVANTAGES OF LEASED LINES

The leased voice line has the following advantages for data transmission over the switched connections:

1. If it is to be used for more than a given number of hours per day, the leased line is less expensive than the switched line. If it is used for only an hour or so per day, then it is more expensive. The breakeven point depends on the actual charges, which in turn depend on the mileage of the circuit, but it is likely to be of the order of several hours per day. This factor is clearly an important consideration in designing a data transmission network.

2. Private lines can be specially treated or "conditioned" to compensate for the distortion encountered on them. The common carriers charge extra for con-

ditioning. In this way the number of data errors can be reduced or, alternatively, a higher transmission rate can be made possible. The switched connection cannot be conditioned beforehand in the same way because it is not known what path the circuit will take. Dialing at one time is likely to set up a quite different physical path from that obtained by dialing at another time, and there is a large number of possible paths. Modems now exist in that condition dynamically and adjust to whatever connection they are used on. These devices enable higher speeds to be obtained over switched circuits, but they are expensive.

3. Switched voice lines usually carry signaling within the bandwidth that would be used for data. Data transmission machines must be designed so that the form in which the data is sent cannot interfere with the common carrier's signaling. With some machines, this operation also makes the capacity available for data transmission somewhat less than that over a private voice line. A common rate over a switched voice line in the 1960s was 1200 bits per second, whereas 2400 bits per second was common over a specially conditioned, leased voice line. As stated earlier, it is probable that up to 3600 bits per second over switched voice lines and 4800 to 9600 bits per second over conditioned, leased voice lines will become common in the 1970s. Already some modems transmit at higher speeds than 3600 over public voice lines.

4. The leased line may be less perturbed by noise and distortion than the switched line. The switching gear can cause impulse noise that results in errors in data. This is a third factor that contributes to a lower error rate for a given transmission speed on private lines.

The cost advantage of switched lines will dominate if the terminal has only a low usage. In addition, the ability to dial a distant machine gives great flexibility. Different machines can be dialed with the same terminal, perhaps offering quite different facilities. A typewriter terminal used at one time by a secretary for computer-assisted text editing may at another time be connected to a scientific time-sharing system and at still another time may dial a computer-assisted teaching program. Machine availability is another consideration. If one system is overloaded or under repair, the terminal user might dial an alternative system. Often this dialing is done over the firm's own leased tie lines.

LINE CONDITIONING

As has been mentioned, private leased voice lines can be conditioned so that they have better properties for data transmission. Tariffs specify maximum levels for certain types of distortion. An additional charge is made by most carriers for lines that are conditioned.

The American Telephone and Telegraph Company, for example, has three types of conditioned voice lines, the conditioning being referred to as Types C1, C2, C4 and C5. A line ideal for data transmission would have an equal drop in signal voltage for all frequencies transmitted. Also, all frequencies would have the same propagation time. This is not so in practice. Different frequencies suffer different attenuation and different signal delay. Conditioning attempts to equalize the attenuation and delay at different frequencies. Standards are laid down in the tariffs for the measure of equalization that must be achieved. The signal attenuation and delay at different frequencies must lie within certain limits for each type of conditioning. The higher the conditioning number, the narrower are the limits. The result of the conditioning is that a higher data speed can be obtained over that line, given suitable line-termination equipment (modem).

Types C1 and C2 conditioning are applicable to point-to-point and multipoint lines. Type C4 is available only on two-point, three-point and four-point lines. Type C5 conditioning can only be applied to point-to-point lines.

TARIFFS FOR WIDEBAND LINES AND BUNDLES

The North American common carriers offer several tariffs for leased wideband lines. Some of these can be subdivided by the carrier into "bundles" of lower bandwidth. Some can be subdivided into channels for voice transmission, telephotograph, teletypewriter, control, signaling, facsimile or data. With some tariffs the user pays a lower price for the bundles than for the individual channels.

The word "TELPAK" was formally used for the "bulk" communication services offered by the telephone companies and Western Union. The word has now been eliminated from the tariffs but is still found in much literature. What used to be TELPAK C is now called a Type 5700 line, and what used to be TELPAK D is now Type 5800. Both can provide a wideband channel or a bundle of lesser channels.

The TELPAK customer pays a monthly charge based on the capacity of the communications channel he selects, the number of airline miles between locations, and the type and quantity of channel terminals. He has use of this channel on a full-time basis.

A Concise Summary of Public, Leased, and Private Transmission Facilities

Originally there were four sizes of TELPAK channels: TELPAK A, B, C, and D. However, in 1964 the Federal Communications Commission ruled that rates for TELPAK A (12 voice circuits) and TELPAK B (24 voice circuits) were discriminatory in that a large user could obtain a group of channels at lower cost per channel than a small user, who could not take advantage of the bulk rates. In 1967 the TELPAK A and B offerings were eliminated.

The Type 5700 line has a base capacity of 60 voice channels (full duplex).

The Type 5800 line has a base capacity of 240 voice channels (full duplex).

Each voice channel in these lines can itself be subdivided into one of the following:

1. Twelve teletype channels, half or full duplex (75 bits per second).

2. Six class-D channels, half or full duplex (180 bits per second).

3. Four AT&T Type 1006 channels, half or full duplex (150 bits per second).

There cannot be mixtures of these channel types in a voice channel. The Type 5700 line can transmit data at speeds up to 230,400 bits per second; the Type 5800 line has a potential transmission rate much higher. Line-termination equipment is provided with these links, and each link has a separate voice channel for coordination purposes.

The TELPAK channels thus serve two purposes. First, they provide a wideband channel over which data can be sent at a much higher rate than over a voice channel. Second, they provide a means of offering groups of voice or subvoice lines at reduced rates—a kind of discount for bulk buying.

Suppose that a company requires a 50,000-bit-per-second link between two cities, together with 23 voice channels and 14 teletypewriter channels, or perhaps 30 voice channels and no teletypewriter links. Then it would be likely to use the Type 5700 tariff. In leasing these facilities, it would have some unused capacity. If it wishes, it can make use of this capacity at no extra charge for mileage, although there would be a terminal charge.

Government agencies and certain firms in the same business whose rates and charges are regulated by the government (e. g., airlines and railroads) may share bundled services. Airlines, for example, pool their needs for voice and teletypewriter channels. An intercompany organization purchases the bundled services and then apportions the channels to individual airlines. Most of the lines channeling passenger reservations to a distant office where bookings can be made are Type 5700 or 5800 lines, and so also are the lines carrying data between terminals in those offices and a distant reservations computer. There has been some demand to extend these shared facilities to other types of organizations that could benefit from them by sharing, but this is not permissible as yet.

TELPAK originally was proposed as an interstate service, but since then it has become generally available intrastate as well.

Although not a TELPAK offering, Series 8000 is another "bulk" communications service in the United States that offers wideband transmission of highspeed data, or facsimile, at rates up to 50,000 bits a second; the customer has the alternative of using the channel for voice communication up to a maximum of 12 circuits. A Type 8801 link provides a data link at speeds up to 50,000 bits per second with appropriate terminating data sets and a voice channel for coordination. A Type 8803 link provides a data link with a fixed speed of 19,200 bits per second and leaves a remaining capacity that can be used either for a second simultaneous 19,200-bits-per-second channel or for up to five voice channels. These links must connect only two cities. The separate channels cannot terminate at intermediate locations.

Most countries outside North America also offer tariffs similar to the Series 8000, and in most locations quotations for higher speeds can be obtained on request. Obtaining a wideband link in many such countries can be a slow process. This is particularly so if the termination is required in a small town or rural area rather than in a city to which such links already exist. Undoubtedly, as the demand for such facilities increases, the service of the common carriers in providing them will improve.

SWITCHED PRIVATE SYSTEMS

Many firms have private leased-line systems that are switched with private exchanges. It is possible to engineer these systems to the same quality as private lines and thus provide a switching system of better quality than the public network.

A Concise Summary of Public, Leased, and Private Transmission Facilities

Some private lines are wholly owned by their users rather than leased. Users are generally prohibited from installing their own lines across public highways, and most privately installed communication links are wholly within a user's premises—for example, within a factory, office building, or laboratory. Railroads have their own communication links along their tracks. Some companies have private point-to-point microwave transmission links or other radio links. Recently infrared and optical links have proven a valuable medium for the transmission of data; line-of-sight links can be established at low cost, capable of carrying up to several million bits per second. Such links require no license, as do private microwave links. Typical devices can transmit and receive a quarter of a million bits per second. They are used in cities for transmission between rooftops. The main drawback is that the link can be put out of operation for a brief period by rain downpours of abnormal intensity—and for longer periods by thick fog.

TELEX

Telex is a worldwide switched, public teleprinter system. It operates at 66 words per minute (50 bits per second) and uses the Murray code. It is operated in the United States by Western Union. Any teleprinter on the system can dial any other teleprinter in that country, and Telex machines can be connected internationally without speed or code conversion. The United States can dial Canada and Mexico directly, but operator intervention is needed when dialing to other countries. Some countries permit the Telex facilities to be used for other forms of dial-up data transmission. Each Telex call is billed on a time and distance basis.

Each subscriber has an individual line and his own number, as with the conventional telephone service. His teleprinter is fitted with a dial, like a telephone, with which he can dial other subscribers. The teleprinter used may or may not have paper-tape equipment also. The teleprinter can be unattended. When a message is sent to an unattended teleprinter, it will switch itself on, print the message, and then switch itself off.

TELETYPEWRITER EXCHANGE SERVICE

The North American common carriers offer a service that is competitive with Telex. Again, each subscriber has a dial-up teletypewriter with his own number listed in a nationwide directory. This service is called the Teletypewriter Exchange Service (TWX), and it uses the telephone circuits combined with several TWX channels so that they can be sent over one voice channel. The combining or "multiplexing" is done at the local switching office, where the dc signals are changed to equivalent bursts of appropriate frequencies. The link between the local switching office and the subscriber is often a conventional telephone line, and in this case the teletypewriter needs a data set to convert the dc signals to appropriate frequencies in the voice range.

Other manufacturers' data transmission equipment can be connected to TWX lines and can transmit at speeds up to 150 bits per second, half or full duplex. This process requires a special terminal arrangement at additional cost. Three types of access lines to the TWX network exist.

1. TTY-TWX.
This is an access line with a teletypewriter provided by the common carrier. The speeds of transmission are either 6 characters per second in Murrary code or 10 characters per second in Data Interchange Code (DIC).

2. CPT-TWX.
CPT stands for "customer provided terminal"; to this access line the customer can attach any device operating with one of the preceding two speeds and codes and adhering to normal TWX line control. The device could be a computer with an appropriate adapter on its input-output channel.

3. CE-TWX (formerly called "TWX Prime").
CE stands for "customer equipment." This can now be any device and is not restricted to a specific code or character speed. Two TWX subsystems are accessible, one operating at speeds up to 45 bits per second and the other up to 150. A CE-TWX terminal can communicate only with another CE-TWX terminal.

TWX directories listing TTY-TWX and CPT-TWX subscribers are published.

SUMMARY

A summary of the main categories of communication links appears in Table 2.□

A Concise Summary of Public, Leased, and Private Transmission Facilities

Types of Link	Comments
Digital link	Designed for digital transmission. No modem required. Are code sensitive in some cases.
Analog link	Transmits a continuous range of frequencies like a voice line. Modem required.
Switched public	Cheaper if usage is low. Switched telephone lines are universally available. It is necessary to avoid signaling frequencies on public telephone lines.
Leased (sometimes called "private")	Cheaper than public lines if usage is high. May have lower error rate. Higher speeds possible on leased telephone lines than switched ones because (1) conditioning is possible and (2) no signaling to avoid.
Leased with private switching	May give the lowest cost. Combines the advantages of leased lines with the flexibility of switching. Public switched wideband lines may not be available, hence private switched wideband networks are built.
Private (non-common-carrier)	Usually only permitted within a subscriber's premises. See next item.
Private (Non-Common-Carrier) Links:	
In-plant	Very high bit rates achievable using coaxial cables or PCM on wire pairs.
Microwave radio	Permissible in special cases for point-to-point links.
Shortwave or VHF radio	Used for transmission to and from moving vehicles or people.
Optical or infrared	Used for short links—e.g., between city roof-tops—at high bit rates (250,000 bps, typical). No license required. Put out of action by thick fog or *very* intense rain.
Speeds	
Subvoice grade	Usually refers to speeds below 600 bits per second.
Voice grade	Usually refers to analog voice lines using modems of speeds from 600 to 10,500 bits per second.
Wideband	Speeds above those of voice lines, most commonly 19,200; 40,800; 50,000, and 240,000.
(For a detailed list see Table 1)	
Mode of Operation	
Simplex	Transmission in one direction only. Not normally used in data transmission except for telemetry or space applications.
Half duplex	One direction or the other; not both at once. The most common in the United States.
Full duplex	Transmission in both directions simultaneously. On the Bell System costs 10% more than half duplex and can give disproportionately higher throughput if the terminal is designed to take advantage of it.
Note: These terms sometimes describe the limitation of a machine rather than the limitation of the line it is attached to.	

Types of Link	Comments
Conditioning	Applied by the telephone company to *leased* analog voice lines to achieve more uniform attenuation and delay so that higher-speed modems can be employed. Extra charge levied.
Types: C1, C2	For 2-point and multipoint channels.
Types: C4	For 2-point, 3-point and 4-point channels.
Types: C5	For 2-point channels only.
Tariffs	
Tariff includes data set	Common carrier provides the data set (modem) and hence the transmission rate is fixed.
Tariff simply for bandwidth	User chooses the modem. Transmission rate depends on modem design.
DATA-PHONE Service	An AT & T service that provides a data set with the line.
Public Telegraph Networks	
TELEX	Worldwide telegraph network, Murray code, 6.7 characters per second.
TWX	North American telegraph network. Murray code at 6 characters per second, or DIC code at 10 characters per second.
TWX: TTY-TWX	With common carrier teletype machines.
TWX: CPT-TWX	With customer-provided terminals adhering to TTY-TWX codes and line control.
TWX: CE-TWX	With customer equipment. Need not adhere to TTY-TWX codes. Speeds up to 150 bits per second. Not listed in TWX directories.
WATS (Wide Area Telephone Service)	An AT & T tariff provided long-distance dial-up service to and from a specified zone. A fixed WATS line goes to and from the area in question.
WATS: INWATS	Any subscriber in the specified zone may call *to* the WATS line.
WATS: OUTWATS	The call originator is connected to the WATS line and may call any subscriber in the specified zone.
	INWATS or OUTWATS must be on separate lines.
TELPAK	A form of quantity discount in which a bundle of lines can be purchased.
Bundling: TELPAK C, Type 5700	Equivalent to 60 voice channels.
Bundling: TELPAK D, Type 5800	Equivalent to 240 voice channels.
	The TELPAK bandwidth can be subdivided in many different ways.

Table 2. Categories of communication lines and tariffs

The Uses of Message Switching in Communications Networks

Problem:

Many of the problems discussed so far have dealt mainly with communications systems that provide a direct connection between one subscriber and another. This kind of connection is important because it permits information to be exchanged immediately between subscribers and allows them to interact until they are satisfied that the information exchanged is exactly as each intends it to be. In this way, errors introduced by the network and by the subscribers themselves can be corrected as soon as they occur. The public switched telephone network is an excellent example of such a network; it operates by setting up a connection between subscribers in the form of a sequence of point-to-point circuits joined together by switches. The disadvantage of needing a connection between subscribers before they can communicate is that all the communication links and switches necessary to make the connection must be available for the duration of the subscribers' conversation, and both subscribers must be free to interact at the same time.

An entirely different communication technique is for the subscribers to exchange information by sending each other complete messages (like a letter). The advantage of this technique is that the subscribers need not be simultaneously involved with each other, and the messages can be transferred whenever it is most convenient. This is particularly attractive when direct interaction is difficult, as for example when the people communicating with each other are on different continents with different local times.

Message switching is not new—it has been around for over a century—but it is being applied in new and novel ways, especially in its latest guise as packet switching. This report offers some definitive insights into the tradeoffs between line switching and message switching design alternatives and traces the very logical evolution of message switching concepts into the even more digital-traffic-compatible techniques of packet switching.

The Uses of Message Switching in Communications Networks

Solution:

The exchange of messages between people requires an intermediary prepared to accept the responsibility for faithfully conveying the information with which it is entrusted. This intermediary must store the messages until the recipient is ready to accept them and, in the case of a mechanized message handling system, it is usual for a copy of each message to be held so that a failure of the communication circuits can be overcome by retransmission of any corrupted portions of a message. If a number of separate connections have to be set up in sequence between subscribers, it becomes feasible to associate message storage facilities with the junction between connections, so that messages are transferred from store to store. Because a direct connection from subscriber to subscriber is not required, it is possible to wait until circuits become free between one store and the next before messages are forwarded through them. For this reason, the method is known as store-and-forward message switching.

In this report, the techniques used in message switching are discussed with particular reference to their use in conjunction with computer systems.

The store-and-forward method first arose in telegraphy when the speed of operators using mechanical keying became a limiting factor on costly long-distance circuits, which were intrinsically capable of much higher transmission speeds.[1] Keyboards were devised to produce a paper tape, punched to indicate the dots and dashes of the morse code, then used for encoding information. The paper tape could be transmitted at high speed by an automatic sender, and several operators working in parallel could supply enough tapes to keep a circuit operating at maximum speed because the tapes were stored and forwarded as the circuit became free. Although this increased the efficiency of line utilization, the operators at the two ends of the line were decoupled from each other and could no longer converse together as they had done previously because of the delay introduced by storing messages on tape and queuing tapes while waiting for transmission over the line.

An alternative method of improving line utilization is to use some form of multiplexer to time share the line between several operators. This has the advantage that each operator appears to have a line to himself, so the ability of operators to interact with each other is retained. A practical method of multiplexing, which allowed up to six operators to use a single line, was introduced by Baudot in 1874. His scheme, shown in Figure 1, used a mechanical scanner that sampled six keyboards in turn, accepting one symbol from each at every revolution. The keyboards comprised five keys, and a five unit code was devised with all symbols represented by constant length combinations of the five units. This departure from the earlier variable length coding, used by Morse to match the shorter codes to the frequently used symbols, was necessary to facilitate the interleaving of symbols from the various keyboards. The output from Baudot's multiplexer was a serial digital binary signal of six channels with five bits per channel; a marker introduced once per revolution of the scanner completed the frame of this early time-division system.

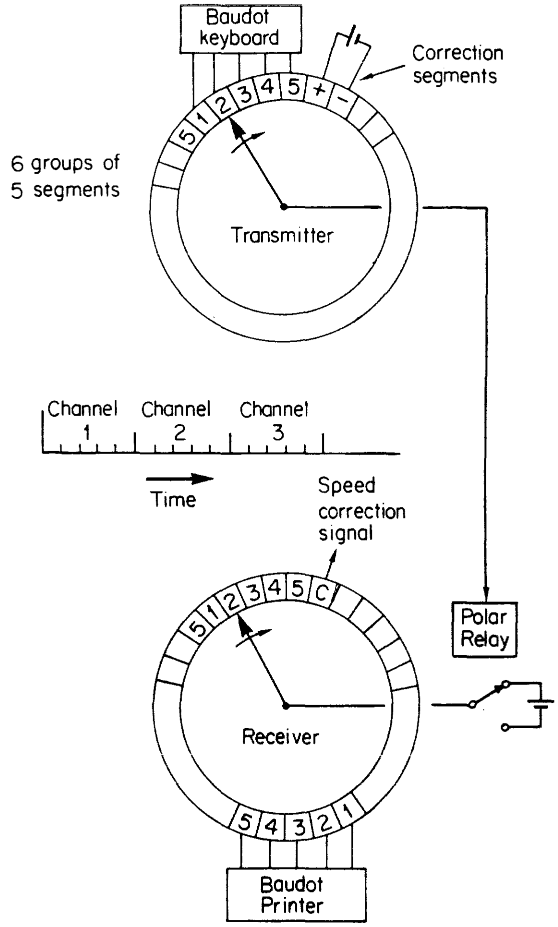


Figure 1. Baudot's multiplexed channel scheme

A similar mechanical arrangement at the receiving end of the line was used to demultiplex or distribute the signals from each channel to particular receiving devices. These were designed to decode the incoming five unit codes into individual symbols and print them onto paper.

It is interesting that, in Baudot's system, the printer made use of mechanical decoding to uniquely determine each symbol, but the encoding was done by the operators. Today, the use of five keys seems cumber-

From *Communication Networks for Computers* by Donald W. Davies and Derek L. A. Barber, Chapter 8, pp 272-309. © John Wiley & Sons. Reprinted by permission of John Wiley & Sons.

some and difficult and, in the later scheme devised by Murray, this feature was replaced by the use of tape punches with typewriter-like keyboards, which prepared tape just prior to transmission. This is shown in Figure 2. A five unit code was also used by Murray, and it became the basis of the International Alphabet No. 2 that is still used today.

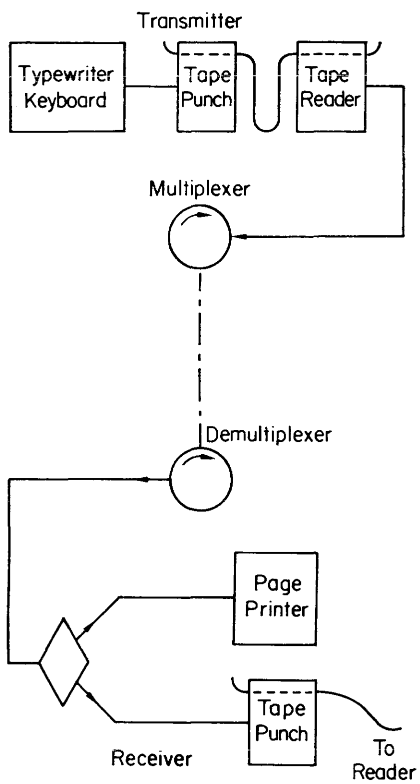


Figure 2. The Murray telegraph

The use of paper tape punches at the receiving terminals of the Murray telegraph system allowed messages to be stored easily in a machine readable form, and this possibility led to the introduction of manual store-and forward switching centers (described later) where messages could be relayed to various destinations after tearing off and sorting incoming tapes ready for transmission to the appropriate destination, rather as letters are sorted in the mail service.

The idea of circuit switching, too, was considered in some of the first telegraph systems; as early as 1850, Dumont devised an exchange with plug boards and wiper switches, which in many respects foreshadowed the later step-by-step switching exchanges used for telephony.[2] But when the telephone was introduced in 1876 there was, naturally enough, an extremely rapid development of circuit switching techniques during which the earlier telegraphy schemes were soon overtaken. No doubt this was due to the simplicity of use and the cheapness of the telephone instrument that could be used, and afforded, by members of the public at large; whereas the skill needed to use the early telegraph keyboard and the cost of the output printers prohibited the use of telegraphy on a wide scale. However, in the early 1930's the invention of the start-stop teleprinter made the public switched telegraph (Telex) service possible, and it has grown steadily in size ever since. The use of frequency-division multiplexing to carry several telegraph channels on a single voice channel has helped by providing economic transmission facilities, and the present growth rate for business use is now quite large.

The earlier developments in telegraph switching were, therefore, mostly concerned with store-and-forward techniques, particularly in the United States where lines were longer and more expensive than in Europe. These techniques developed independently of line switching ideas and progressed through manual, semi-automatic, and fully-automatic systems to the modern message switching systems. But it seems increasingly likely that in Europe there will be a gradual integration of both telegraph message and circuit switching into the data networks of the future. For although (as will be shown later) the traffic characteristics of messages between people differ from those between computers, the volume of traffic between computers may well become increasingly large compared with that between people. It is quite likely, therefore, that all digital information that is not time dependent will eventually be combined and handled by the packet-switching type of network.

EARLY SWITCHING CENTERS

The early store-and-forward telegraph message switching centers were known as torn-tape centers because operators literally tore tape from punches on incoming circuits and placed them in the appropriate transmitters after examining the destination address in the message header. Some torn-tape centers still exist, and the same principles, of course, are found in the later mechanized switching centers.

In torn-tape centers, the torn-off tapes are stored or queued in baskets while waiting to be placed in the appropriate paper tape reader. Queues may also form at the incoming tape punches during peak loading times when several messages may arrive before an operator can get round to inspecting them and sorting them to the right tape reader. A typical torn-tape center is shown in Figure 3.

The first mechanized telegraph message switching centers were, naturally enough, based on torn-tapes centers; they were organized in the same way, and paper tape was still used for storing messages. An arrangement typical of that time is shown in Figure 4. A set of busbars was arranged to make it possible

The Uses of Message Switching in Communications Networks

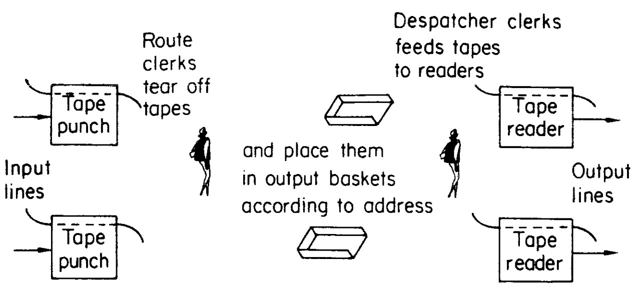


Figure 3. Manual message switching

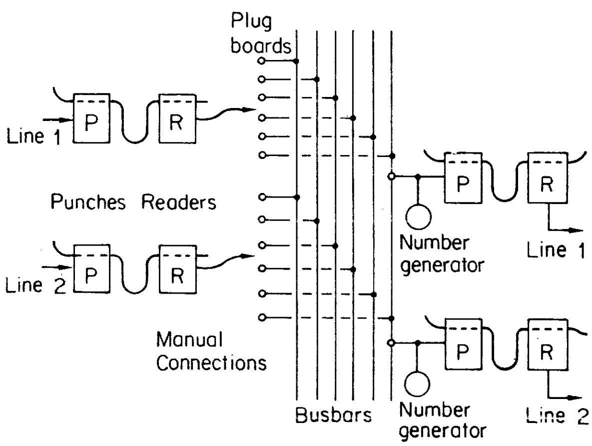


Figure 4. Partially automated scheme

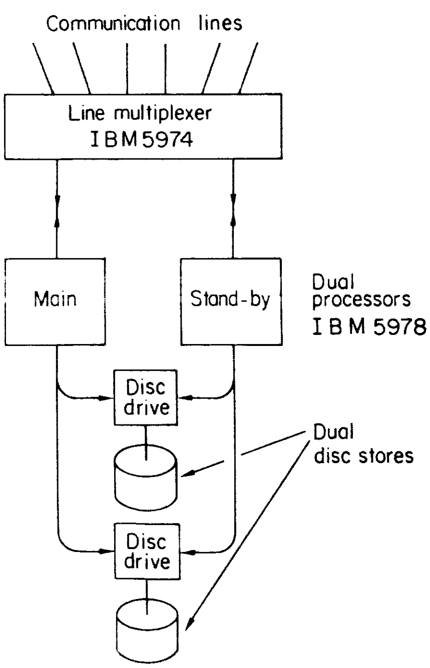


Figure 5. The IBM 5910 message switching system

to connect any of several incoming lines with the various outgoing lines to other switching centers. Each incoming line had a tape punch that reproduced the messages arriving at the center. The tape from the punch was automatically fed to a tape reader that read it out into the appropriate outgoing line as selected by an operator. The operator would make a note of the destination address and serial number of each input message and connect the reader output through a plug and jack arrangement to the busbars. As soon as the busbar became free, it was seized by the next waiting tape reader, and a message number generator was activated to transmit the next outgoing message serial number, which was followed by the output from the tape reader. A group of symbols indicating the end of the message was recognized by suitable decoding circuits and arranged to stop the reader and release the busbar as soon as the message had been sent.

In these semi-automatic message switching centers, the operators kept track of the serial numbers of incoming messages and could request repetitions if any were lost. They also set-up the routing connections inside the center according to the destination address accompanying each message. A certain degree of variation was tolerable in the format of the messages while human operators were involved at the switching centers, but when the remaining functions of checking serial numbers and routing were mechanized, more standardization became necessary; this process of agreeing on standards to permit the introduction of increased mechanization is, of course, still a major problem today.

MODERN MESSAGE SWITCHING SYSTEMS

The advent of the digital computer and the rapid development of new storage techniques made possible radical improvements in message switching systems; it is now common for computers to be used for switching messages and for the messages to be stored on magnetic discs or drums while waiting to be processed or to be transmitted onward to their destination. As these improved systems were made by modelling their structure on the earlier manual ones, computer programs had to be developed to perform the functions originally done by people; in big systems these programs can become extremely complex. However, the speed and flexibility of the computer-based systems allows a wide variety of services to be offered to the users of modern message switching systems.

A typical present day system is the IBM 5910 (Figure 5). It comprises the 5974 modem adaptor and multiplexer; dual 5978 processors, and duplicated disc backing stores. The replication of equipment gives a very high reliability, and this may be readily extended to the subscriber by the use of separate lines from the multiplexer. Each channel in the multiplexer can

The Uses of Message Switching in Communications Networks

carry out low-level processing tasks such as character assembly and code conversion, and can also handle transfers directly to and from the main stores of the two processors. Extensive use is made of microprogramming techniques, which may be changed easily, to give a highly flexible input/output system.

At any one time, one processor is on-line, and the other is in the standby state. The active processor updates the standby, which has a time-out of about one second. If no information is received within this time, the standby processor takes over control.

The capacity of the system is 480 telegraph channels or the equivalent in the range 50 to 9600 bit/s. Sixty lines are handled by a modular block in the multiplexer; each of these lines can have a variety of modems. With an average message length of 300 characters and an average multiple address of 1.3 destinations, the switch can typically handle five messages per second with a transit delay of around 300 milliseconds. This is when a fast store, which can handle a byte in about 0.6 microseconds, is employed. There is a useful 'break-in' facility for interrupting very long messages to allow the processing of urgent short ones.

The connection of the 5910 system to a computer may be done in two ways; it may be made identical to a remote terminal of the computer, or an IBM standard selector channel may be used to link the Central Processor Unit of the 5910 to that of the computer.

The organization of the software is shown in Figure 6. Only the portions shown shaded are purpose-designed; the remainder is common to most applications. The message reception and assembly and any necessary code translation functions are mostly microprogrammed, and after these processes have operated on the message, it is stored ready for analysis and routing. At this stage special application dependent tables are consulted. The processed message is then placed on backing store to await delivery. The queue and de-queue processes are responsible for writing and reading discs, and, when a message is read, it is passed to the disassembler for breakdown into characters which are handed to the transmission processes for despatch to line.

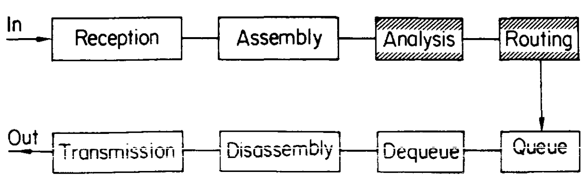


Figure 6. The software organization

The extensive use of modularity in the 5910 hardware makes for high reliability and flexibility. The use of modular software allows a variety of standard programs to be married together, with a minimum of new software, when a new application is required.

The services offered or planned by the Western Union Company in the United States are typical of what has been made possible by the digital computer. Western Union is, of course, traditionally a message carrying organization and, because the long distances between cities in the U.S.A. have always encouraged the use of store-and-forward techniques to maximize line utilization, this same technique has, naturally enough, been the basis of their new computer-based services. These are:

1. INFOCOM, which will be described in detail below.

2. SICOM—the Securities Industries Communication System which gives information on the state of the stock market to its subscribers.

3. Bank Wire Service—the need for this service arises from the fact that there are 14,000 separate banks in the U.S.A. that need to exchange credit information rapidly. The banking situation in the U.K. is quite different.

4. Legal Citation Service—this is a data bank of Case Law accessible to subscribers to the service.

INFOCOM is a general-purpose communications service offering a range of data or telegraph private switched services through a common shared network connecting mainly low-speed keyboard machines. All subscribers are physically joined to this common network, but by establishing closed sub-networks using restricted addressing, INFOCOM provides each subscriber with what seems to him to be a private network. This appears to have limited outlets and is accessible only to authorized members of the group to which he belongs.

The INFOCOM service handles Telex to Telex messages, and Telex to TWX (the A.T. & T. version of Telex that has been sold to Western Union). A third possibility is that Telex messages may be translated into telegrams that are delivered manually. These messages may be sent 'paid' or 'collect.' It is also possible for customer's computers to be linked directly into the INFOCOM system by using an agreed fixed format message from the customer.

The following types of terminal and service may be connected to the INFOCOM network:

1. A private wire to a teletype machine operating at 150 baud using 8-bit ASCII code.

2. A party line to a group of teletype machines (up to six terminals can share a single access to the network).

The Uses of Message Switching in Communications Networks

3. A Telex service at 50 baud using CCITT alphabet No. 2, each terminal having access to the network through a Telex exchange. Subscribers connected to the network can have access to other similar subscribers, to the public telegraph system, to the Telex system, to TWX, an international carrier, or the Canadian network.

4. The final possibility is a 2400-baud link to connect INFOCOM to a subscribing computer. This last service will enable the INFOCOM system to act as a communications processor for computer service bureaux.

The facilities offered by the INFOCOM service are:

1. Simplified message format compared with earlier systems.

2. Code and speed conversion between five-level code, 50 baud, Telex stations, and the eight-level code, 150 baud, Telex machines.

3. Complete privacy between any nominated group of stations.

4. Two levels of message priority.

5. Multiple address, whereby a message may be sent to a group of addresses.

6. Alphanumeric addresses.

7. Message storage and retrieval—messages are held on drum storage for four hours and on tape for ninety days.

8. Itemized billing for individual stations—there is no objection to stations in a private group being rented by different companies, or one company renting stations to another.

9. One station in a group may be nominated as the control station. It can obtain status reports and can direct the system to pass the message for a station that is out of order onto any other station.

10. The delivery time for a 50-character message is less than two minutes.

Computer Centers are located in Chicago, New York, San Francisco, and Atlanta, with concentrators in other cities. All four message switching Centers are joined by 4,800 bit/s data links to form a fully connected mesh network.

The Computer Centers are based on the use of two UNIVAC 416 machines with 65,000 word core stores and 16 input/output channels. One of the machines is a processing computer, the other a communications controller. These two machines are joined by a 66,000 word/second data link, and a third standby machine can be rapidly switched in to replace either of the other machines. For back-up storage there is a small Sperry-Rand type 330 drum of 165 K characters. This contains the operating system programs, which may be reloaded if it becomes necessary to restart the system. It also contains special subroutines used in collecting statistics and similar tasks.

There is a large 44 million word Sperry-Rand drum on which incoming messages are stored and a number of back-up tape decks on which messages are also stored, so that the integrity of the system may be maintained in case of equipment failure.

There are various diagnostic facilities for compiling error statistics for checking on customers' complaints and tracing whether messages came into and left the system, and a pair of printers which print out logging and fault messages to the operating staff.

THE CHARACTERISTICS OF NETWORKS AND USERS

The preceding discussion of telegraph message switching contrasted the techniques of circuit-switching and store-and-forward switching. It is worthwhile now to examine the properties of the two techniques from the point of view of the user to see how they influence the way he may operate; equally important, and more usually considered, are the characteristics of the users behavior and their impact on the design of a communications system.

The way two people communicate and interact depends almost entirely upon the delay inherent in the communication channel between them. A useful way to think about human behaviour is to liken it to a multi-access system handling multiple interrupts. The degree of conscious attention given to any task depends on the priority assigned to it. Normally, face-to-face interaction with another person is a high-level activity, often involving an exchange of concepts with very high information content efficient coding. A high proportion of available attention is given to this kind of interaction.

When a telephone call is made, the channel characteristics do not unduly restrict the ability to communicate. The call set-up procedure and time taken are acceptable, and the direct connection with no apparent delay permits immediate interaction; so both parties in communication give their full attention and cooperate together in the exchange of information. Only when delays approaching one second are introduced is any effect observed, probably caused by the lack of immediate feedback of the sort usually given

The Uses of Message Switching in Communications Networks

by one party when the other is speaking. The introduction of much more delay, as when a telephone answering machine is employed, turns the channel into a kind of message switched system, and it is very unlikely that people would communicate verbally to any extent through such a system.

A Telex call, being a direct connection, is sometimes used by operators to hold conversations, but the delay due to typing makes it rare for ordinary subscribers to do so. Usually, the Telex network is used as a high-speed telegram exchange service, which is very convenient for business because of the rapid interchange of messages possible. However, the immediacy of the communication process is lost, so the subscribers do not interact together; instead, each message is treated by them as an interruption of their other activities and is dealt with as soon as convenient after it arrives. There are definite advantages in this mode of co-operation because neither subscriber is committed at a particular instant and may divide his time between the 'conversation' and other activities. This permits the transaction to proceed at a reasonable pace, and subscribers are not held up unduly while waiting for a response to their last message.

When a subscriber is engaged in transactions with several other subscribers during the same period, telephone and Telex calls may be difficult to make due to the high probability that one or other of the terminals will be busy when an attempt is made to call it. Some private circuit-switched systems offer the possibility of queuing callers for a busy terminal, called 'camping on busy,' but generally a subscriber who finds another engaged has no alternative but to start all over again later.

A message switching system using store-and-forward switching obviates the need for repeated attempts at establishing a call by accepting the message and undertaking to deliver it when delivery becomes possible. Sometimes different grades of service are available, and an express message will be guaranteed an attempt at delivery at the earliest possible moment. Often, in a large organization, the Telex service can appear as a store-and-forward system because a single terminal and operator is made to serve the whole organization through the internal mail. In this case, the behaviour to the user is indistinguishable from a message store-and-forward service to which he has a direct connection. Finally, whenever more delay is tolerable, the ordinary public mail service may be used; this is, of course, also a store-and-forward system.

In addition to the possibility that the called subscriber may be unavailable or engaged, there is a chance that the communication service itself may be congested. With a circuit-switched network the result of congestion is the inability of one subscriber to make a call

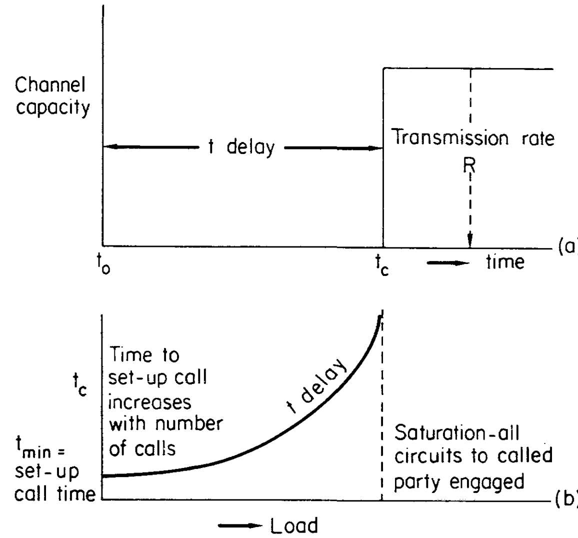


Figure 7. Characteristics of circuit-switched networks

to another, even though that subscriber would be quite able to accept a call.

The bottleneck may be in the switches or, more often, in the junction or long-distance trunk circuits; but either way, the appearance to the user is as shown in Figure 7. There is a connection delay that increases with the total load on the network as given at (b). But, as is shown at (a), once a connection is established the information exchange may take place at a rate R governed only by the physical properties of the particular circuit and the subscribers themselves or their equipment.

With a store-and-forward system an increase in the overall load naturally results in congestion, but the appearance to the user may not be immediately apparent, for the system continues to accept messages, at least for a while. The local switching center usually can take and store further information, but may not be able to forward it for some time, although eventually, of course, no further messages will be accepted until some relaxation of the congestion allows those taken earlier to be despatched. With the mail service, the normal outward sign of congestion is the increase in delay. For example, at Christmas, letters take longer to arrive; but the peripheral stores, in the form of public post boxes, are so large that they never, under usual circumstances, refuse to accept messages. However, during the strike of British Post Office workers in 1971, the boxes had to be sealed to prevent any input into the system while it was not operational. The appearance to the user of a store-and-forward system is illustrated by Figure 8. There is a minimum time necessary to transfer a message through the network. This comprises the time to travel between switches at the transmission speeds of the intervening links and a minimum message handling time at each

The Uses of Message Switching in Communications Networks

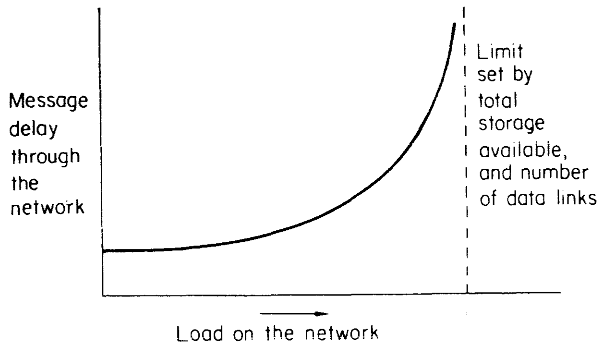


Figure 8. Characteristics of message-switched networks

switching point, including the time taken to read a message into and out of the stores associated with the switches. As the load on the network increases, queues develop at switches while messages wait to be processed or for links to become free. This causes an increase in the time for messages to pass through the network as shown in the diagram.

Another way of looking at the store-and-forward network is to regard it as a large distributed set of stores that gradually fill as the load on the network increases. The peak load is therefore shifted in time by an amount depending on the available storage. Figure 9 illustrates this point by showing how a peak load is reduced, but at the expense of being delayed.

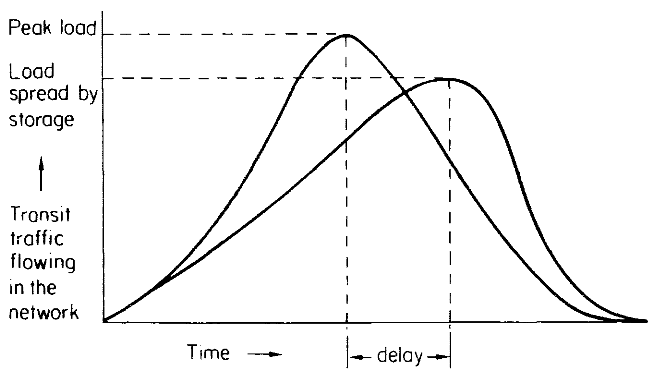


Figure 9. Effect of storage in message switch

The area under each of the two curves represents the total number of messages handled and is, of course, the same for both curves. But, in a practical network, there will be some upper limit to the instantaneous capacity of the network, perhaps due to limited transmission speeds or switching capability. Clearly, the peak demand on the network is less with the increased use of storage that spreads the load over a longer time period. However, the delay increases as the load is spread, and the time taken for messages to pass through the network also increases, as was shown in Figure 8.

If the load on almost any public service is plotted over a twenty-four hour period, a demand curve similar to Figure 10 is obtained. The capital cost of providing the service is governed by the peak amplitude of the demand curve, and it is obviously an advantage if the curve can be made more uniform in amplitude by shifting some of the peak load to other times. The use of differential tariffs with reduced rates outside peak hours is a means of encouraging a shift of peak traffic to less busy periods.

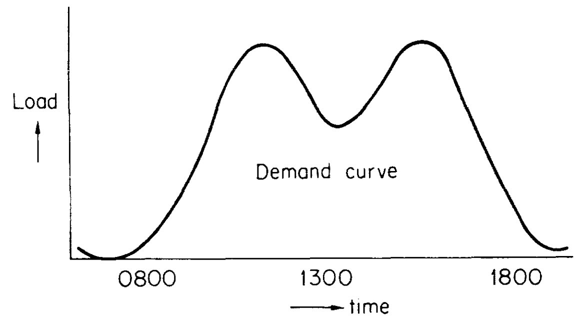


Figure 10. Load on a public service

If the demand curve is examined more closely in a reduced time interval, it is seen to combine the sum of very many individual demands with an average value that fluctuates rapidly. The upper limit of this fluctuation represents the peak demand, and there is an advantage to be gained by smoothing these rapid fluctuations to reduce the peak.

With the telephone network, the smoothing occurs because subscribers obtain a busy tone when they dial a call and try later: this is the way a circuit switched system appears to users at peak load times. With a store-and-forward network, the messages will be accepted (if the store at the first switch is large enough), but delivery will be delayed. In the latter case, the originator of the message is spared the task of repeatedly trying to begin an interaction with another subscriber, but the interaction will proceed more slowly as the network load increases. With the circuit switch, the start of an interaction will be delayed as the load increases, but once begun it may proceed at a speed unaffected by loading on the network.

THE BEHAVIOUR OF DATA SYSTEMS

In discussing the characteristics of networks, an analogy was made between a human being and a multi-access computer to describe the way people communicate with each other. But the computer works much more rapidly, so, although the problem of meeting the communication needs of computers is essentially very similar to that of suiting people, the time scale is much shorter. In fact, three conditions may be identified: interaction between a person and a computer; interaction between two or more computers; and interaction between a computer and a remote sensor or peripheral device.

The Uses of Message Switching in Communications Networks

To a large extent, the interaction between a person and a computer is governed by the characteristics of human beings. The average rate of input and output obviously are determined entirely by the person (unless the whole system of communication links and remote computer is so ill-designed as to hamper his performance), and the nature of his interactions will depend on the job he is doing. The present level of sophistication reached in the design of interactive systems makes it necessary to have a frequent interchange of messages between the user and the computer. The kind of interaction that can occur during a transaction with an information service is shown in Figure 11. Notice that quite short commands (underlined) from the user can result in long replies from the computer, and that these may require some thought on the user's part before he again instructs the computer. In essence, therefore, the interactions are sporadic and are characterized by a variability of lengths with a tendency towards short messages. This has been observed experimentally by workers at the Bell Telephone Laboratory.[3]

```
HONEYWELL


ON AT 10:10 G265 A 27/04/72 TTY 7

USER NUMBER --
SYSTEM -- BASIC
NEW OR OLD -- NEW
NEW FILE NAME -- DLAB
READY.


10 PRINT "X" , "X↑2"
20 LET X=1
30 LET Y = X↑2                    Note  User's input is underlined
40 PRINT X ,Y
50 LET X=X+1
60 IF X>10 THEN 80
70 GO TO 30
80 END

RUN


DLAB     10:15     G265  A  27/04/72

X          X↑2
 1          1
 2          4
 3          9
 4          16
 5          25
 6          36
 7          49
 8          64
 9          81
 10         100
```

Figure 11. An interaction with a computer bureau

The interactions that take place between one computer and another are somewhat different from those between two people or between a person and a computer. To some extent, the form of interchange depends on the tasks performed by the computers and, indeed, on the way the exchanges are constrained by the intervening data links. If both computers are multi-access computers engaged in handling transactions and a particular transaction demands that one system communicates with the other, the interchange is likely to, in the form of short messages, transfer data from one system to the other. On the other hand, one computer may need to transfer a complete file to another machine, perhaps to up-date information about a company's operations. In this case, it is often assumed that much longer messages would pass between them. Even so, it is likely that the messages would be broken-up into shorter blocks with intermediate validity checks carried out to make sure all was well, as it is generally unsafe to rely on the links' ability to provide error-free data transfers.

When a validity check is to be made, the use of short blocks has two advantages. Firstly, it can economize on high-speed storage in the two computers, for it reduces the amount of store occupied by a data block at the sending computer while a copy is transferred to the other machine and checked for correct arrival. Secondly, a short block has less chance of being corrupted by a line disturbance than a long block, so the likelihood of having to retransmit it is lower. For these reasons, the shortest possible block size is to be preferred. However, when the computers are far apart, the long signal propagation time between them makes it desirable to use long blocks to maximize the quantity of information transferred for each handshake acknowledgement. A compromise block length therefore has to be accepted.

Apart from the advantages attached to the transfer of long data files by relatively short messages, the need for long transactions between computers is likely to be reduced as more on-line usage is introduced. At present, much updating of files occurs in a batch processing manner after business hours are ended. This calls for the transfer of long files. However, if the records were amended on a transaction basis as the business was carried out, there would be no need for the subsequent updating of files. As there is, potentially, an economic advantage in keeping up-to-date files, the trend towards on-line operation with short transactions may be expected to continue.

The interactions possible between a computer and remote peripheral devices other than those used by human operators could be of many kinds. There are a growing number of private networks for telemetry and control of oil and gas pipelines, electricity grids, and so on, and there are already a few cases where the

The Uses of Message Switching in Communications Networks

public telephone network is used; for example, river level measuring devices are available that can dial-up a flood-warning center when a critical water level is reached. The interactions occurring between such devices are also generally likely to be of short duration; but in any case, the total volume of traffic arising from such sources is unlikely to be great, though the reliability and accuracy required are high.

It was the appreciation of the likely characteristics of the data traffic that would arise with remote access computer systems and the probable rapid growth in the volume of such traffic that led to proposals for packet-switching data networks. But before examining these proposals it is desirable to consider in more detail the basic characteristics of data traffic.

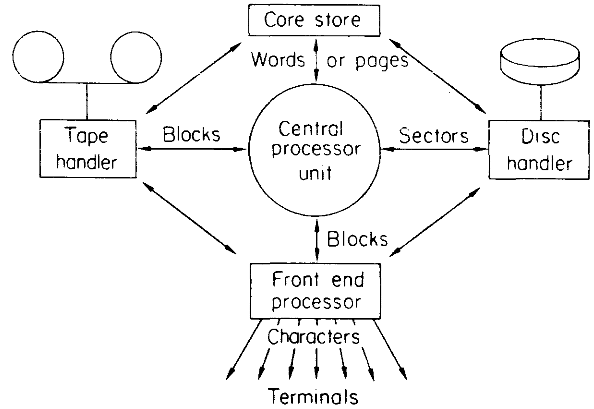


Figure 12. Internal message paths

THE CHARACTERISTICS OF DATA TRAFFIC

It is necessary to understand the nature of data traffic to enable the appropriate choice to be made between possible alternative methods of carrying it, and to examine whether new techniques need to be developed to handle the traffic more effectively.

The behaviour of computer-based systems was shown to depend on three factors; the computer itself, the remote peripheral devices it serves, and people making use of the services it provides.

A computer is characterized by the handling of information internally at very high rates, typically tens of Megabits per second. The operation of the various peripheral devices is scheduled to keep them as active as possible, and this leads to the transfer of information in relatively short blocks of up to a few thousand bits. When a computer is connected by telephone lines to remote terminals, the disparity between the internal speed of the computer and the lines makes it necessary to introduce storage and sometimes auxiliary communications processors. The communication between the main computer and the storage or auxiliary computers may then be of the same form as other internal traffic, i.e., relatively short blocks. This is illustrated by Figure 12, which shows information flow between the central processing unit and the other parts of the computer.

The behaviour of peripheral devices depends on their principle of operation. Mechanical devices like printers and paper tape punches and readers are constrained by strength of materials to operate up to a few thousand bps. Magnetic tapes and discs can range from a few tens of kbps to a few mbps, while electronic devices, like cathode-ray-tube terminals, may be operated at speeds approaching that of the computer itself. Very often peripheral devices contain buffers to decouple them from the data link. With slow mechanical peripherals, the buffers fill rapidly through the link, which is then idle while the device reads the data from the store. Where the link is dedicated to such a peripheral device, little advantage is gained from the use of buffer storage; but when several such devices are grouped, the use of storage allows the link to be shared by them. The flow of data to a terminal fitted with a buffer is illustrated by Figure 13.

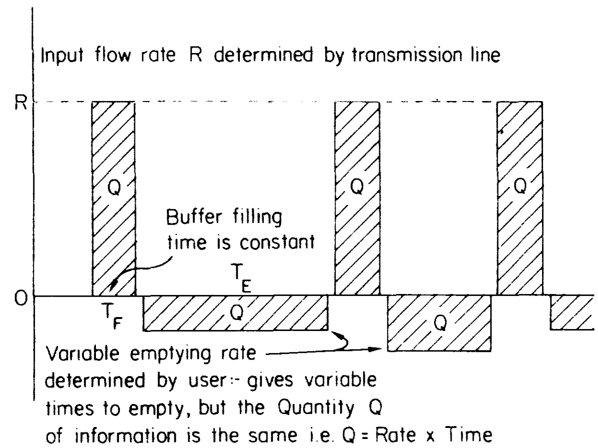


Figure 13. The use of a line buffer

The third factor affecting data traffic is the way people behave when using data terminals. If they are using keyboards, the rate will be only a few tens of bps and it is often the case that buffers are added to collect a block of data rather than allow characters to be sent as they are typed. Again, this is only useful if one data link, or computer input-output channel, can be shared between several keyboards by interleaving the blocks that each produces. The rate of receipt of information by people depends on how it is presented. Pages of text, if read carefully, would require an average rate of transmission of about two hundred bps. However, sometimes people scan pages quite quickly, and it is preferable to have them produced rapidly; there is then a pause before the next is required. Much the same happens when graphical information is presented; it is usually better for a picture to appear at once and then be held while the user examines it at leisure.

From the above considerations it is apparent that data traffic, although covering a range of characteristics depending on its source, is broadly characterized by a tendency to fairly short messages often exchanged at quite high speeds. This is illustrated in Figure 14. This figure shows qualitatively how the number of messages varies with the length of message averaged over all types of system involving data transmission; both the message length distribution and the cumulative distribution are shown, and it is clear that the majority of messages are less than about 2,000 bits in length.[3]

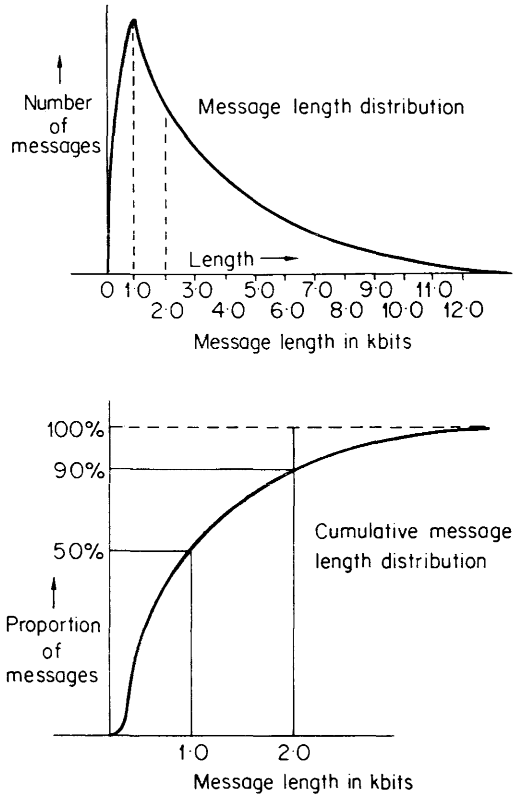


Figure 14. Message length distribution

The time taken to transmit messages of various lengths is shown in Figure 15 where the time is plotted against message length for different speeds of transmission. The loop delays for 10, 100, 1000, and 10,000 mile circuits are also shown. It is worth noting that, for a message length of 1,000 bits and a transmission speed of 48 kbits/s, the message length is equal to the number of bits that could be stored in a 1,000 mile circuit loop.

Because the message transmission times are so short, the use of circuit switching is really inappropriate, for

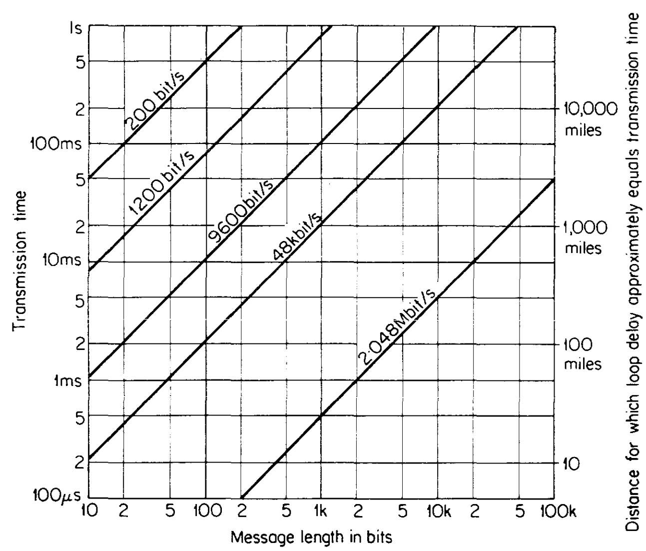


Figure 15. Message transmission speeds

the time to set-up and clear circuits is much longer than the message transfer time. Indeed, the end of a message may well have left the transmitting device before the beginning reaches the receiver. The message switching principle is more appropriate for data traffic, particularly the high-speed, fixed-format message technique known as packet switching.

THE PACKET SWITCHING PRINCIPLE

The principle of packet switching was first put forward in the early 1960's by Paul Baran while working at the RAND Corporation on military voice communications networks.[4] Baran proposed a mesh network of computers joined by communications circuits that handled digitized speech in short bursts. The mesh form of network gave many alternative paths between any two subscribers trying to communicate with each other, and an adaptive routing method called the hot potato technique was proposed to give reliable communication even if parts of the network were destroyed by enemy attack while calls were in progress. Only when no possible path remained were two subscribers prevented from conversing.

Each burst of speech was preceded by a destination address and was directed by each node in the general direction of the destination. The nodes handled the bursts like hot potatoes, getting rid of them as soon as possible. This gave a rapid progress of information through the network, which behaved as a very high speed store-and-forward information switching network. The speed of processing required to handle speech in this way is very high, and the delays in mesh networks using store-and-forward techniques are generally rather long for satisfactory speech communication. But it seems very likely that improving technology will eventually change this situation.

The idea of a mesh network that handles information in short bursts seems very well suited to the communication needs of remote access computer systems, and, since 1966, the concept has been developed both theoretically and practically at the U.K. National Physical Laboratory, in the context of a possible National Data Network.[5] The technique has also been used in private networks, notably that built by the Advanced Research Projects Agency (ARPANET) in the U.S.A.

Packet switching is often regarded as a particular form of message switching especially suited for handling data traffic. Superficially, this is indeed the case, but there are very important differences between conventional message switching and packet switching, and the latter is, in reality, an entirely separate method of communication. In fact, there are resemblances between packet switching and circuit switching also, because the packet switching principle combines those major advantages of both circuit switching and message switching that are most appropriate for handling the interchange of information between computer systems.

With conventional message switching systems, messages of any length are accepted in their entirety and are stored as such at each switching point during their passage from switch to switch towards their destination. The network takes full responsibility for maintaining the integrity of the message, and elaborate procedures are employed to ensure that this is achieved. The accent is on reliability, rather than speed, in the information transfer between subscribers, and the messages are held by the network until the recipient is ready to accept them, however long this may take. It is not intended that subscribers should interact rapidly with each other through a message switching network, so the type of interaction common between telephone users to overcome errors introduced by the network, or indeed their own mistakes, is not possible. This is why the accuracy of information transfer through a message switching network is so important.

With packet switching systems, information is exchanged in the form of short packets, and the transit time of these messages through the network is kept low; the subscribers are expected to interact with each other by exchanging packets, in much the same way as they would interact by exchanging information through a circuit-switched connection. Because the subscribers interact together, they may take part in the validation of the information exchange procedure; this can cope with their own errors, as well as any introduced by the network. However, this does not imply that a packet switching network is unreliable; indeed, because information is not retained in it for long periods the chance of information being corrupted while it is static is reduced, while the probability of corruption during transmission can be the same as for a conventional message switching network.

It is worthwhile pointing out here that long messages are readily handled by a packet switching network if they are broken up into short packets for transmission and reassembled again at the destination. This method of operation is often called cut-through and is shown in Figure 16. The use of cut-through has important advantages: the fixed packet structure permits efficient handling, and the absence of indefinitely long messages prevents the blocking of transmission links and keeps the queues at switching points small.

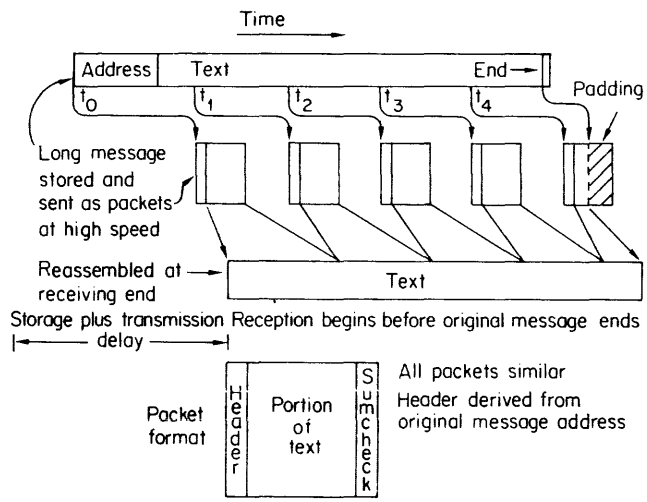


Figure 16. The packet switching principle

Storage at switches is made sufficient only for a few packets, and the total amount of information stored in the network is low. The result is that the delay through a packet switching network is much smaller than through a normal message-switching network, and the rate of throughput of information can be much higher. Where message storage is required as a service to subscribers, it must be provided externally to the network rather than as an integral part of the switches in the manner usual in a conventional message-switching network. This separation of the storage in the network into two parts, that used primarily for scheduling the efficient use of communication lines and that used for providing services and storing messages, clearly confers some very important benefits for handling data traffic, and a considerable amount of attention has recently been given to the design of packet switching networks.

A COMPARISON OF MESSAGE AND PACKET SWITCHING

The difference between packet switching and message switching begins with the messages and packets themselves. Messages are the units of information recognized by the users of networks. Consequently, their length must be moderately unrestricted. To make them suitable for formatting by people, messages usually have a format depending on certain markers or codes.

Packets are designed for handling by computer and so have a fixed format. To keep their transit time low, a maximum length is set, and if messages longer than this maximum have to be carried, they are split into packets for transmission. Their format is not suitable for use by people.

Message switching emphasizes the responsibility of the network for the message, because no response or feedback is expected—the delivery of one message is a complete transaction. Because message systems were not associated with computers but with big organizations, the need to replicate messages, sending identical copies to many locations, has been met. Computers would distribute information selectively, not using replication. Packet switching serves interactive systems where a response is expected, so the loss of a packet is generally less troublesome than a data error introduced in transmission.

Packet switched networks aim to deliver packets with minimum delay and do not hold them for delayed delivery. If the called terminal (or computer) is busy, the conversation will probably not even begin, but when it does begin it is likely that many packets will be exchanged before it ends.

The kind of transit delay expected in a packet switched system can be estimated roughly by noting that a 2048-bit packet can be sent over a 2048 Mbps link in one millisecond. The transit delay for one node will be a small multiple of this service time, determined by the queue length plus the processing time for each packet. Because of the fixed packet size, service time is constant, and it can be shown that even for 80% saturation the mean queue length is two (not including service position), so the mean queuing component of the transit time is 2 milliseconds. The processing component is likely to be comparable with this time also. In practice, a much lower occupancy is used, less than say 25%, to provide a margin for error in traffic estimation. The contribution of queuing to mean transit time for a U.K. national network with such links is less than 5 milliseconds, and 10 milliseconds is rarely exceeded. This is in marked contrast with message switching, where transit delays of minutes are normal.

So, although the store-and-forward principle is employed in both kinds of systems, their objectives are different, and the performance expected is very different.

SHARED PRIVATE NETWORKS

Essentially, private data communications networks with lines leased from common carriers use store-and-forward message switching techniques and, being entirely the concern of a single organization, may use any message format and other standards most suited to the particular needs of the organization. The Western Union INFOCOM service is typical of several other message switching networks that attempt to serve the needs of more than one organization. This task is complicated by the problem of agreeing on standards for a wide range of users, and this is made even more difficult if the basic problem of transferring messages between subscribers is combined with that of offering services that are provided as an integral part of the network.

However, because a group of organizations with similar requirements is more readily able to devise common standards, there are a number of shared private networks in use or planned. There will undoubtedly be many such similar, but incompatible, data networks in a few years time. This situation is likely to worsen unless the PTTs manage to agree on really comprehensive, advanced standards and begin to implement public data communications networks.

Of the many shared networks now in existence, two have been chosen for closer study; one is the airline network established in Europe by the Société Internationale de Télécommunications Aéronautique, the other is the Advanced Research Projects Agency network in the U.S.A.

The SITA Network

The Société Internationale de Télécommunications Aéronautique (SITA) was originally established in 1949 by a group of airlines as a non-profit making organization that would provide them with a cheaper means of exchanging messages. This was needed to facilitate the sale of seats on their aeroplanes, to exchange operational information, and for the location of baggage that had gone astray. For these purposes, SITA organized a common service in the form of a worldwide low-speed message switching network handling teleprinter traffic; this was well established by the time computer-based airline seat reservation systems began to be implemented. Because of these new systems, a review of the SITA network was undertaken. In 1964, the decision was made to adapt its design to keep pace with the new patterns of traffic that were expected to arise during the next ten years.

The anticipated traffic to be handled by the modified SITA network was of three kinds:

The Uses of Message Switching in Communications Networks

1. Type A: data traffic between computer systems requiring a rapid response; this would comprise single address messages using various types of 5, 6 or 7 unit information codes.

2. Type B: conventional teleprinter traffic using either CCITT alphabets No. 2, or No. 5, with messages having single or multiple addresses and of lengths up to 4000 characters.

3. Type C: single address data traffic requiring a response time similar to type B, and using CCITT alphabets No. 2 and No. 5.

The transit times through the network for type A traffic was to be of the order of 3 seconds, while the other two categories would have three levels of priority indicated by a label—QU, 2 minutes; QN or no label, 30 minutes; QD, 12 hours. All messages handled by the network were to be protected from loss or mutilation, and facilities were required for the repetition of type B and C messages that had already been delivered and for holding them if the destination was unable to accept them.

The computer-based seat reservation systems already in use controlled their remote terminals by polling through networks of leased lines. The redesigned SITA network was required to provide for handling this kind of terminal. In addition, the new network would have to match into the existing telex systems operated by PTTs and into other shared networks such as its American counterpart ARINC (Aeronautical Radio Incorporated), which serves some 90 U.S. airlines, and any private networks operated by companies such as BOAC, PANAM, and KLM, to name just a few.

When the SITA network was being reappraised in 1964, it was already connected to over 100 centers, and about 100 million conventional telegrams per year were being handled. This number was increasing at a rate that would double it in less than four years. To handle this growth in existing traffic and to satisfy the anticipated new requirements, a high-level or trunk network was planned; this would provide an improved service and new facilities to a large part of Europe, and also North America, and would join together areas of smaller message concentration using the existing communications methods. High-level network centers were chosen in New York, London, Paris, Amsterdam, Brussels, Rome, Frankfurt, and Madrid, each having a defined area for message collection and delivery. These main centers were to be joined by data links, which needed to operate at, at least, 2400 bps to achieve the short response time specified for type A traffic. This response time criterion led to a decision to use store-and-forward block transmission in the high-level network with blocks of variable length up to a maximum determined by the response time and transmission efficiency. Each block would have a header giving addresses and control information to guide it through the switching centers, and the overall efficiency would, clearly, be increased by using longer blocks for a given header size. On the other hand, the shorter the block, the less likely would be the chance of its corruption by noise, with the consequent need for its retransmission. Also, shorter blocks would allow a more rapid response for high priority messages, because the waiting time for a lower priority block to end would be shorter. However, an overriding factor that influenced the choice was the average message length of the existing telegraph traffic, which was less than 200 characters. Indeed, 95% of all messages had less than 250 characters. It was, therefore, decided to use a 256 character block with a 250 character information field available for carrying message. A long message would be transmitted as a sequence of two or more blocks.

The requirements to handle codes of up to seven units led to the adoption of an eight-bit character with seven information bits and an eighth bit giving overall odd character parity. An alphabet similar to CCITT No. 5 was chosen for control purposes, and all other characters in the information field were to be padded up to seven bits where necessary. For example, with the existing telegraph messages in alphabet No. 2, the start and stop bits would be removed from each character, and the remaining five information bits would be augmented by a sixth bit indicating the shift case and a seventh bit which was always 'one.' This would prevent an alphabet No. 5 control character being falsely simulated by a message character.

Following feasibility studies,[6] a system was designed, and a specification was prepared for the computer centers. This allowed them to be introduced into the existing network of low-speed circuits in such a way that modification to form the high-level network was readily possible later. Univac 418 II systems were chosen for New York, Frankfurt, Brussels, Rome and Madrid, while Phillips DS714 Mk II systems were installed in the larger London and Paris centers, and at Amsterdam. The high-level network shown in Figure 17 was completed in 1970 and has been in service ever since.

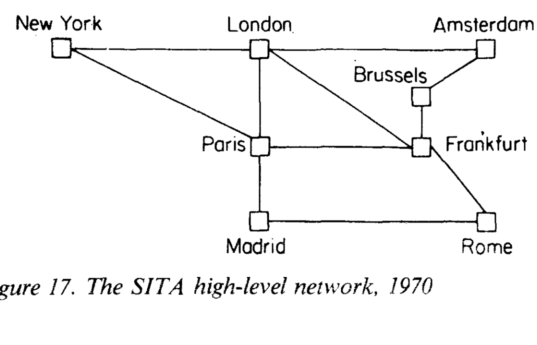


Figure 17. The SITA high-level network, 1970

The Uses of Message Switching in Communications Networks

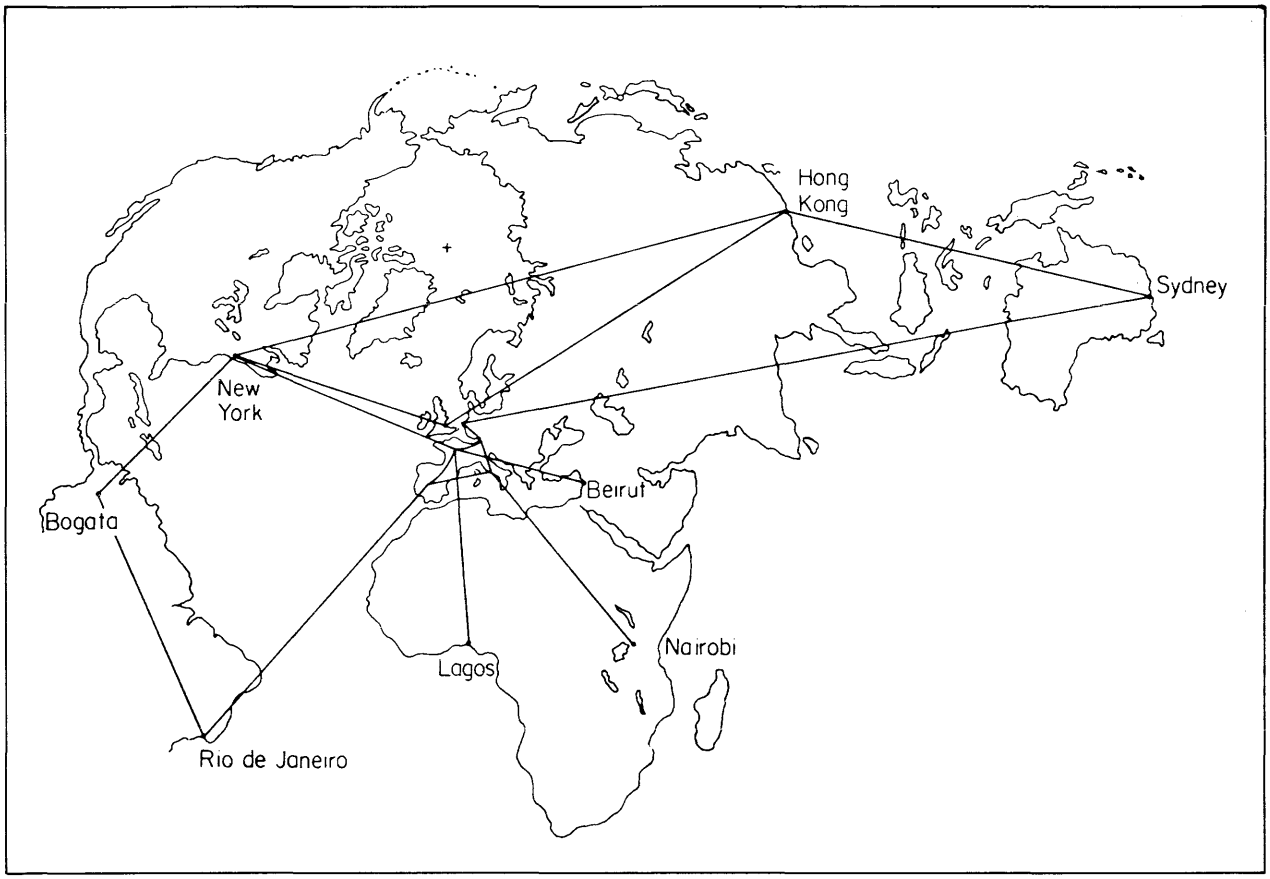


Figure 18. Some planned SITA developments

In parallel with the development of the main network, the existing networks were progressively modernized, a process that is still continuing. At first, the manual message switching centers were replaced by multiplexers capable of handling several low-speed devices such as teleprinters, and later satellite processors were introduced. These are able to cope with a wide variety of traffic and can control the input-output terminals designed specifically for airline automatic systems. Figure 18 shows some possible extensions of the SITA network likely to be made in the next few years.

The High-level Network

The SITA high-level network was designed with three aims in view. Firstly, to improve the service for conventional traffic; this has been done and transit times have been reduced from hours to minutes. Secondly, to provide a service for handling short data messages; this, too, has been done and has enabled British European Airways to extend its data transmission network into Europe. Thirdly, to offset rising cost of conventional communications facilities; this has more than been achieved, for the efficient use of circuits and the reduced number of operating staff has brought about a fall in the cost per message.

Because the SITA network has proved so successful, it is worthwhile examining the procedures devised for ensuring the integrity of the messages it handles. This is done on two levels; the individual blocks passing along each link are protected against mutilation, loss or duplication, while complete messages are similarly protected between their points of entry to and exit from, the network.

The protection of blocks in transit between centers is achieved by adding to each block an even parity check character; this augments the odd parity bit used with each character and much improves the chance of detecting errors. As long as blocks are available, they are sent continuously and are checked for correct parity at the receiving end of a link.(If no blocks are available the transmitter sends a link message every three seconds.) An acknowledgement is generated for every received block indicating whether it was correct (ACK) or incorrect (NAK); these acknowledgements are interleaved between blocks being received on the return channel. If a full length block were being received on the return channel, the acknowledgement for several short blocks sent on the forward channel would be delayed. A block numbering scheme is, therefore, employed and the acknowledgement signals carry the number of the last block correctly received.

The Uses of Message Switching in Communications Networks

The loss or duplication of blocks is prevented by arranging for each transmitting center receiving a correct acknowledgment (ACK) to erase the associated information blocks, while if an incorrect acknowledgement (NAK) is received the center repeats all blocks with numbers following that carried by the (NAK) signal. At the receiving end of a link, the detection of a faulty block inhibits reception, and all further blocks are rejected until the block previously found to be faulty has been correctly received.

If after retransmission a series of blocks is not correctly acknowledged, it is repeated a further three times. If a correct acknowledgment still fails to appear, the communications circuit is assumed to be faulty, and traffic is diverted to another route. Meanwhile, a print-out is produced advising operators of the situation, and check messages are sent continuously into the faulty circuit until it is found to be operating correctly again.

When a link between centers becomes faulty or is restored following a fault, a status message is sent automatically to all other centers. This message causes all routing tables in the network to be modified accordingly. Should all the links to a center, or indeed the center itself, fail, the low priority messages for the center are stored at entry points until the center becomes operational again. However, the high priority messages must be dropped, and the polling of all remote terminals belonging to computers served by the out-of-action center has to cease.

The automatic rerouting of blocks during fault conditions makes it possible for some blocks of a long message to arrive out of sequence, and to overcome this problem an entry-exit block numbering scheme is used. This operates in a similar manner to that used on individual links, but a separate series of numbers is used for each type of traffic and for each pair of network centers. The entry center holds all blocks until they have been correctly acknowledged as received by the exit center. Individual blocks are not acknowledged, but there is at least one acknowledgment provided for every 16 blocks. The entry-exit block numbering arrangement enables the blocks of a multi-block message to be reassembled correctly and ensures that none are lost or duplicated.

It is interesting to contrast the SITA network that has evolved from the needs of a specific group of users—the airlines—with the ARPA network, designed as a research project, which is discussed in the following.

The ARPA Network

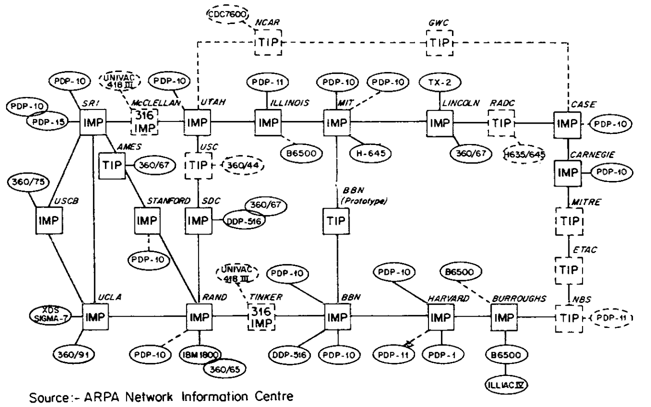
The Advanced Research Projects Agency is operated by the Department of Defence of the United States. It was formed to promote advanced projects of all kinds, following the early and unexpected success of the U.S.S.R. in launching the first man-made earth satellite. Computer science was an obvious field where new projects could readily be conceived; the renowned Project MAC at the Massachusettes Institute of Technology was one, ILLIAC IV at the University of Illinois another, while several other computer projects were also begun at universities all over the U.S.A. Of these, some twenty continue to be supported. In 1966, at MIT, experiments were conducted with two computers joined by a data link; this work led to the proposal for a network of data links connecting several of the centers where ARPA funded projects were in progress. This ARPA network, begun in 1967, is possibly the most important computer project in the world today, because it links together and is beginning to coordinate the work of advanced research establishments spread all over the U.S.A. Its role as a cohesive force associating otherwise separate groups of research workers is a most important aspect of the project.

When initially conceived by L.G. Roberts[7] in 1966, the ARPA network was based on the use of 2.4 kbps data links, but further consideration, encouraged by the theoretical studies that had been carried out independently in the U.K. by the National Physical Laboratory, led to the adoption of high-speed group-band links. These now connect small computers in a mesh network forming a packet switching communications system shared by the sites it joins. All the sites are funded by ARPA, and the network is, therefore, a private shared network. However, it is, at present, unique in the use of high-speed lines, which give a coast-to-coast transit time in the region of 100 ms. This is much less than is achieved in other high-level networks—such as the SITA network—and allows new concepts in interactive computer-to-computer communications to be developed. In fact, most other data networks have been designed for connecting distributed terminals to central computer systems, whereas the ARPA network was intended from the beginning to link a relatively large number of advanced data processing systems, many of which already supported their own networks of terminals.

Initially the ARPA network comprised two parts: a network of data processing systems called HOSTs and a communications subnetwork of packet switching node computers known as IMPs (Interface Message Processors). In August 1971, a new type of node computer for handling terminals was introduced for sites that had no HOST system. This TIP (Terminal Interface Processor) makes the powerful new services, which are being developed on some of the HOST systems, readily accessible at sites which could not otherwise afford such facilities.

Because the ARPA network is itself a research project, it is developing continually, but a typical example

The Uses of Message Switching in Communications Networks



Source:- ARPA Network Information Centre

Figure 19. The ARPA network, 1972

of its structure (in mid-1972) is given in Figure 19. This shows the large number of IMPs and TIPs already installed and indicates the variety of computer systems that can now be joined together. The communication sub-network has undergone some changes since it was originally designed, but essentially it has proved very effective in coordinating research at the centers it connects. These centers are developing techniques for using the network effectively, and several projects are being conducted aimed at providing new services and facilities that would not be feasible without the sub-network. The design of the ARPA communications network and the way it is being used are discussed in the following.

The ARPA Communications Sub-network

The contract for the design and implementation of the communications network was awarded to Bolt, Beranek and Newman of Boston, Mass. They designed the IMP shown in Figure 20 around an augmented, ruggedized, version of the Honeywell DDP 516 computer using 12K words of store with 16 multiplexed channels, and 16 levels of priority interrupt.[8] Special hardware to control interfaces to HOSTs and data links, and to monitor the IMP's internal behavior was designed by B.B. & N.

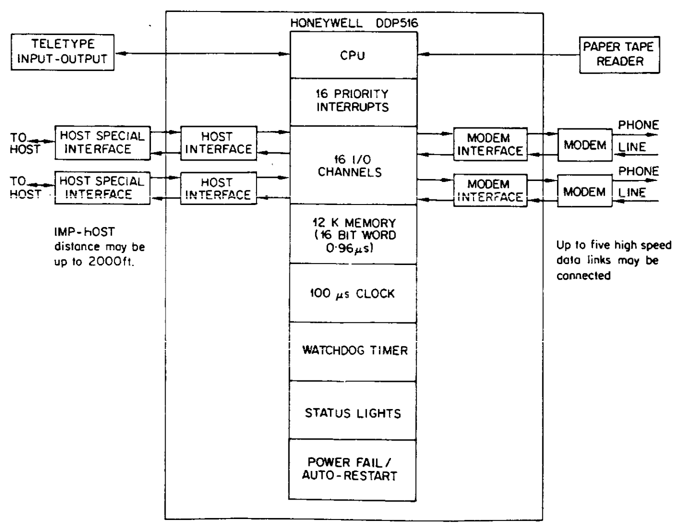


Figure 20. The ARPA IMP

An IMP can serve up to four HOSTs, provided the total number of HOSTs and data links does not exceed seven. An identical operating system, occupying some 6K store words, is used for all IMPs leaving 6K words for message storage; but a protected 512 word block contains programs which allow an IMP detecting an internal software corruption to reload a copy of the operating system from another IMP.

The messages passing between a HOST and an IMP are controlled by a HOST and IMP protocol, and may vary between 1 and 8095 bits in length. These messages are partitioned into packets with a maximum size of about 1000 bits, each having a cyclic sum check added by hardware during transmission into a data link. The packet format is shown in Figure 21. The particular link used for each packet is selected by the IMP according to its estimation of the transit time to the packet's destination using each available link. The transit time estimates are recomputed every half second and are based on an interchange of estimates and past records with neighboring IMPs. The dynamic reestimation of transit delays allows links to be selected to minimize the transmission delay, and to maximize the total throughput of the network; but, of course, the successive packets of one message may follow different paths.

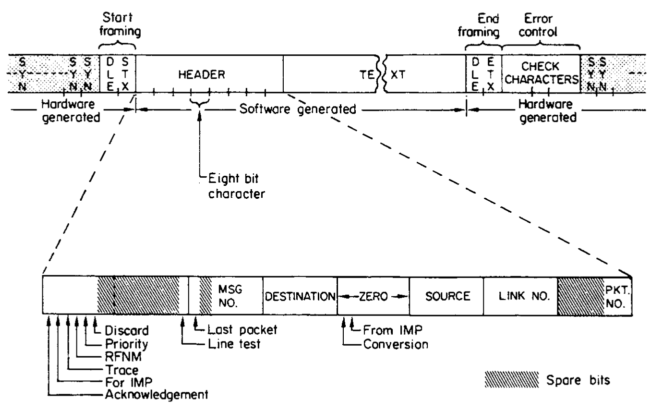


Figure 21. The ARPA packet format

When an IMP receives a packet from another IMP an error check is performed. If the packet is error free, it is stored, and a positive acknowledgment is returned to the sending IMP allowing it to release the storage area occupied by its copy of the packet. However, if the received packet is in error or if the receiving IMP is unable to accept it, the packet is ignored. The sending IMP waits a preset time for a positive acknowledgment; if one does not arrive, the packet must be assumed lost, and a copy is retransmitted, possibly by another route.

The Uses of Message Switching in Communications Networks

Once a packet has been stored and acknowledged, a receiving IMP must determine, by examining the destination address, whether the packet is to be delivered to a local HOST or forwarded to another IMP. In this latter case, the packet is placed in the queue for transmission in common with any outgoing messages from local HOSTs. But if the packet is for local delivery, the IMP checks to see whether all associated packets have arrived; if so, it reassembles them into the proper order and delivers the complete message to the HOST. The HOSTs control the interchange of message between themselves using a HOST to HOST protocol.

As well as handling messages, an IMP is arranged to detect and report data link failures and also gather performance statistics. In the absence of normal traffic, each IMP transmits idle packets on unused lines at half-second intervals. The lack of a return packet or incoming traffic on any line for more than a preset time indicates a faulty line and allows routing tables to be updated accordingly. The return to normal of the line is indicated by the resumption of idle packet traffic. Internal performance statistics are also collected by each IMP and are automatically transmitted to a specified HOST for analysis. This HOST is thereby able to formulate a picture of the overall state of the whole network. A useful feature novel to the ARPA network is the trace message. Any HOST message may have a trace bit set; when each IMP processes any packet of such a message it records the packet arrival time, the queues on which it resides, the time spent in these queues, and the departure time. These records are also sent to a specified HOST for evaluation; they allow a very detailed picture to be formed of how the network handles messages.

The Terminal IMP (TIP)

The Terminal IMP is a lower power version of an IMP based on the Honeywell DDP316 computer, fitted with a multi-line controller (MLC) designed and built by B.B. & N. to handle up to 64 channels for asynchronous terminals; these may be connected directly, or through telephone circuits, see Figure 22.

The MLC has a cycling buffer store that samples each channel every 50 us, giving a maximum rate for output or synchronous input of 19.2 kbps. However, the aggregate rate must not exceed about 100 kbit/s, although a wide variety of terminals of differing speeds may be mixed together within this overall limit. The asynchronous format uses start and stop bits, and because each character is sampled eight times, the maximum rate is 2,400 bps for this type of terminal.

The TIP is designed to appear to the main network as a HOST, so the TIP program must implement the HOST to IMP protocol internally between its ter-

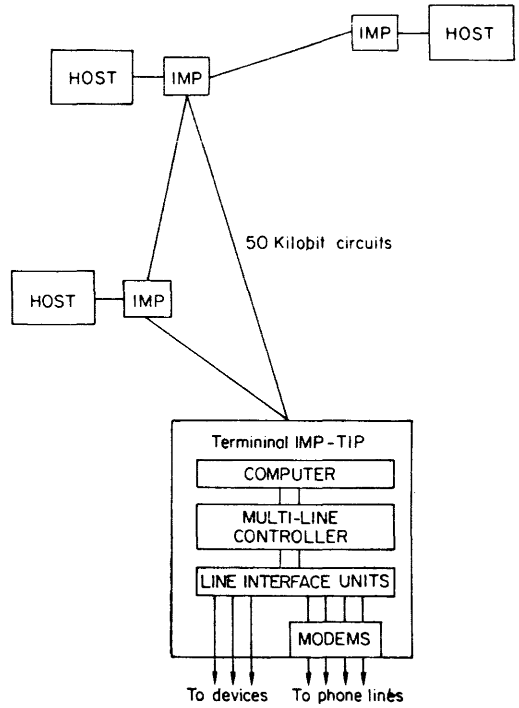


Figure 22. Terminal IMP connections

minal handling, and trunk-line handling partitions. In addition, it has to satisfy the HOST to HOST message exchange protocol and the TELNET protocol designed to convert the messages from terminals—which of course, use their own codes and procedures—into a common network form called the Network Virtual Terminal (NVT). This allows terminals to use services with which they could not communicate directly because of hardware incompatibilities.

An interesting method is used to handle the variety of terminals that may call a TIP. As soon as a connection has been made to a channel of the MLC, the user enters a character specific to that type of terminal. The MLC samples this at a high speed, generating a bit pattern which identifies the terminal uniquely. The TIP software then sets the correct sampling rate for the channel and selects the code-conversion tables appropriate for the terminal type.

Applications of the ARPA Network

Once the communications sub-network was in operation, it encouraged people to think positively about ways in which it could be used. It had always been intended to gather information about the behaviour of the sub-network and this is being done by the Network Measurement Center at the University of California, Los Angeles (UCLA), which is responsible for system modeling, and the Network Control Center operated by B.B. & N. in Boston, where the states

The Uses of Message Switching in Communications Networks

of the IMPs, the lines and HOSTs are monitored. As a result, changes in the way the sub-network operates are being made.

The construction of the communications sub-network and its assessment by performance monitoring were, naturally enough, the major initial activities; although some early use was made of the network as a facility by the RAND corporation in conjunction with the Universities of California at Santa Barbara and Los Angeles, and by Stanford Research Institute. However, many other applications soon began to be developed and will continue for a number of years, particularly when the ILLIAC IV (which is a very powerful parallel-processing computer system) and the 10^{12} bit direct-access store being developed for use with it, are available on the network. To a large extent, the more advanced uses depend on the agreement of high-level procedures for exchanging information and controlling interactions through the network, and it is interesting to see how this is being achieved with such a large group of users, operating many different types of computer.

The Network Working Group (NWG) with members from each HOST site is responsible for technical coordination and has several sub-groups working on various protocols. These are discussed below. The vital dissemination of technical information throughout the growing network community is achieved by a three-level documentation scheme. The most formal papers are known as 'Documents' which are issued as a statement of network policy by the chairman of the NWG. A 'Request for Comments' (RFC) may be issued by any member of the NWG as a means of proposing new technical standards and promoting the exchange of ideas between members of the group. A guide to RFCs is published monthly by the MITRE Corporation of Boston. Finally, a complete collection of all Documents, RFCs, memoranda, etc. is deposited at the Network Information Center (NIC) operated by Stanford Research Institute (SRI), which publishes a comprehensive index and has a computer-based network information system accessible from remote terminals through the network. The excellence of these arrangements for information dissemination and document control plays an important role in the success of the project by coordinating research at the ARPA sites. These sites are given in Table 1.

High-Level Protocols

When the network was designed the interface between the IMPs and the HOSTs was defined by B.B. & N and procedures or protocols were worked out for controlling the flow of packets between IMPs (the IMP to IMP protocol) and the flow of messages between HOSTs (the HOST to HOST, or network control protocol). But the way in which the hetero-

ARPA site identifier	*ARPA network organization*
ABERDEEN	Aberdeen research and development center
AFETR	Air Force Eastern Test Range
ALOHA	Aloaha network, University of Hawaii
AMES	NASA Ames Research Center
ARPA	Advanced Research Projects Agency
BBN	Rolt Beranek and Newman
BELVOIR	USAMERDC, Fort Belvoir
BURR	Burroughs Corporation, Paoli
CASE	Case Western Reserve University
CCA	Computer Corporation of America
CMU	Carnegie-Mellon University
DCAO	Defense Communications Agency Operations
DOCB	Department of Commerce, Boulder
ETAC	USAF-ETAC
FNWC	Fleet Numerical Weather Center.
GWC	Air Force Global Weather Center
HARV	Harvard University
ILL	University of Illinois
ILLIAC	NASA Ames Research Center
LBL	Lawrence Berkley Laboratory
LLL	Lawrence Livermore Laboratory
LL	M.I.T. Lincoln Laboratory
MCCL	McLellan Air Force Base
MIT	M.I.T.
MITRE	Mitre Corporation
NBS	National Bureau of Standards
NYU	New York University
RADC	Rome Air Development Center
RAND	Rand Corporation
SAAC	Seismic Analysis Array Center
SDC	System Development Corporation
SU	Stanford University
TINK	Tinker Air Force Base
UCLA	University of California Los Angeles
UCSB	University of California Santa Barbara
UCSD	University of California San Diego
UCS	University of Southern California
UTAH	University of Utah

Table 1. Organizations in the ARPA network

genous HOST systems were to communicate could not, initially, be determined. However, under the Network Working Group a 'layered' approach to the specification of protocol has been adopted. The inner layers are the original HOST to IMP and HOST to HOST protocols, and several higher-level protocols have since been defined by the community of HOST sites. These form a hierarchy of protocols covering various kinds of interaction between HOSTs. These protocols are:

1. The initial Connection Protocol (ICP), providing a standard method for processes in different HOSTs to establish a connection.

2. The Telecommunication Network (TELNET) protocol, used to provide communication between a keyboard terminal and a terminal-serving HOST. This uses the Network Virtual Terminal to overcome terminal hardware differences.

The Uses of Message Switching in Communications Networks

3. The Data Transfer Protocol (DTP), specifying standard methods of formatting data for passage through the network, allowing it to be used to implement higher level protocols.

4. The File Transfer Protocol (FTP) defines standard methods for reading, writing and updating files stored at a remote HOST in an endeavour to shield users from the differences between filing systems at various sites.

5. The Data Reconfiguration Service (DRS) attempts to deal with the problem of reconciling the different input-output data formats that are used by various applications programs. It uses an interactive procedure between user processes and an interpreter called the Form Machine (analogous to filling out a form by hand).

6. The Mail Box Protocol (MBP) provides a service for passing messages between people, and is widely used to facilitate the work of protocol development by members of study groups.

7. The Graphics Protocol (GP) is being considered by a group thinking about the difficult area of specifying standard ways of handling graphical information.

It is clear that the development of computer usage in the U.S.A. is bound to profit markedly from the coordinated efforts of so many people in developing these advanced concepts. It is also clear that without the ARPA network project to act as a catalyst, the coordinated approach would have been difficult, if not impossible, to organize. As the project develops other people will undoubtedly wish to join and it is not easy to see, at this stage, how far-reaching it will eventually become.

References

[1]Freebody J.W., *Telegraphy,* Pitman, London (1958).
[2]Barton, R.W., *Telex,* Pitman, London (1958).
[3]Jackson, P.E., and Stubbs, C.D., 'A study of multi-access computer communications', *AFIPS Conference Proceedings Spring Joint Computer Conference,* 34, 491 (1969).
[4]Baran, P., 'On distributed communication networks', *IEEE Trans. Communication Systems, CS-12,* 1-9 (March 1969).
[5]Davies, D.W., 'Communication networks to serve rapid response computers', *Proc. IFIP Congress, Invited Papers,* 72 (1968).
[6]Silk, D.J., 'Routing doctrines and their implementation in message switching networks', *Proc. Inst. Elect. Engrs.,* 116, No. 10, 1631 (October 1969).
[7]Roberts, L.G., 'Computer network development to achieve resource sharing', *AFIPS Conference Proceedings Spring Joint Computer Conference,* 36, 543 (1970).
[8]Heart, F., 'The interface message processor for the ARPA computer network', *AFIPS Conference Proceedings Spring Joint Computer Conference,* 36, 551 (1970).

The Techniques and Applications of Packet Switching

Problem:

With digital channels proliferating and the use of computer terminals growing by leaps and bounds, a particularly important technology is the building of data transmission networks. Packet switching is one way to build a switched data network capable of handling the burstiness of interactive computer traffic. This report offers a general introduction to packet-switching networks.

Solution:

The CCITT definition of a packet is as follows: A group of binary digits including data and call control signals which is switched as a composite whole. The data, call control signals and possibly error control information are arranged in a specified format.

The associated CCITT definition of packet switching is: The transmission of data by means of addressed packets whereby a transmission channel is occupied for the duration of transmission of the packet only. The channel is then available for use by packets being transferred between different data terminal equipment. Note—The data may be formatted into a packet or divided and then formatted into a number of packets for transmission and multiplexing purposes.

These definitions apply to the networks for which the terms were originally used, such as the ARPA network, the Telenet network, the Datapac service in Canada, and the British Post Office EPSS (Experimental Packet-Switched Service), and they apply also to networks with conventional data concentrators or to hierarchical network protocols such as IBM's SNA (System Network Architecture), or to conventional message switching networks. A major difference between such networks is the geographical layout. A network such as ARPA or Telenet serves many computer centers with a mesh-structured line layout. A typical data concentrator network or message switching network has a tree structure, taking traffic to or from one computer center or possibly a few interlinked centers.

Future Developments in Telecommunications (2nd edition) by James Martin. Chapter 29, pp 539-557 © 1977 Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

MESSAGE SWITCHING

Packet switching is a form of store-and forward switching. Messages are stored at the switch node and then transmitted onwards to their destination. Store-and-forward switching has existed for decades in telegraphy where it is called message switching. There are, however, major differences between packet switching and conventional message-switching.

Whereas message switching is intended primarily for nonreal-time people-to-people traffic, packet switching is intended primarily for real-time machine-to-machine traffic, including terminal-to-computer connections, and is employed to build computer networks. These differences in purpose are such that there are major differences in operation between message-switching and packet switching networks. One important difference is in the speed of the network. A packet-switching network may be expected to deliver its packet in a fraction of a second, whereas a message-switching system typically delivers its message in a fraction of an hour. Each node passes the packet to

The Techniques and Applications of Packet Switching

the next node quickly like passing on a hot potato. Another important difference is that a message-switching system files a message for possible retrieval at some future time. A packet-switching system deletes the message from memory as soon as its correct receipt is acknowledged. Because a message-switching system files messages, usually at one location, it tends to use a centralized star-structured or tree-structured network. A packet-switching network usually has an amorphous structure with no particular location dominating the structure.

In many message-switching systems, long messages are sent as a single transmission. In packet-switching systems, long messages are chopped up into relatively small slices—1008 bits per packet on the U.S. Telenet system; 24 8-bit bytes on Canada's Datapac system. Because the packets are of limited size, they can be queued in the main memory of the switching nodes, and passed on rapidly from node to node. The original message has to be reassembled from the slices at its destination.

The locations connected in the Telenet network are linked with wideband lines operating at 50,000 and 56,000 bits per second. Each of these terminates in a small network computer. The network computer has two main functions. First, it acts as a link between the network and the data processing equipment that uses the network. Second, it carries out the switching operation, determining the route by which the data shall be sent and transmitting it.

The customer's computers that the network serves are called host computers. When one host computer sends data to another, it passes the data with a destination address to its local network computer. The network computer formats the data into one or more packets. Each packet contains the control information needed to transmit the data correctly. The packets are transmitted from one network computer to another until they reach their destination. The final network computer strips the transmission control information from the packets, assembles the data, and passes it to the requisite host computer.

A network computer receiving a packet places it in a queue to await attention. When it reaches the head of the queue, the computer examines its destination address, selects the next network computer on the route, and places the packet in an output queue for that destination. A packet-switched network is usually designed so that each network computer has a choice of routing. If the first-choice routing is poor because of equipment failure or congestion, it selects another routing. The packet thus zips through the network, finding the best way to go at each node of the network and avoiding congested or faulty portions of the network.

We might compare a telecommunications network with a railroad network. With circuit-switching there is an initial switch setting operation. It is like sending a vehicle down the track to set all of the switches into the desired position; the switches remain set, and the entire train travels to its destination. With packet switching, the cars of the train are each sent separately. When each car arrives at a switch, the decision is made where to send it next. If the network is lightly loaded, the cars will travel to their destination by a route close to the optimum. If the network is heavily loaded, they may bounce around or take lengthy or zig-zag paths, possibly arriving in a different sequence to that in which they departed.

A train with only a single car can head off into the network with no initial set-up operation. However, if the train has many cars, it should not start its journey until it is sure that there is enough space for all of the cars at the destination. As we shall see, on some networks an engine has to be sent to the destination and return with a go-ahead message before the train can set off.

PACKETS

The packets might be thought of as envelopes into which data are placed. The envelope contains the destination address and various control information. The transmission network computers should not interfere in any way with the data inside the envelopes. In fact the system should be designed with security safeguards so that network computers cannot pry into the contents of the envelopes.

Figure 1 shows the structure of the ARPANET, or Telenet, packet. It has a maximum length of 1008 bits. The text (data being sent) is preceded by a start-of-message indicator and a 64-bit header. It is followed by an end-of-message indicator and 24 error-detection bits. The header contains the destination address, the source address, the link number and packet number used to ensure that no packets are lost and that packets in error are transmitted correctly, a message number with an indication of whether there is more of the message following in another packet, and some special-purpose control bits.

PACKET CONTROL PROCEDURES

The transmission of the packets through the network requires three types of control:

1. Error control, to deal with any transmission errors that occur.

2. Routing control, to determine the routes over which the packets are transmitted.

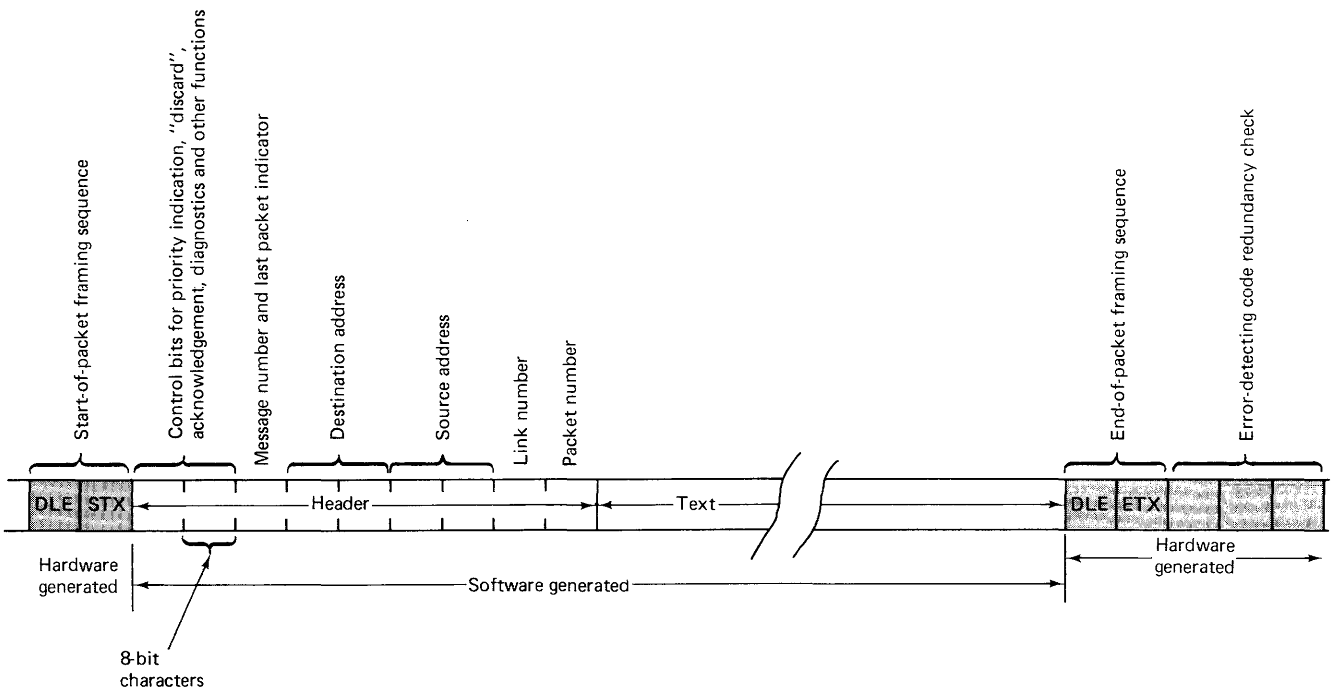


Figure 1. A typical packet format

3. Flow control, to avoid congestion in the network which could cause lockouts or traffic jams.

ERROR CONTROL

Error control procedures are applied to each point-to-point link. When a node receives a packet, it checks its accuracy using the error-detecting code bits. Error-detecting codes can be made very powerful so that the probability of a transmission error being undetected is very low. It is estimated that the probability of an error bit being undetected on a link in the ARPA or Telenet system is 10^{12}.

If a node finds the packet it receives to be correct, it transmits an acknowledgment message to the sender. If it finds the packet to be incorrect or garbled, it ignores it. Whenever a node transmits a packet, it retains it in storage until it receives the acknowledgment of correct transmission. If it does not receive such an acknowledgment within a specified time period (say 100 milliseconds), it automatically retransmits the packet. If repeated attempts to transmit the packet receive no acknowledgment, the sender tries to transmit by a different route.

In some packet-switching systems there exists this point-to-point error control between the switching nodes but there is no end-to-end error control between the destination machine and the originating machine. Consequently the malfunctioning of a switching node can, on rare occasions, result in the loss of a message.

ROUTING CONTROL

When a packet-switching computer receives a packet addressed to another location, it must determine which of the neighboring nodes of the network to send it to. The computer will have a programmed procedure for routing the packet. A variety of different routing strategies are possible:

1. Predetermined Routing—The route may be determined before the packet starts on its journey. The packet then carries routing information, which tells network computers where to send it. The determination of the route may be done by the originating location, or it may be done by a "master" station controlling the entire network.

Most of the traffic consists not of lone wandering packets but of "sessions" in which many packets pass back and forth between the same two locations. Typical sessions are dialogues between a terminal user and a computer in which many messages and responses are sent, data entry operations in which an operator keys in many transactions, file transfers in which many packets of data are sent, and monitoring operations in which a remote computer monitors or controls a process. With predetermined routing, the routing decision can be made once for the entire session. Any packet that is part of a session contains the same routing instructions. If a line or network computer fails during the session, a recovery procedure will be necessary to establish a new route.

The alternative to predetermined routing is that each network computer makes its own routing decision for each packet. ARPANET and most proposed packet-switching systems employ this nonpredetermined routing. With nonpredetermined routing, the network computer has more processing to do, but the packet envelope may be shorter because it contains a destination address not an entire route.

2. Calculated Routing—The address of the destination nodes in a network may be chosen in such a way that it is possible for any interim node to determine which way to send a packet by performing a simple calculation on its address. If a node has received information about a failure in that direction, it may calculate a second-best routing.

Calculated routing is simple but in general too inflexible and hence unlikely to be used in practice.

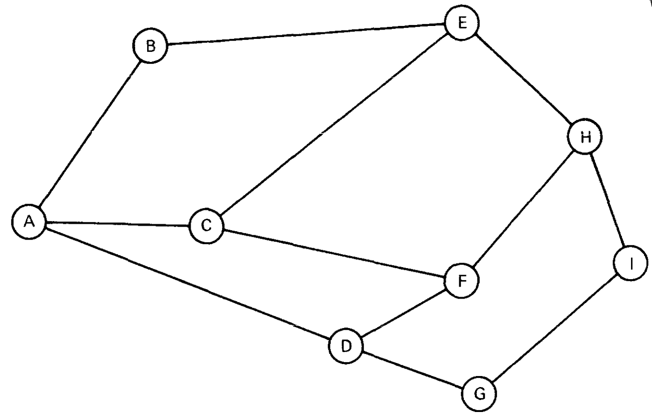
3. Static Directory Routing—With directory routing, each node has a table telling it where to send a packet of a given destination. Figure 2 shows a possible form of such a table. The table shown gives a first choice and second choice path. If the first choice path is blocked or inoperative, a node will use the second choice path. (There is no need for the table in the illustration to give a third choice because no node has 3 lines going from it.) As a packet travels through the network, each node does a fast table look-up and sends it on its way.

4. Dynamic Directory Routing—The previous method uses a fixed table. A more versatile method is to use a table which can be automatically changed as conditions of the network change.

There are several possible criteria that could be used in selecting the entries for a table such as that in Figure 3. They include:

- Choosing a route with the minimum number of nodes.
- Choosing routes which tend to spread the traffic to avoid uneven loading.
- Choosing a route giving a minimum delay under current network conditions.

The last of these conditions implies that the table will be constantly modified to reflect the current delays on the network caused by congestion or failure. Figure 3 shows figures proportional to the delays on the network at one time, and a routing table for node C that takes these delays in consideration. As part of the time delay at each node is caused by queuing, the delays and the optimum routing table will change as the traffic patterns and volumes change.

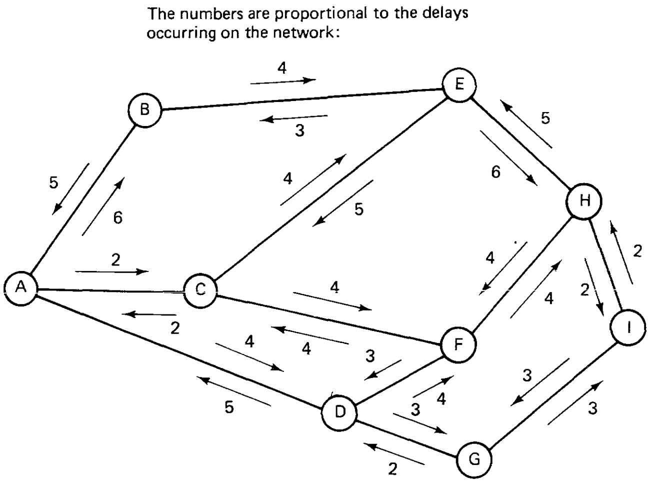


Packet destination	First choice node	Second choice node
A	A	F
B	A	E
D	A	F
E	E	A
F	F	A
G	F	A
H	E	F
I	E	F

Figure 2. A static routing table for node C in the above network, giving the node to which C should route a packet for a stated destination

The question now arises: how should a node be informed of what the network delays are? Several methods have been suggested. Paul Baran suggested that the delay on any route x→y is similar to the delay traveling in the opposite direction y→x. Consequently, a node could obtain approximate information about network delays by finding out how long it had taken each packet reaching it to travel from its originating node. The time of departure would be recorded in each packet so that each node could determine this transit time. The method is rather like asking travelers on a rush-hour traffic system. "What is it like where you came from?"

Another system is for each node to send a service message at intervals to each of its neighbors. The message will contain the time it was originated, and also that node's knowledge of whether the delays have changed. The recipients will record how long the message took to reach them, and if this is substantially different to the previously recorded delay they will



Packet destination	Next node
A	A
B	E
D	A
E	E
F	F
G	A
H	F
I	F

Figure 3. A dynamic routing table for node C intended to minimize transit delays under current network conditions

include this information in the next service message they send. Knowledge of changes in transit times will thus be disseminated throughout the network.

ADAPTIVE ROUTING

A scheme in which the routes selected vary with the conditions of the network is called adaptive routing. The ARPA and Telenet network use adaptive routing, each node sending a service message every half second. Most proposed public networks in other countries also use adaptive routing.

Adaptive routing sometimes results in oscillatory behavior, the routing pattern oscillating rapidly backwards and forwards under peak conditions. Minor changes in the routing algorithm can affect the routing behavior under heavy loading in ways that are difficult to predict without simulation of the network. Detailed study of routing algorithms is being made and can substantially affect the peak network throughput.

It is possible that packets in such a network could fail to reach their destination because of temporary equipment failure or a data error in the address. Such packets might be passed indefinitely from one node to another if something did not stop them. To prevent this occurrence, a count field is used in each packet, and the number of nodes that have relayed that packet is recorded in it. When the count exceeds a certain number, the packet is returned to its point of origin. This process protects the network from becoming clogged with roving, undeliverable messages.

FLOW CONTROL

Flow control is desirable to prevent too many rackets converging on certain parts of the network so that traffic jams occur. The control messages passed between nodes to control the packet routing play a part in avoiding traffic jams. However, if too many packets enter the network heading for a given destination, the routing control alone will not prevent a traffic jam. Traffic congestion can be harmful because packets bounce around from node to node occupying an excessive share of the transmission capacity. The network performance degenerates out of all proportion to the increased load, like the roads out of a large city on a Friday evening.

The best way to prevent congestion is to control the input to the network Control messages can warn all input nodes that congestion is beginning to build up. The most common cause of potential traffic jams is that one host computer suddenly sends a large volume of traffic to another. If the packets for this traffic follow each other at the speed of the input node, there may be a traffic jam on the route. The rate of input needs to be controlled rather than merely opening a sluice-gate wide.

The ARPA network controls such surges by permitting only one message at a time to be sent from an originating point to any one destination. Each message can consist of up to eight packets. When the destination node has completely received and assembled the message and delivered it to the machine for which it was intended, the destination node sends a control message back to the originating node saying that it is permitted to send another message. This is called a "Ready-for-next-message" (RFNM) signal. The RFNM is formatted similarly to the other packets and finds its way through the network using the same protocols as any other packet.

Figure 4 shows the passage of a message through a packet-switching system with protocols similar to the original ARPA network. The message goes from the host machine on the left of the diagrams to the host machine on the right. Three switching nodes, or IMPs, are shown. The first IMP is connected directly (i.e.,

The Techniques and Applications of Packet Switching

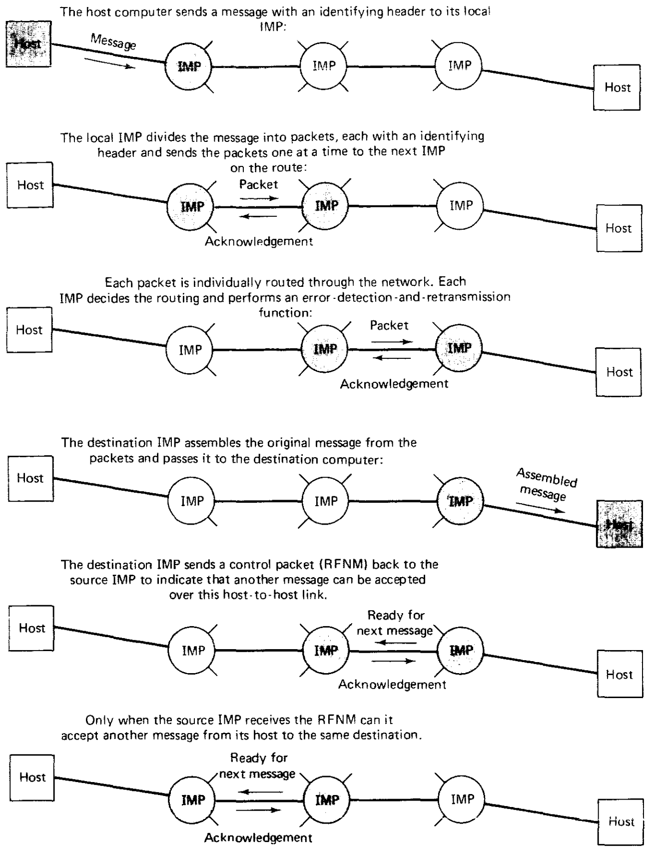


Figure 4. The flow of data on the ARPA network. A problem exists with the original ARPA control mechanism shown here, in that shortage of memory in the destination IMP can cause a "reassembly lockout" in which messages partially received from different sources cannot be completely received and a traffic jam builds up. An additional reservation mechanism is necessary to prevent such lockouts

not over a telepone company line) to the originating host machine. The host sends a message, which on the ARPA network can be up to 8063 bits, to its local IMP with a header saying where the message is to be sent. The IMP chops the message into slices, each slice becoming a packet with the format shown in Figure 1. The IMP determines which line to transmit on and sends the first packet. When it receives an acknowledgement from the next IMP saying that the packet was received correctly, it sends the next packet. The packets eventually arrive at the destination IMP, the third IMP in the figure. They may have come by different routes and hence could possibly arrive out of sequence. The destination IMP waits until it has received all the packets for this message: it assembles the message, removing the packet envelopes, adds a message header, and delivers the message to the destination host. The receiving IMP then sends a RFNM addressed to the originating IMP on the left. The RFNM bounces like any other packet to the IMP it is addressed to. When the lefthand IMP receives the RFNM correctly, it tells the host that it is ready to accept another message for transmission.

REASSEMBLY LOCKOUT

As the operators of the ARPA network discovered, the protocol shown in Figure 4 was not enough to prevent congestion problems. A serious form of jam could occur when long messages from different sources converged on one destination.

The destination IMP would receive and acknowledge the first packets of several messages. It has only a certain amount of memory available for the assembly of messages, and when it starts to receive a message it does not know how many packets will arrive for that message. It may begin the reception of more messages* than it can complete. It is then in serious trouble. It runs out of memory for reception of more packets but cannot deliver to the destination host the incomplete messages clogging its memory. The neighboring nodes continue to transmit packets to it, but it cannot accept them. After several attempts, the neighboring nodes try to send the packets to it by a different route. But the destination IMP is locked solid. The traffic piles up in neighboring areas like city streets after an accident in the rush hour.

This condition is called a reassembly lockout and was regarded as a bug in the protocol design of the ARPA network.

Reassembly lockout can be avoided in several ways. First, it does not occur if only single-packet messages are sent. The originating hosts could be made to slice up their own messages, so that this was possible. Second, the lockout could be avoided if the destination IMPs could pass incomplete messages to the receiving host. Both of these solutions were regarded as unsatisfactory because they complicate the protocols needed in the hosts. The intent of the design was that the network should appear transparent to the hosts, and that they should not have to be concerned with slicing up or reassembling messages.

A third solution is to give the nodes a backing store so that they can temporarily move incomplete messages out of their main memory when the lock-outs occur. This would increase the cost of the nodes substantially.

The solution adopted on the ARPA and similar networks was to prevent the originating nodes from sending a multipacket message until they were sure that enough space has been reserved in the destination node for its reception. To accomplish this, more control messages were needed in addition to those in Figure 4. Before sending a multipacket message, the IMP on the left sends a reservation message to the destination IMP telling it how much space to reserve. The destination IMP sends an acknowledgment if it has the space, and then the originating IMP sends the packets. The overhead associated with the reservation process is not as severe as it may sound, because most of the messages sent are small enough to fit in a single packet.

It will be observed that the control of a mesh structured network is substantially more complex than that of the more conventional tree- or star-structured networks.

TWO TYPES OF PACKET SWITCHING

Because of the problems associated with message reassembly, two types of packet switching have evolved. The first handles multipacket messages and so has to have protocols that permit error-free message assembly without causing traffic jams. The second handles only single-packet messages, and hence avoids complex protocols that increase the network overhead. Canadian common carriers coined the term datagram to relate to the second kind of service. In a datagram service, users can send messages up to but not exceeding the maximum capacity of one packet. This permits a network to be built with simple control procedures and switches, low overhead, and fast transit times.

A datagram network is of value for many applications, including the vast future needs of electronic fund transfer. It could be designed to operate with very inexpensive terminals. Other applications need to send longer messages—including the vast future needs of facsimile mail services.

NETWORK TRANSPARENCY

It is the intention of the designers of many packet-switching networks to make the communication techniques as unobtrusive as possible to the users. The network operation should be independent of the nature of the computing operations that employ the network. Many new types of computers and new types of computers and new types of operation could then employ it. The network should connect two computer processes, perhaps thousands of miles apart, as though they were directly interconnected via a precisely defined interface.

This illusion of direct interconnection is referred to as network transparency. To make the network appear transparent, the transmission must be fast, and the software must hide the complexity of its operations from the process which uses it.

On the ARPANET and Telenet systems there are three layers of communication protocol illustrated in Figure 5. On the outside of this diagram are two computer processes communicating by sending streams of bits to one another. In the innermost

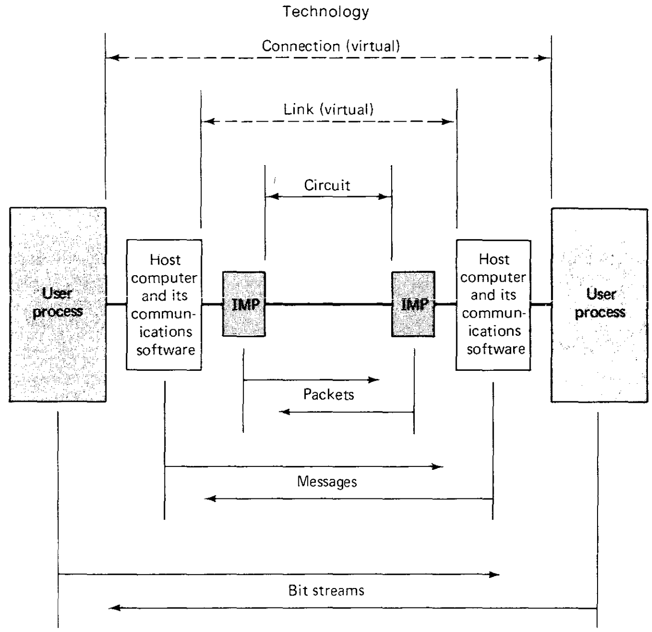


Figure 5. Three layers of protocol on the ARPA and Telenet networks

layer, the IMPs are transmitting to one another. In between the IMP network and the user process are protocols which reside in software and hardware in the host computers.

The connection between the IMPs is a physical circuit over which packets are transmitted. The connection between hosts is referred to as a "link," and conveys messages. The link is referred to as a "virtual" path, because in reality there is no physical path between hosts; it only appears that way. The connection between user processes, passing bit streams, is likewise referred to as a "virtual" path.

Three protocols are needed to permit communication between the user processes:

1. The IMP-to-IMP protocol which permits the packets to be routed between IMPs, free of errors.

2. The HOST-to-IMP protocol which permits the HOST and IMP to pass messages to one another. This requires appropriate hardware and software in the host.

3. The HOST-to-HOST protocol which provides rules permitting the user process to talk to one another.

TWO SEPARATE FUNCTIONS

Two separate functions must be carried out by nodes of a packet-switching network: communication with the users, and the switching of the packets. In the ARPA network, both functions are carried out by the same machines—the IMPS or TIPS. Other proposals,

The Techniques and Applications of Packet Switching

including proposals for public switched networks in several countries, have employed separate machines for the two functions.

If the functions are separated, the switch can be a relatively small computer. The steps to be executed in switching one packet are relatively few, so a high data throughput is possible with a fast machine. It has been estimated that more than a million bits per second could be switched using today's technology, and so the switch might be geared to PCM transmission channels of 1.544 or 2.048 million bits per second. Unlike a message-switching computer, the packet-switching computer does not have any accounting or filing functions to perform. It needs no peripheral storage. It merely receives packets, queues them, determines their routing, and retransmits them. During idle moments, it will send diagnostic packets and update its routing table.

The subsystem that interfaces with the user has somewhat more complex functions than the switch. They may include:

1. Creating the Packets—Data from a terminal or host is placed into one or more envelopes and control information needed for transmission is written.

2. Reassembly of Data—After transmission, data sent in more than one packet is reassembled. For some uses the entire message may be assembled prior to delivery: for others it may be better to deliver the data a block at a time as it arrives, provided that the blocks are in sequence.

3. Host-Network Protocol—The interface computer will observe a protocol for communicating with the host computers to ensure that the interchange functions correctly and that no data can be lost. Different protocols may be needed for different types of host.

4. Terminal Protocol—A protocol for communicating with user terminals will be observed. Some of the user terminals may be far away, connected to the interface computer by links incorporating multiplexer, concentrators, polling, public network dial-up or other procedures.

5. Special-Function Protocols—Special protocols may be used for functions such as the transfer of or the use of graphics, "conference calls" in which more than one user participates in one dialogue, mail-box services in which the network passes messages between users and holds then until the users see them, facsimile transmissions, transmission of exceptionally high security, and others.

6. Session Control—Many transmissions, as discussed earlier, are part of a "session" in which multiple messages are sent, as in a human telephone conversation. In this scheme, the interface computer may control the session. It will store an envelope header for the session so that it does not have to be recreated for each stage. It may use a session-oriented protocol with the host computer or terminal. It may allocate a high priority to the packets of certain sessions, possibly after establishing whether this priority is sustainable with the current network load.

TWO-LEVEL NETWORK

Various authorities, including D.W. Davies and his co-workers at the British National Physical Laboratory and the designers of the PCI (Packet Communications Incorporated) network, the first of the U.S. value-added common carriers to receive FCC approval, have advocated that a packet-switched network should be constructed having two levels. The lower level would be the local area network, corresponding broadly to the central office and local loops of the telephone network. This would carry out the function of interfacing the host computers and terminals. The higher level would be the long-distance packet-transmission and switching network, corresponding broadly to the trunking network of the telephone system.

A wide variety of different terminals can be attached to the lower level interface computer. Indeed, one of the chief advantages of this type of network is that entirely different types of terminals can intercommunicate. They can have widely different speeds and use different codes. They can be synchronous or start-stop. They can use polling, contention, or be alone on a line. They can have different types of error control. If desired, they can be connected via multiplexers or concentrators. Above all, they can be inexpensive, for most of the common terminal features like buffering and elaborate line control are not really necessary. The interface computer maintains a list of the characteristics of all the terminals attached to it, their control mechanisms, speeds, and transmission codes. If the code differs from the network transmission code the interface computer converts it as the packet is being assembled. When the packet is received by the destination line-control computer, it is converted, if necessary, to the code of the receiving terminal. It is estimated that one interface computer could handle more than a thousand terminals in this way.

An additional function of the interface computer is to collect the information necessary for logging and billing subscribers.

The high-level network can use links of different speeds for moving the packets. As satellite technology develops, the high-level network of such a system would probably incorporate satellite transmission. Where the high-level network is installed by the same authority that operates the telephone network, the links can be the PCM links that carry telephone voice traffic. Packet transmission and voice transmission can thus share facilities.

THE FUTURE OF PACKET SWITCHING

Many countries have announced an intention to build or experiment with a packet-switching network. There will continue to be arguments about the relative merits of packet-switching and fast circuit-switching networks.

It seems likely that the existing packet-switching networks will steadily grow, acquiring more traffic and more nodes. If they become large and ubiquitous, economies of scale may make some form of switched data network replace most of the private leased-line networks that corporations and government departments use today with techniques such as concentrators and polling.

There are several future directions in which packet-switching networks will probably evolve if their traffic grows sufficiently:

1. High-speed PCM links may become the links used by packet-switching networks. Transmission rates of millions of bits per second will permit very fast response-time systems to be built.

2. To fill such high speed links, the networks will have to attract a high traffic volume (as will Datran to fill its 44 mbps trunks). Much of this traffic may come from relatively new uses of data links such as electronic mail, electronic fund transfer, and other forms of message delivery.

3. As networks grow very large, it is economical for them to become multilevel networks with a hierarchy of switching offices. Just as the telephone network has five classes of office, so data networks may acquire two, three, and eventually more levels.

4. Several classes of traffic may be handled, differentiating perhaps between datagrams and long messages.

5. Several classes of priority may be handled, including perhaps immediate delivery, 2-second delivery for interactive computing, delivery in minutes, and overnight delivery.

6. Some message traffic may be filed as on a message-switching system. Messages intended to be read on visual display units or spoken over the telephone may be filed until the recipient requests them. Distributed storage rather than centralized storage may be used, especially for bulky data, depending upon the relative costs of storage and transmission. A hybrid between message-switching and packet-switching may thus emerge.

7. Fast-connect circuit switching has advantages over packet-switching for some types of traffic. The nodes of a large data network may be designed to select whether a circuit-switched or packet-switched path is used. A hybrid between circuit-switching and packet-switching may emerge.

8. The user interface computer may become separate from the switching computer and have an entirely different set of functions. It may be designed to convert the transmission of all terminals to a standard format, code, and protocol so that completely incompatible machines can be interconnected. A telex machine using Baudot code can transmit to a visual display terminal using the CCITT Alphabet No. 5.

9. Elaborate security procedures may become used.

10. The user interface devices may be designed to receive from and transmit to conventional facsimile machines or other analog devices. I.E. codecs may be built into the nodes.

11. The interface machines may be designed to compress messages before transmission to increase the transmission efficiency. This is valuable with data, but especially valuable with facsimile messages.

12. The interface machines may be designed to handle packet radio terminals or controllers. Portable data terminals may be linked to the system.

13. One of the most cost effective data transmission facilities will be the satellite. Packets will probably be sent via satellite. To use future satellites in an optimal (broad band) fashion will substantially change the topology and protocols of packet-switching networks.

14. Economies of scale and flexibility may require that telephone or continuous-channel traffic and burst traffic be intermixed. Networks, especially satellite networks are capable of handling both continuous-channel and packet-switched traffic may employ both.

15. An interlinking of separate national networks will occur. Satellites will interconnect nodes in many countries, giving users of packet-switching the capability to contact computers around the world and send messages worldwide.

16. It is extremely important that there should be internationally agreed standards for the interface to the networks user machines employ.

17. When vast numbers of computers are available on the networks, directory machines will be very important for enabling users to find the facilities they need.□

Operating Systems for Computer Networks

Problem:

Large data communications networks tend to become highly complex even when all of the computers and terminals are provided by one manufacturer. Complexities are multiplied when different manufacturers' mainframes, front-ends and terminals are intermixed. In fact, there are severe limitations on what can and cannot be done in a commercially available heterogeneous network. The operative word here is "commercially available."

The problems of sharing resources within a network of heterogeneous components has been addressed and solved with some degree of satisfaction within the Arpanet. Arpanet, however, was developed within an academic environment and is operated and maintained with government funds. Consequently it was not, and is not, subject to the budgetary limitations which commercial manufacturers must face. It has, and will continue to provide valuable information for manufacturers in the development of their own network architectures and software.

This report describes a network operating system developed for Arpanet. It is presented as representative of the types of network operating which may be developed by major computer manufacturers within the next few years. It is not a commercial product or even a prototype, however, and that is why it is located in the Concepts section of our service.

Solution:

The last decade has seen the rapid evolution of computer communication networks from research curiosities to operational utilities. The Arpanet,[1] for example, currently supports communication among more than 100 computer systems and is used daily by hundreds of users. Commercial networks, such as Telenet[2] in the United States and Datapac[3] in Canada, have very bright futures. One attraction of these networks is their ability to provide access to a wide variety of resources distributed among the connected computers.

A typical network (Figure 1) includes a communication subsystem, to which a collection of computers, called hosts, are connected. This subsystem usually consists of communication processors interconnected by communication links, such as coaxial cable, telephone lines, or satellite channels. Com-

"Operating Systems for Computer Networks" by Harry C. Forsdick, Richard E. Schantz, and Robert H. Thomas of Bolt, Beranek and Newman, Inc. © 1978 IEEE. Reprinted, with permission, from COMPUTER, January 1978.

Operating Systems for Computer Networks

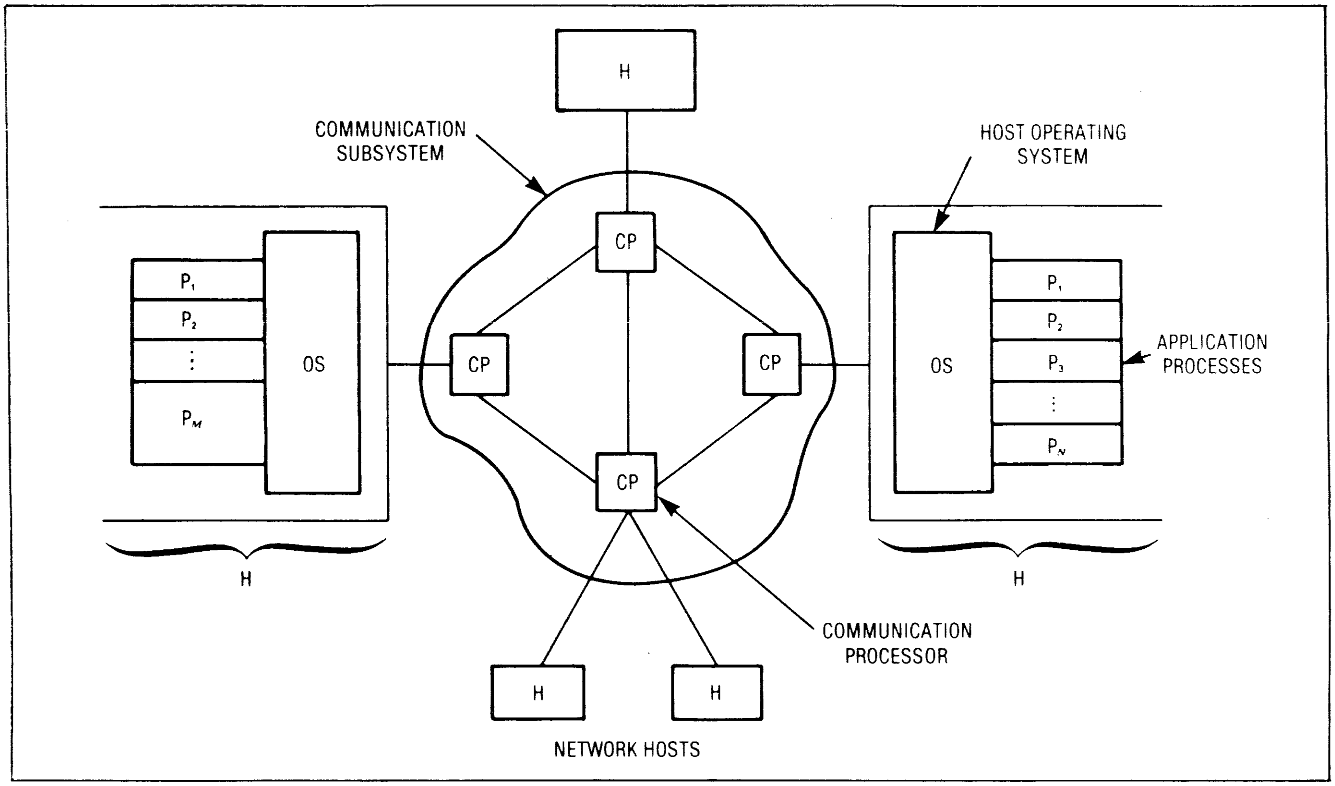


Figure 1. In a network, a communication subsystem interconnects host computers, host operating systems, and application processes with one another.

munication processors have two functions: they cooperate to support communication between the hosts, and they provide the interface through which the host computers are connected to the communication facility. Each host includes an operating system that supports one or more application processes. The principal purpose of the communication network is to permit access by a user, or by a process acting on his behalf, to resources of the other machines.

Networks such as the Arpanet have stimulated the development of protocols that support terminal access to remote hosts[4] and file transfer between hosts.[5] However, at present the amount of resource sharing that occurs on computer networks falls far short of that which is possible—for several reasons:

- To use the network resources effectively, the user must know not only the access mechanisms of the network, but also the operating systems of the hosts whose resources he wishes to use.
- Resources of the various hosts are generally not compatible with each other.
- Information about the available resources and how to use them is difficult to obtain.
- Accounting and billing for resource utilization are generally maintained by each host, so that a user must establish a separate account with each organization whose host he plans to use.

The fact is that a network, even with terminal access and file transfer protocol implementations, is not an integrated operating system.

The concept of a network operating system represents a promising approach for overcoming these difficulties and thus realizing the full potential of computer communication networks. A network operating system (NOS) is a collection of software and associated protocols that allow a set of autonomous computers, which are interconnected by a computer network, to be used together in a convenient and cost-effective manner. An NOS would help users and their programs exploit the resources distributed among network hosts in much the same way that a single-host operating system provides its users with controlled access to local resources. It would tend to make the underlying network and the boundaries between host systems transparent to users. It would remove many of the logical distinctions between resources that are local and those that are remote. Such a system would provide an environment within which uniform accounting throughout the network would be feasible. In addition, an NOS environment could support a wide range of easily accessible information services, which might range from current information on the status of constituent systems to information about the hardware and software resources available. Because of the potential for redundancy and the inherent autonomy of the constituent host systems, an NOS can, in principle, provide service of far higher availability than can a single-host operating system. However, the potential for decreased availability also exists, if the NOS depends on system components that are subject to failure. Thus, an NOS must be carefully designed to produce a system with increased, rather than decreased, availability.

Several projects have been undertaken to study, design, and build such systems. The technical problems uncovered in these efforts are illustrated in two projects with which we have been involved: the Resource Sharing Executive Project, supported by the Defense Advanced Research Projects Agency (DARPA), and the National Software Works Project, supported jointly by DARPA and the Air Force Rome Air Development Center.

In these projects, intensive study was devoted to the first two of the four previously mentioned factors: facilitating access to network resources and resolving incompatibilities among different host systems. Although the other two factors are equally important, a thorough discussion of them is beyond the scope of this report. Although both systems provide an environment capable of supporting effective information services, these services have not been extensively developed. With regard to the fourth factor, the accounting problem, we feel that the unresolved issues are largely political and administrative rather than technical.

THE RSEXEC SYSTEM

The Resource Sharing Executive System[6] was developed as an experimental vehicle to explore a wide variety of NOS issues, ranging from the types of features that would be useful to network users, to system structures and mechanisms for implementing those features, to strategies for ensuring reliable, fail-soft performance.

The initial emphasis of the RSEXEC System was to couple the operation of the many PDP-10 hosts connected to the Arpanet which use the Tenex operating system.[7] The goal was to provide access to the combined resources of these hosts much as the Tenex operating system and command language interpreter provide access to a single Tenex host (Figure 2). The RSEXEC includes both a command language interpreter for users and an execution environment for user programs.

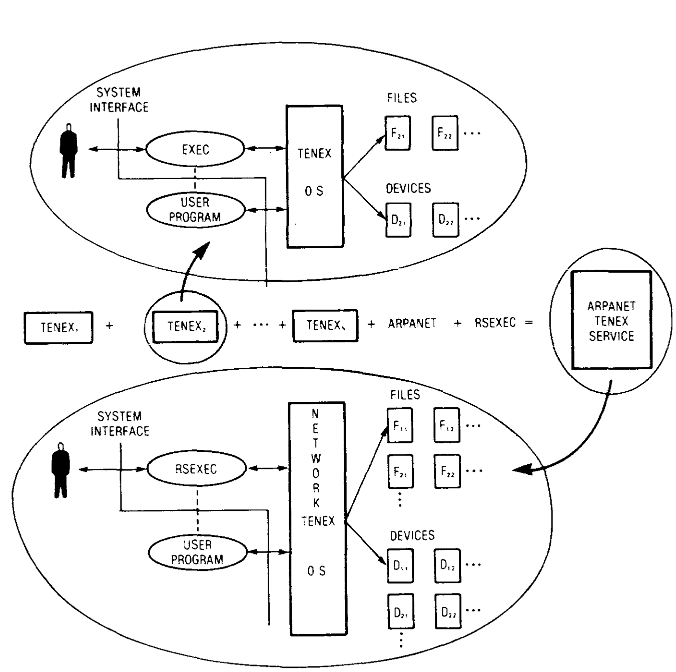


Figure 2. The RSEXEC couples the operation of Tenex hosts on the Arpanet. One Tenex system includes the components in the upper part of the diagram; when combined with other systems, the network, and RSEXEC, it becomes effectively a much larger system, shown beneath.

The system has three principal components: (a) the RSEXEC program, which is an interface between each active user and the system resources; (b) the RSEXEC server (RSSER) program, which runs on each of the hosts and makes the resources of that host accessible to remote users; and (c) the RSEXEC/RSSER protocol, the set of conventions governing the interactions between the RSEXEC and RSSER programs that is necessary to support the various system features.

Although RSEXEC was designed primarily for a homogeneous system composed of similar Tenex hosts, dissimilar host systems (Multics, IBM 360/TSO, PDP-10/ITS, Arpanet TIPs) have been partially integrated into it. Currently over 40 hosts are part of the system. To integrate a new type of host into the system, RSEXEC and RSSER programs are prepared for it. The protocol, which is independent of host type, specifies the functional requirements of the RSSER program. For a given host type the RSEXEC program is intended to provide access for local users and programs to the

Operating Systems for Computer Networks

combined resources of the RSEXEC hosts in the same style that the local operating system provides access to local resources. Although the form that user interactions take could be expected to vary from one host type to another, the functionality supported by RSEXEC programs for different host types is similar.

RSEXEC features. RSEXEC has a distributed file system that supports access to files that reside with the constituent hosts. The file system satisfies two principal goals: to make all files regardless of location uniformly accessible, and to provide convenient access to frequently used files. Uniform file access is achieved through a file-naming syntax applied to all hosts that extends the syntax within each single host by adding a field for network location. From the operating system viewpoint, a file name implies a series of steps through the file system hierarchy to reach the file, and is therefore called a pathname. Convenient file access is achieved through the use of partially qualified pathnames (which need not include network host location) to specify frequently used files. RSEXEC interprets these partial pathnames with the help of a private working directory, which it maintains for the user. The working directory, in general, spans host boundaries in the sense that it catalogs files from several hosts. Thus, knowledge of a file's location is not essential to make use of it. The host and its local operating system, acting alone, would normally reject such a partial pathname as an error; RSEXEC translates it into the full pathname for the host storing the file.

A third goal of the RSEXEC file system was to support multiple copies of critical files, so as to ensure that they would be accessible even when one or more of the hosts storing them had failed. The RSEXEC does help a user maintain multiple file copies by, for example, attempting to update all copies of the file when the user updates one copy. However, if a host in which a copy is stored is inaccessible when the attempt is made, the RSEXEC simply informs the user and takes no further responsibility for updating that particular file copy. To that extent it falls short of the third design goal.

Implementation of the file system features are largely the responsibility of the RSEXEC program. When a file is requested, the RSEXEC program's first step is to determine where the file resides, by interpreting the file name. File location is explicit in a full pathname. Partial pathnames are interpreted in the context of the current RSEXEC working directory, which is maintained locally for each user. A local copy ensures faster response by allowing partial pathnames to be interpreted without incurring the delay and other overhead of querying one or more remote hosts. Since the working directory spans several hosts, it is, in effect, a composite directory whose components are individual directories at these several hosts.

RSEXEC maintains a profile of each user that ordinarily has access to RSEXEC resources. Each profile includes a list of the component directories the user customarily accesses together with securely encrypted passwords to gain access to these directories. When the user begins an RSEXEC session, the RSEXEC program builds his working directory for that session from the profile. As he continues, the working directory expands and contracts, acquiring and discarding catalog information as necessary, and going out of existence at the end of the session, to be reconstituted when needed at the beginning of the next session. For component directories maintained by remote hosts, this catalog information is obtained from RSSER programs; for local component directories, it is obtained directly from the local operating system. For each file, the composite working directory includes the site at which the file is stored, along with other file descriptor information. After the RSEXEC program determines the location of a requested file, it initiates file operations at the appropriate host.

The features of the RSEXEC file system are supported at both the command language level and the executing program level. The RSEXEC program interprets commands typed by the user for manipulating files, while operations initiated by programs are performed in the context of the distributed file system.

An example of the former is a "RUN" command, with which a user executes a program within the RSEXEC environment. The program is retrieved from the RSEXEC file system and thus may reside on any participating host. The current implementation of RUN requires that the program execute on the local (RSEXEC program) host, where the user is working. (The underlying system structure will support remote program execution—that is, the RSEXEC/RSSER protocol includes commands for placing programs into execution under the control of an RSSER program at a remote host. However, the RUN command currently does not use that protocol feature.) When a remote file is the parameter of the RUN command, the RSEXEC program retrieves a copy of the file for its own direct control.

File operations initiated by executing programs use a technique called "encapsulation."[8] An application program executing on a stand-alone computer system issues many calls to its operating system to have tasks such as file operations performed for it. For the program to execute successfully in a multi-host system, these systems calls must be interpreted in the context of the larger set of multi-host resources. A section of the RSEXEC program acts to encapsulate application programs so that calls are interpreted from the RSEXEC viewpoint of distributed resources. RSEXEC intercepts certain file operations and interprets them in the context of the network file system as previously described. Operations referring to local files are then passed directly to the local operating system. Remote operations are forwarded to the appropriate remote RSSER server program. An important benefit of this encapsulation technique is that programs written for a single host environment need not be rewritten to operate within the network environment. Thus the value of existing software, such as text editors and language processors, is significantly enhanced by being able to work with remote as well as local data files.

RSEXEC includes commands with which users can obtain status information, such as current loads of constituent host systems—for example, the number of active users, load factors, etc. In addition, a user can link his terminal directly to that of another user, who may be logged into either the local system or a remote system, in order to engage in an on-line dialog. Users initiate and terminate remote links and local links in the same way. However, remote linking requires cooperation with a remote RSSER program, whereas local linking is directly supported by the local operating system.

Augmenting small host capabilities. One of the ideas investigated in the context of the RSEXEC system was to expand the usefulness of small machines by letting them obtain services from larger ones.[9] This has been particularly beneficial to some users who access the Arpanet and its hosts via TIPs.[10] The TIP is a small host system designed to provide terminal access to the network for users who are remote from the hosts they normally use. There are currently more than 20 TIPs in the Arpanet. The TIP itself offers no computational services and no user services beyond the ability to connect a terminal to a remote host through the network. Thus, the TIP does not directly provide status information or interact with the user.

However, the TIP does have a command that obtains access to an RSEXEC, which can provide these services. Several hosts on the network offer this service; they form a pool, to which the TIP broadcasts requests, selecting the one that responds first. This simple selection mechanism tends to distribute the service load among the hosts in the pool, and ensures that the service is available as long as at least one of the hosts is accessible.

An important application of this approach of using large-host resources to support small-host functions was the implementation of access control and accounting for Arpanet TIPs. Because TIPs have no long-term storage, they are incapable of authenticating their users or maintaining long-term usage information. Therefore, when an Arpanet user activated a TIP terminal port, perhaps by dialing it over a telephone line, the TIP would automatically connect to an RSEXEC, which would attempt to authenticate the user's identity. If the user failed to log in successfully, further use of the TIP would be denied—that is, the TIP would break its connection to the user. After logging in, the user would be permitted to use the TIP in the normal way. As part of the log-in procedure, the RSEXEC would pass the user's unique network identification to the TIP for accounting purposes, and the TIP would account for the user's connect time and network message traffic. Periodically, the TIP would send the usage data and the user's identification to an RSEXEC host for storage and subsequent processing and billing. (This access control and accounting procedure, though it was operational at one time, has been disabled for a number of non-technical reasons. Its mechanism, however, illustrates the potential of the approach.)

THE NATIONAL SOFTWARE WORKS

The National Software Works[11] is a network operating system to support software development. Designed and implemented by personnel from Bolt Beranek and Newman, Inc., Massachusetts Computer Associates, MIT, SRI International, and UCLA, NSW provides managers of programming projects access to a collection of management tools for monitoring and controlling project activities and gives programmers uniform access to a wide variety of software production aids. These include conventional tools such as text editors, simulators, compilers, interactive debuggers, emulators, and test data generators, as well as experimental aids being developed as part of various research projects, such as program verification systems and systems to support program development methods.

Although tools such as these, which span the software development process from design through implementation and checkout, have been available on a variety of computers for many years, they are seldom applied together in an effective way to support software implementation projects. One reason is that existing tools have been implemented for different computer systems, and programmers typically do not have access to the range of machines that house them. Furthermore, even if they did, programmers would have to master a variety of

Operating Systems for Computer Networks

different host operating systems and command languages, and deal individually with basic interhost incompatibilities to use the tools. The NSW system addresses both of these problems. It supports access to a wide variety of tool-bearing hosts and provides a simple, uniform means to invoke a tool regardless of the host that provides the tool.

Scenario of NSW application. The initial release of the NSW system included as tool-bearing hosts several PDP-10's operating under Tenex, some Honeywell 6000 computers operating under Multics,[12] and one IBM 360/91 operating under OS. Part of the user community is a group developing software for the AN/UYK-20 computer, a small machine that does not itself support a wide range of software development aids and is generally configured for production use rather than program development. This group's use of the NSW in the edit-compile-debug cycle illustrates the system capabilities.

Programmers prepare and modify their routines with interactive text editors that run on Tenex and Multics. Then they compile their source programs using language processors that run on the 360/91, and form executable AN/UYK-20 object modules from the compiled output with loaders that can run either on the 360/91 or on Tenex. Finally they debug the AN/UYK-20 programs interactively within NSW using a debugger tool that runs on Tenex. This interactive debugger is, in fact, a multicomputer tool. Part of it runs on Tenex and part of it runs on a microprogrammable computer, the MLP-900, which is connected as a peripheral device to one of the Tenex hosts and is programmed to emulate the AN/UYK-20. After making several edit-compile-debug cycles and producing a debugged AN/UYK-20 program, the user can "export" his modules from the NSW for final checkout and operation on a real AN/UYK-20 machine.

To apply any tool, a user simply specifies its name—such as CMS2M, an AN/UYK-20 compiler—and the NSW system starts the tool and connects the user to it. The user need not know which tool-bearing host supports the tool nor the command language of that particular host. He need only learn the NSW command language. To support tool execution the NSW implements a distributed file system. All NSW tools utilize this single system for their file references. When a tool calls for a data file (for example, a source program for a compiler), the NSW system ensures that the access refers to the correct file, independent of its actual location. If the addressed file is stored on another host, the file is automatically transported to the requesting host. In addition, NSW may transform certain aspects of the file to compensate for incompatibilities between the hosts or their file systems. Thus, the user and the tools he employs need not concern themselves with the location of data files, the details of the file systems on the various constituent hosts, the movement of files between hosts, or the data translations that may be required to make a file created on one host usable on another host. However, the file translations currently included in NSW are limited to those that support the existing set of NSW tools; NSW does not address the general problem of data translation, which is a difficult one.

It should be emphasized that the primary objective of the NSW project is not to develop tools, but rather to develop a system framework for tools—a framework to support uniform access to and integrated use of a diverse collection of tools, some already existing and some to be developed, all of them programs that run on one or more of a wide range of host processors under a variety of different operating systems.

NSW system structure. A useful view of the principal components of the NSW system is as processes that cooperate to provide NSW services. These components include front end (FE), works manager (WM), tool-bearing host foreman (FM), and file package (FLPKG) processes (Figure 3).

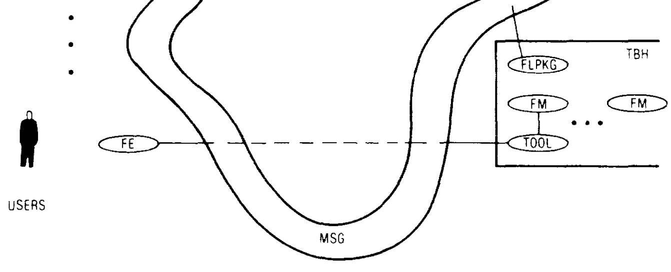


Figure 3. NSW components include front end (FE), works manager (WM), foreman (FM), and file package (FLPKG) processes. NSW tools run on tool-bearing host (TBH) computers under control of FM. The MSG interprocess communication facility supports communication among the distributed system components.

Operating Systems for Computer Networks

Each active user has a dedicated front-end process which is his interface to the NSW system. The principal function of the FE is to interpret the NSW command language and make requests of other components as necessary to satisfy user commands.

The works manager is the resource allocation and access control module for the NSW system. All requests for NSW resources, such as tools for files, must be authorized by the WM. To perform its task, the WM maintains data bases such as an NSW file system catalog, tool descriptor information, and user password and account information. These data bases and the WM software reside in one of the hosts included in the NSW system—possibly but not necessarily one of the tool-bearing hosts.

Interactions between WM processes and other system components occur as required, for each transaction involved. That is, the system dynamically allocates and deallocates WM processes as necessary to support a user session, rather than dedicating a single WM process to each active user for the duration of his session. For example, when a user initiates a command that requires access to an NSW resource, a WM process is allocated to handle requests related to that command. Upon completion of the command, the WM process is deallocated—that is, either returned to a pool of free WM processes or effectively destroyed. In the latter case, it is unavailable for reuse except by repeating the start-up procedure. Continuity across such instances of WM service is achieved by sharing a dynamic data base.

The tool-bearing host foreman[13] is the tool's interface to the NSW. When a user requests the start of a tool, an FM process on the appropriate TBH is allocated for the duration of the tool session. The FM process provides an execution environment for the tool and controls its operation. This execution environment differs somewhat from the standard environment of the local operating system. For example, a "file open" operation must be carried out in the context of the entire NSW, taking into account all of the network resources, not just those at the tool's host. The FM process responds to the file opening by calling on a WM process to complete the file reference. The WM process consults the NSW file catalog to verify the existence of the file specified by the FM, and to verify that the user and tool are authorized to use the file.

Next, the WM acts to ensure that the file can be physically accessed by the FM/tool. This may require movement of the file to the FM host and possible translation of the file data to a form usable by the tool. The WM process arranges for FLPKG[14] processes at the file source and destination hosts to cooperate to accomplish this movement and translation. Like WM processes, FLPKG processes are allocated on a transaction-oriented basis. When the file movement is completed, the WM process is notified by one of the FLPKG processes and then both FLPKG processes are deallocated. The WM process then informs the FM process that the file reference can be completed and is itself deallocated. Finally, the FM uses information provided by the WM—for example, the name of a local temporary copy of the file, accessible to the tool—to complete the tool file reference.

Communication between the various FM system processes, as well as the allocation and deallocation of those processes, is the responsibility of a component called MSG.[15] Two addressing modes, generic and specific, are supported by MSG. Generic addressing is the means by which a process initiates a transaction with another previously unrelated process. It is used when any process of a given type is acceptable. For example, to initiate the file operation in the AN/UYK-20 scenario, the FM process sends a generically addressed message to a WM process. Similarly, the FLPKG processes on the file source and destination hosts are activated by generically addressed messages. MSG allocates a process of the appropriate class, such as WM or FLPKG, to receive such messages.

Specific addressing, the second mode, is used when the sending process must communicate with a particular process. In the preceding example, the replies by the FLPKG process to the initiating WM and by that WM to the initiating FM are specifically addressed messages. When a process executes a receive operation, it declares whether a generically or specifically addressed message is requested.

The FM component provides two different tool interfaces to NSW. One of these is an encapsulation interface, which permits software packages that were developed to run under a particular TBH operating system to be installed as NSW tools. As in the encapsulation used in the RSEXEC system, the FM intercepts certain requests made by the tool of the local tool-bearing host operating system and transforms them into equivalent NSW operations. Ideally, the encapsulation interface would allow existing software to be installed into NSW without modification. In practice, the success of encapsulation for a given TBH depends upon the functional commonality between the TBH operating system and NSW, and the degree to which various tools exercise TBH functions not naturally mapped into NSW functions. Encapsulation has been fairly successful for the Tenex hosts; however, the integration of a new Tenex tool sometimes requires minor enhancements to the FM encapsulation interface.

The second tool interface provided by the FM is for tools developed specifically for operation within the NSW environment. This direct interface permits tools to explicitly call NSW functions. Generally the FM must interact with other NSW processes to satisfy these tool calls. The design intent is for the direct tool interface provided at different tool bearing hosts to be functionally equivalent. The manner in which tools invoke NSW system operations, of course, may vary from one host to another.

The NSW supports a wide variety of tool-bearing host types. As mentioned previously, the system includes three different types at present. For a host to be integrated into the NSW, it must have its own MSG, FM, and FLPKG components.

NSW file system. The NSW file system is built upon the file systems of the constituent hosts, which are used simply as a storage pool for NSW files. For each NSW file, the file catalog maintained by the WM contains an entry that includes the NSW name for the file, the location of the file in a constituent host file system, and various other file attributes. Like RSEXEC, the NWS file system supports the use of both partial and complete pathnames; but unlike RSEXEC, complete NSW file pathnames do not include a host name component.

NSW file operations are logically centralized in that the central WM file catalog is almost always used to resolve file references. (Exceptions occur with some tool-bearing host FM processes that maintain a "cache" of information about previous references to NSW files. This cache eliminates the need to interact with the WM when a tool refers to a file previously used by a tool during a particular session.) This approach is somewhat different from that of the RSEXEC file system, which makes more direct use of the constituent host file systems in the sense that it does not maintain a central file catalog independent of those of the constituent host operating systems. In RSEXEC, file operations are decentralized. The RSEXEC program acquires file catalog information with which it maintains the user's composite working directory, by interacting as necessary with various RSSER server programs. This direct use of the constituent host file systems is more feasible in the RSEXEC system because it is homogeneous, designed primarily to work with a network of similar systems—whereas NSW was intended to work with a heterogeneous network of dissimilar systems. On the other hand, the NSW design goal to isolate itself entirely from syntax specific to any host precludes direct use of the constituent file systems.

When a tool opens a file, a copy of the file is moved from its storage space into a workspace maintained by the tool's FM. The file manipulated by the tool is truly a copy, in that any modifications the tool makes to it will not be reflected in the NSW file maintained by the WM unless the FM or the tool explicitly returns the copy into the NSW file system. This "copy on open" mechanism prevents tools from sharing the same file dynamically in the way permitted by many modern single-host operating systems, such as Multics and Tenex. However, it does ensure that the NSW always contains an internally consistent version of each file.

Because of its modular structure, the NSW system architecture is potentially resilient to individual host failures. Indeed, the intercomponent protocols make continued system operation possible in the presence of failures in the front end or tool host processors. In addition, mechanisms have been developed to recover files trapped in a tool workspace by a tool-bearing host crash. When the crashed host is restarted, tool workspaces that were in use when the crash occurred are preserved in a way that allows a user to retrieve selected files, either upon restart or later.

At present the WM is the principal weakness of the system from a reliability point of view. Because the WM and its central data bases reside in a single host, the NSW system is vulnerable to failure of that host. A multi-host implementation of the WM function is being designed that would mitigate this weakness. In addition, it would also allow the system to expand gracefully as the number of users grows, by permitting the WM load to be distributed among several hosts.

The principal technical problem in the multi-host WM results from the fact that the WM data base is logically centralized. To distribute the WM, its data base must be distributed. This distribution requires a synchronization mechanism that ensures that the distributed parts are consistently maintained. Several synchronization mechanisms of this sort have been developed recently.[16]

POINTS OF COMPARISON

As of October 1977, an RSEXEC implementation supporting the features previously described had been operational for over two years. An initial NSW operating capability is just emerging, and implementation efforts are continuing. Some consequences of the different design techniques represented by

Operating Systems for Computer Networks

the two systems give insight into alternative designs for network operating systems. By focusing on the differences between the systems we hope to illustrate the great flexibility open to the designer of an NOS.

Degree of host coupling. One point of comparison is the degree to which the constituent host computers are operationally and administratively coupled. How autonomous are the host computers in the larger context of the network operating system? RSEXEC is completely distributed in concept and implementation. Each constituent host maintains its own descriptive and state information about its own resources. Thus, any RSEXEC host can operate autonomously, although obviously with diminished scope, compared with cooperative operation with other RSEXEC hosts. Because no attempt is made to maintain a long-term system-wide data base of network resources, such autonomous operation may lead to conflicting interpretations of global file names. For example, the existence of multiple copies of a file is evident with RSEXEC only when reestablished dynamically each time an environment is configured for a user.

The RSEXEC hosts permit non-NOS computer activity on NOS resources. In particular, since RSEXEC uses local host file systems directly for maintaining its own file system status information, "network files" may be manipulated from outside the context of the network operating system. This provides more flexible access to these resources, but may also lead to occasional inconsistencies in network operation. These inconsistencies can occur because, as just mentioned, no attempt is made to keep long-term data bases of network resources, nor is file activity that occurs outside of RSEXEC synchronized with the short-term dynamically acquired RSEXEC data bases. The impact of such inconsistencies usually is not severe. Typically, an inconsistency results in the failure of an attempt to obtain a file because the RSEXEC file catalog entry for it is no longer valid. The RSEXEC itself is resilient to such failures; it merely returns an appropriate error indication to the user or his program.

The user can recover from the failure by instructing RSEXEC to rebuild the portion of his file catalog that is no longer valid. In principle, this reconstruction of the file catalog could be automatic. When a file operation fails after having been expected to succeed, on the basis of RSEXEC file catalog information, RSEXEC could automatically reacquire the appropriate file catalog information and reinitiate the operation. The current RSEXEC system does not behave in this way; instead it requires the user to initiate recovery action.

The total autonomy of the RSEXEC hosts has the further administrative implication that a user must register individually with each host whose resources he plans to use. This involves establishing valid accounts, acquiring sufficient resource guarantees, and setting up appropriate access controls for each constituent host. Once this has been done, RSEXEC facilitates the use of the resources on these hosts. When looking at RSEXEC in this context, one should keep in mind the initial concept of RSEXEC as a tool for helping users cope with the administratively autonomous, distributed environment brought about by the advent of the Arpanet

In contrast to the distributed nature of RSEXEC, the initial concept of the NSW is a centrally administered service facility. The NSW has a logically centralized system structure supporting a collection of distributed system resources. The responsibility for managing the distributed resources lies with the WM. Since the WM maintains state information concerning all the distributed resources, a WM host must be part of every NSW configuration and takes part in every operation that utilizes a resource. Thus, autonomous tool-bearing host operation within NSW is not feasible, except in the special case where a WM host is also a tool-bearing host. However, since the state information is maintained apart from the tool-bearing hosts, NSW operation in the absence of a tool-bearing host cannot lead to data base and system inconsistencies. Maintenance of this information external to the tool-bearing hosts is also the basis for achieving uniform NSW file naming.

An individual computer system can be a tool-bearing host within NSW and at the same time directly service non-NSW users. However, the resources supporting the computer as a tool-bearing host are dedicated exclusively for this purpose, and cannot ordinarily be manipulated from outside the NSW environment. This helps maintain the integrity of the globally maintained system resource data base.

The administrative centralization of network resources in NSW means that procedures for establishing access rights to these distributed resources are quite simple. One merely negotiates with a single NSW administrative organization to arrange potential access to all NSW resources. The logical centralization of authentication, resource allocation, and access control functions within a single component simplifies the software implementation of these functions. It also means that a convenient framework is available for preparing a single accounting of NSW resources utilized by an individual user.

Operating Systems for Computer Networks

Visibility of distribution. An important issue to resolve in the design of a network operating system is the nature of the interface to the network resources as seen by a user or a program. Part of this issue is the degree to which the distributed nature of the underlying system is visible. At one extreme is complete invisibility, where a user is not necessarily aware of the network operations being performed on his behalf, nor is he able to exercise direct control over the use of distributed resources. At the other extreme is complete network visibility, where the user is aware of the network and its constituent hosts, and must directly assert control over the selection of resources to service his requests.

The NSW attempts to mask almost all details of network operation for the user, and thus tends toward the invisible end of the design spectrum. This position is based on the belief that dealing with the network or the constituent host systems at any level would impair a user's ability to easily utilize combinations of software development tools. The NSW system itself currently assumes all responsibility for selection and placement of system resources, without notifying or consulting the user. Neither the syntax nor the semantics associated with the NSW user interface includes any provision to specify actions directly relating to the distributed nature of the system. (However, the network aspects of the system intrude somewhat into the user's conceptual model in such areas as the movement of files between NSW file space and secondary storage not under NSW control, and also with the use of tool workspaces. The visibility in these areas represents dealing with the operational realities rather than a change in philosophy.) As a result, all network resources are uniformly and easily accessible within a simple user framework.

A number of factors suggest that a more visible approach to network systems is sometimes more appropriate. In currently available systems, performance frequently decreases appreciably with access to remote resources. Similar resources may have slightly different characteristics on different host types. Additionally, certain operations may be supported only for co-located entities—that is, resources in the same host. For these and other reasons, some system designers have opted for user interfaces that acknowledge the distributed nature of the system and allow users to influence, if not specify, resource selection patterns. Usually coupled with this is a set of judiciously selected system defaults, which specify certain "standard" resources if the user's command omits them. These are chosen to involve the user as little as possible in resource selection decisions. RSEXEC is visibly distributed; users deal with the distribution in terms of physical host names. It also provides mechanisms for operating in an NSW-like fashion where the user is not compelled to specify details relating to distribution.

Extendable program set. Although the designs for both RSEXEC and NSW permit users to execute programs on the network hosts, there are somewhat subtle design differences between them. For RSEXEC, the object program that is to be executed is obtained from the distributed file system. To add to his collection of executable programs, a user need do nothing more than create a new network file. In NSW, however, the object program that a user can execute can be selected only from a tool list, separately maintained by the WM for the entire system. New programs can be added to the available tool set only by the NSW administrative staff. Although at first this might seem arbitrary, it is a result of the NSW philosophy of providing access only to "well debugged, documented programs," and of extending the concept of managing software projects to include the ability to control the use of certain programming tools.

The somewhat closed tool set supported by NSW exemplifies another distinguishing characteristic. NSW is not designed to be the execution facility for completed software, nor to be a general purpose computer utility. As mentioned earlier, its purpose is to support software development. The software created within NSW is meant to be exported to machines outside of the NSW domain for production runs. RSEXEC was conceived in more general terms, and can be viewed as an attempt at a self-contained system which would ultimately lead to a general-purpose computer utility.

Implementation strategies. There are several alternative NOS implementation strategies. One approach is to determine the operations basic to the NOS and implement them directly on the hardware base. Another strategy is to use the basic functions provided by the constituent host operating systems as the building blocks for the NOS; any operations required for the NOS not provided by the base operating system must be included in the software that implements the NOS. This approach was chosen for both RSEXEC and NSW implementations in order to take advantage of the large investment in proven base operating system and application-level software. Furthermore, for these systems the first approach was viewed as infeasible, particularly for NSW, which utilizes several different base hardware systems. However, building upon the hardware base might have resulted in a set of basic functions somewhat better matched to the needs of the NOS.

An example of such a system is Mininet,[17] a special purpose system designed for transaction processing on distributed data bases. This implementation approach is feasible, in part, because the hardware on which it runs is homogeneous, and because there is no large investment in existing application software.

Because they are built upon existing operating systems, neither RSEXEC nor NSW directly performs certain traditional operating system functions, such as processor scheduling or memory management. Rather they rely upon the underlying operating systems to provide these functions. The emphasis of both implementations is to provide the expanded functionality, such as distributed file systems and interhost interprocess communication, required to support users in a network environment.

Another aspect common to both systems is the use of processes to structure the implementations. For example, the RSSER program runs as a service process on each site that participates in RSEXEC. Processes are used to an even greater extent in NSW, where, for example, WM, FE, and FM are all separate processes. The use of processes is attractive for several reasons. First, it isolates one function from the effects of another. Second, the resulting logical separation forces a more explicit definition of the interactions between NOS components.

AN NOS FOR PERSONAL COMPUTERS?

Despite their different and sometimes very specific design goals, RSEXEC and NSW do exhibit general approaches to many of the common problems inherent in building a network operating system. They also have in common the use of large, general-purpose, shared computer systems as basic building blocks. But just as a technological base for such systems is becoming established, other factors are causing reappraisal. For example, system economics are apparently changing, in a way which reduces the need for large, costly system hardware shared by many users, and which leads to the feasibility of dedicated, general-purpose personal computers.

Although hardware economics were the prime motivation for developing large shared facilities, users soon discovered that the systems could also support other forms of sharing. The ease with which users could share programs and data of all forms within these systems has left an indelible impression and set an irreversible trend. Because of this need to share, the widespread use of personal machines is likely to stimulate even more demand for systems that couple the operation of such machines. Thus, a whole new set of NOS design issues providing new challenges for system designers can be expected.

Acknowledgment

This work was supported by the Rome Air Development Center under contract number F30602-76-C-0425, and by the Advanced Research Projects Agency of DOD under contract number N00014-75-C-0773.

REFERENCES

[1]L. G. Roberts and B. D. Wessler, "Computer Network Development to Achieve Resource Sharing," *AFIPS Conference Proceedings, Vol. 36, 1970 SJCC,* pp. 543-549.
[2]R. B. Hovey, "The User's Role in Connecting to Value Added Networks," *Data Communications,* May/June 1974.
[3]W. W. Clipsham, F. E. Glave, and M. L. Narraway, "Datapac Network Overview," *Proceedings Third International Conference on Computer Communications,* Toronto, Canada, August 1976, pp, 131-136.
[4]J. Davidson, N. Mimno, R. Thomas, D. Walden, W. Hathaway, and J. Postel, "The Arpanet Telenet Protocol: Its Purpose, Principles, Implementation and Impact on Host Operating System Design," *Proceedings Fifth Data Communications Symposium,* Snowbird, Utah, September 1977, pp. 4-10—4-18.
[5]E. Feinler and J. Postel, "File Transfer Protocol for the Arpanet," *Arpanet Protocol Handbook,* 1976 Edition, pp. 117-235, National Technical Information Service, Accession No. ADA027964.
[6]R. H. Thomas, "A Resource Sharing Executive for the Arpanet," *AFIPS Conference Proceedings, Vol. 42, 1973 SJCC,* pp. 155-163.
[7]D. G. Bobrow, J. D. Burchfiel, D. L. Murphy, and R. S. Tomlinson, "Tenex, a Paged Time Sharing System for the PDP-10," *CACM,* Vol. 15, No. 3, March 1972, pp. 135-143.
[8]R. H. Thomas, "JSYS Traps—A Tenex Mechanism for Encapsulation of User Processes," *AFIPS Conference Proceedings, Vol. 44, 1975 NCC,* pp. 351-360.
[9]B. P. Cosell, P. R. Johnson, J. H. Malman, R. E. Schantz, J. Sussman, R. H. Thomas, and D. C. Walden, "An Operational System for Computer Resource Sharing," *Proceedings Fifth ACM Symposium on Operating System Principles,* published as *ACM Operating Systems Review,* Vol. 9, No. 5, November 1975, pp. 75-81.
[10]S. M. Ornstein, F. E. Heart, W. R. Crowther, S. B. Russell, H. K. Rising, and A. Michel, "The Terminal IMP for the ARPA Computer Network," *AFIPS Conference Proceedings, Vol. 40, 1972 SJCC,* pp. 243-254.
[11]S. D. Crocker, "The National Software Works: A New Method for Providing Software Development Tools Using the Arpanet," presented at Consiglio Nazionale delle Richerche Instituto Di Elaborazione Della Informazione Meeting on 20 Years of Computer Science, Pisa, Italy, June 1975.
[12]F. J. Corbato, J. H. Saltzer, and C. T. Clingen, "Multics—The First Seven Years," *AFIPS Conference Proceedings, Vol. 40, 1972 SJCC,* pp. 571-583.
[13]R. E. Schantz and R. E. Millstein, "The Foreman: Providing the Program Execution Environment for the National Software Works System," BBN Report No. 3442; also available as Massachusetts Computer Associates Document No. CADD-7701-011, January 1977.
[14]P. M. Cashman, R. A. Faneuf, and C. A. Muntz, "File Package: The File Handling Facility for the National Software Works," Massachusetts Computer Associates Document No. CADD-7612-2711, December 1976.
[15]NSW Protocol Committee, "MSG: The Interprocess Communication Facility for the National Software Works," BBN Report No. 3483; also available as Massachusetts Computer Associates Document No. CADD-7612-2411, December 1976.
[16]R. Peebles and E. Manning, "System Architecture for Distributed Data Management," *Computer,* January 1978, pp. 40-47.
[17]J. Labetoulle, E. Manning, and R. Peebles, "A Homogeneous Network for Data Sharing—Modelling and Analysis," *Computer Networks,* Vol. 1, No. 4, 1977.□

Understanding Satellite-Based Telecommunications Systems

Problem:

Satellites are functionally similar to the more familiar land-locked microwave relay towers, but they offer a much wider transmission path and introduce certain unique delay problems into communications links that are not found in ordinary microwave links. You will probably not be responsible for the detailed, evolving design of satellite communications systems, but you probably will, at some point, become a paying member in a network that uses satellite relays to form all or part of the network linkages. This report explains the basic theory of transmission through satellites and clarifies many of the very basic limitations and capabilities of satellite-based communications systems. Your understanding of satellite-based systems will be an enormous asset in planning for future traffic growth.

Solution:

A communication satellite is, in essence, a microwave relay in the sky. It receives microwave signals in a given frequency band and retransmits them at a different frequency. It must use a different frequency for retransmission; otherwise the powerful transmitted signal would interfere with the weak incoming signal. The equipment that receives a signal, amplifies it, changes its frequency, and retransmits it is called a transponder.

Most satellites have more than one transponder. The bandwidth handled by a transponder has differed from one satellite design to another, but most contemporary satellites (e.g., INTELSAT IV, ANIK, and Western Union's WESTAR) have transponders with a bandwidth of 36 MHz. How this bandwidth is utilized depends on the earth station equipment. The WESTAR satellites, which are typical, may be used to carry any of the following:

- One color television channel with program sound, or
- 1200 voice channels, or
- A data rate of 50 Mbps, or
- The center 24 MHz of each band may relay either

  (a) 16 channels of 1.544 Mbps, or
  (b) 400 channels of 64,000 bps, or
  (c) 600 channels of 40,000 bps.

The WESTAR satellites each have 12 such transponders, two of which are spares used to back up the other 10 in case of failure. Future satellites will have a larger number of transponders.

EARTH STATIONS

A satellite earth station consists of a large dish, which points at the satellite in basically the same way that

---

James Martin, *Telecommunications and the Computer*, Chpt. 15, pp 280-301 © 1969, Second Edition, 1976. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Understanding Satellite-Based Telecommunications Systems

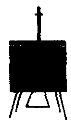
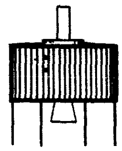
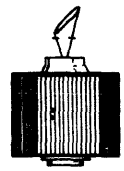
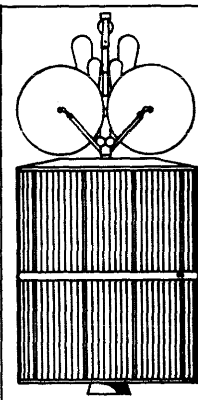
					Being designed
Name	Intelsat I (Early Bird)	Intelsat II	Intelsat III	Intelsat IV	Intelsat V
Year of launch	1965	1967	1968	1971	≈ 1978
Diameter	28 inches	56 inches	56 inches	93 inches	
Height	23 inches	26 inches	78 inches	111 inches	
Weight in orbit	85 lbs	192 lbs	322 lbs	1547 lbs	> 2000 lbs
Number of antennas	1	1	1	3	Multiple
Primary power (watts)	40	75	120	400	1000
No. of transponders	2	1	2	12	≈ 50
Bandwidth of transponder	25 MHz	130 MHz	225 MHz	36 MHz	Variable
Cost of satellite	$3.6 million	$3.5 million	$4.5 million	$10.0 million	≈ $15 million
Cost of launch	$4.6 million	$4.6 million	$6 million	$16 million	≈ $20 million
Design lifetime	1.5 years	3 years	5 years	7 years	10 years
Total cost per year	$5.47 million	$2.70 million	$1.90 million	$3.71 million	≈ $3.5 million
Average No. of voice circuits	240	240	1200	6000	≈ 60,000
Cost/voice circuit/year	$23,000	$11,000	$1600	$618	≈ $58

Figure 1. The INTELSAT birds—four generations of satellites in six years

an earthbound microwave relay dish points at the next tower in the chain. The earth station antenna, however, is larger, giving a narrower beam angle.

The first earth stations were massive. The earth station at Andover, Maine, originally built for AT&T's Telstar satellite and then used with Early Bird and its successors, has a dome 18 stories high housing a huge steerable horn-shaped antenna weighing 380 tons. Its electronic circuits were cooled by liquid helium. Many of today's earth stations, are small enough to be erected quickly in a parking lot behind a factory or office building. Dishes of 20 feet or less are used; whereas the large Comsat earth stations use 100-foot dishes. Initially, earth stations were owned only by the common carriers (and military). Now the small earth stations are owned or leased by private industry and access common carrier satellites.

While most earth stations simply transmit and receive the telecommunication signal with a fixed antenna, at least one must carry out the additional function of controlling the satellite. Western Union's earth stations, for example, are unmanned, except for one station at Glenwood, N.J. that monitors both the satellites and the other earth stations.

During a satellite launch, the Glenwood station has the critical function of maneuvering the satellite into position once it has separated from its launch vehicle. For 7 years or more after the launch, it must maintain the satellites in its correct position by occasionally firing small gas jets on board the satellite. It can send commands to the satellite to turn transponders on and off (to save power), to switch to redundant and backup equipment, and to control the charging of batteries and positioning of antennas.

THE DROPPING COST OF SATELLITE CHANNELS

The first four generations of INTELSAT satellites carried increasing numbers of channels and had progressively longer design lives, as shown in Figure 1. Consequently, the cost per voice channel per year dropped dramatically. The process will continue with INTELSAT IV and V.

Understanding Satellite-Based Telecommunications Systems

The bottom line of Figure 1 shows the drop in cost per satellite voice channel per year. Figure 2 plots the trend. The figure shown is in the investment cost of the satellite and its launch. The cost to a subscriber will be much higher because it must include the earth station and links to it, and must take into consideration the fact that the average channel utilization may be low.

The satellites and their launch costs are referred to as the space segment of satellite communications. The comment is sometimes made among systems planners that the space segment costs are dropping to such a low level that overall system costs will be dominated by the organization of the ground facilities.

The cost of an earth station, however, has dropped much more spectacularly than that of a satellite.

The first Comsat earth stations cost more than $10 million. (The first Bell System earth stations for TELSTAR cost much more.) Earth stations have dropped in cost until now a powerful transmit/receive facility can be purchased for less than $100,000. Receive-only facilities are a fraction of this cost. At the same time, the traffic that can be handled by an earth station is increasing as satellite capacity increases. Combining these two trends we find that the investment cost per channel per earth station is dropping as shown in Figure 3.

The total earth segment costs are not dropping because to provide increased accessibility to the satellites many earth stations are being built. Prior to 1973 the United States had only a handful of earth stations. Now many corporations are setting up their own satellite antennas.

There is a trade-off between the cost of the satellite and the cost of its earth station. If the satellite has a large antenna and considerable power, smaller earth stations can be used. if the satellite makes higher use of its bandwidth allocation, the cost per channel will be lower. There is a limit to bandwidth utilization, and so the main effect of increasing satellite cost will be to reduce earth antenna size and cost.

As the earth facilities drop in cost, more antennas will be constructed and more traffic will be sent, making it economical to use more powerful satellites, which will make the earth facilities drop further in cost.

SATELLITE ORBITS

Modern communications satellites have orbits very different from their experimental predecessors such as AT&T's Telstar satellites and RCA's Relay satellites. The latter traveled rapidly around the earth at a

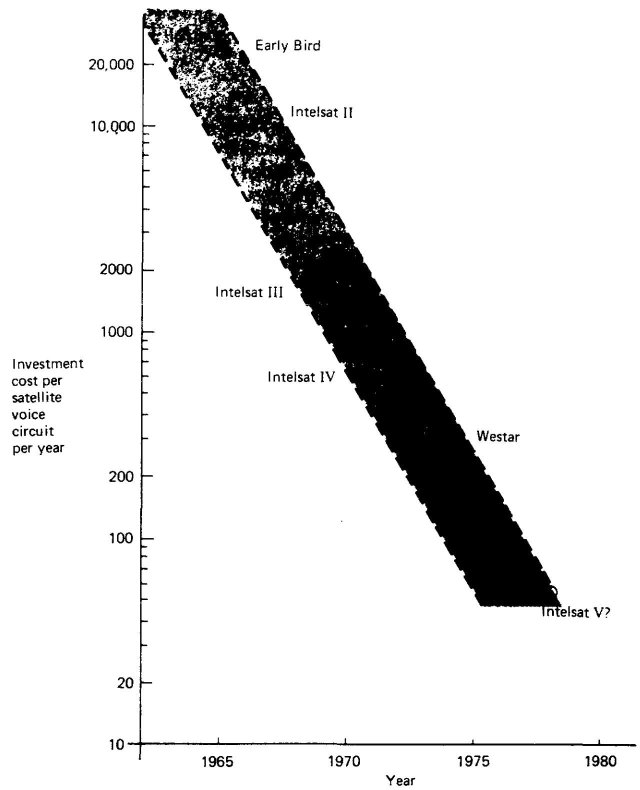


Figure 2. Satellite costs per circuit are dropping rapidly

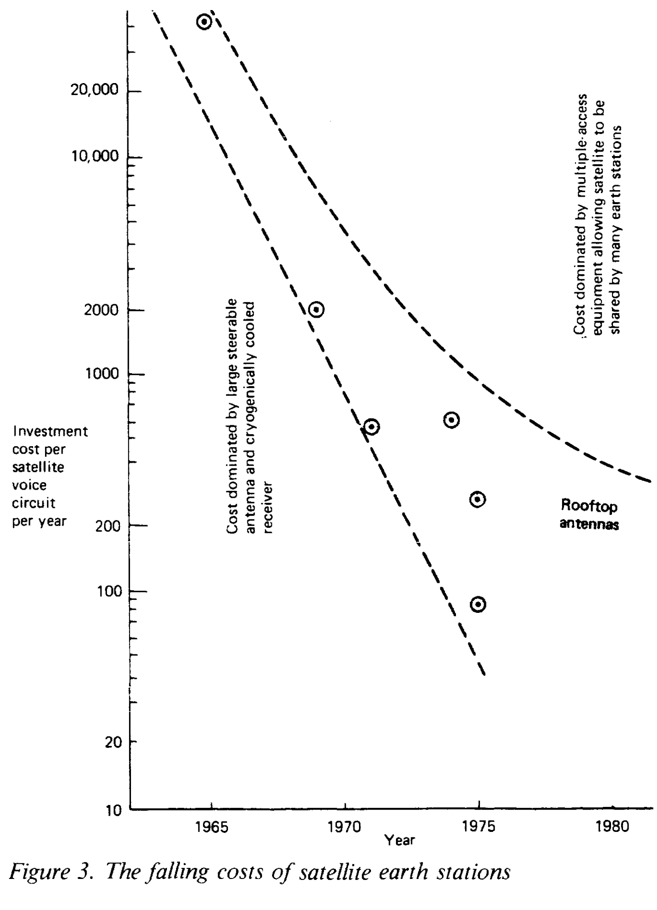


Figure 3. The falling costs of satellite earth stations

Understanding Satellite-Based Telecommunications Systems

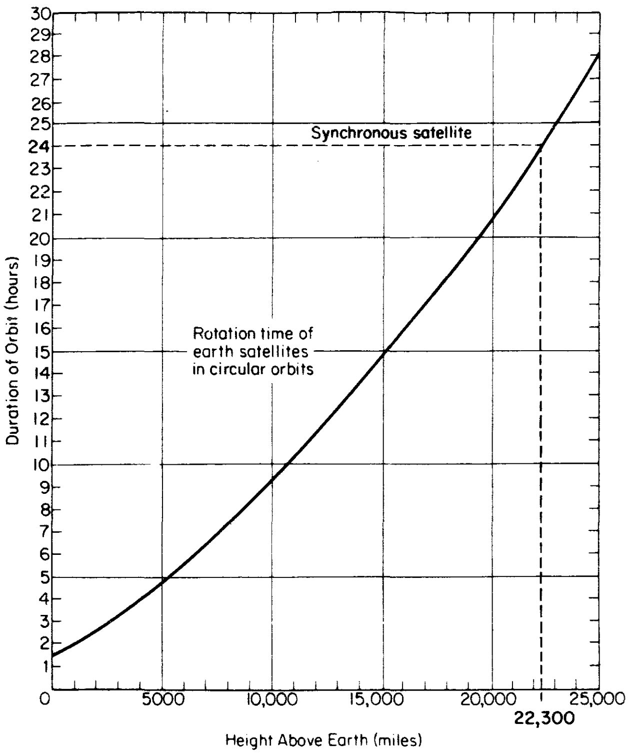


Figure 4. Rotation times of earth satellites in circular orbits

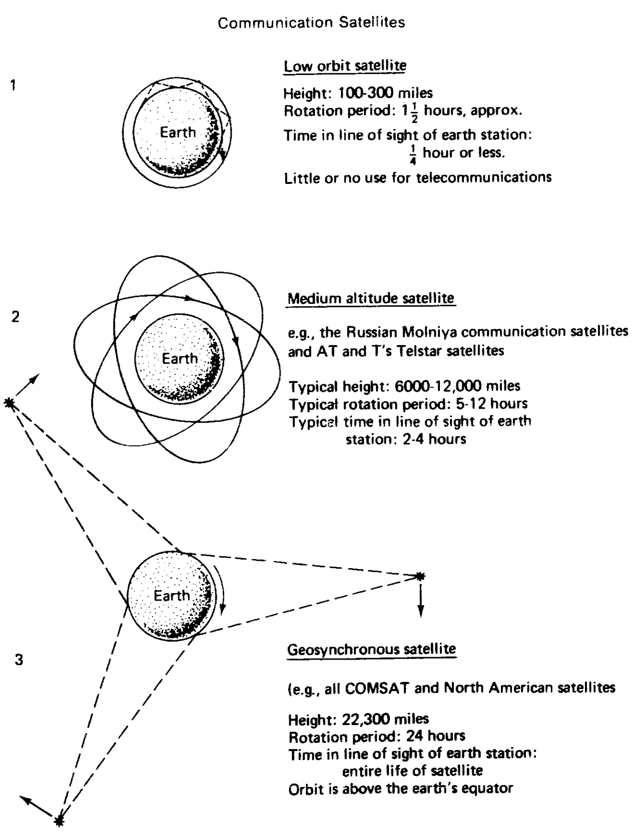


Figure 5. Satellite orbits

relatively low altitude. The Telstar satellites had highly elliptical orbits, Telstar I from about 600 to 3800 miles and Telstar II from 600 to 6200 miles. The apogee of the ellipse was positioned so that the satellite was within line-of-sight of certain stations for as long as possible. As with early manned orbital flights and most other satellites launched in the first decade of space flight, they traveled around the earth in a few hours: Telstar I, 2 hours and 38 minutes, and Telstar II, 3 hours and 45 minutes. Herein lay their disadvantage for telecommunications; they were within line of sight of the tracking station for only a brief period of time, often less than half an hour. The Russians also use elliptical orbits for their Molniya communication satellites, but their orbits are larger, so that the satellites are within sight for longer periods.

Figure 4 plots the time a satellite takes to travel around the earth versus its height. The orbit at a height of 22,300 miles is special in that a satellite in that orbit takes exactly 24 hours to travel around the earth—the earth's rotation time. If its orbit is over the equator and it travels in the same direction as the earth's surface, then it appears to hang stationary over one point on earth. This orbit is called a geosynchronous orbit. The apparently stationary satellite is called a geosynchronous satellite.

The INTELSAT satellites hang stationary in the sky over the Atlantic and Pacific oceans. The U.S. domestic satellites hang over South America or the Pacific ocean west of Equador. Figure 5 shows satellite orbits. It is clear now that the right place for a communications satellite is in geosynchronous orbit.

As shown at the bottom of Figure 6, three geosynchronous satellites can cover the entire earth with the exception of almost unpopulated regions close to the poles. An advantage of using satellites in such a high orbit is that they cover a large portion of the earth. Figure 7 shows the maximum spacing between earth stations for different satellite heights, assuming that 5° is the minimum angle of elevation of the ground station antennas.

The placement of a satellite in a synchronous orbit needs high-precision spacemanship. The launch vehicle first places it into a lengthy elliptical orbit with the highest part of the ellipse about 22,300 miles from earth. This orbit is then measured as exactly as possible, and the satellite orientation is adjusted so that it will be in precisely the right attitude for the next step. When the satellite is at the farthest end of its ellipse, traveling approximately at right angles to the earth's radius, a motor is fired at precisely the right instant to put the satellite in a circular orbit around the earth. The satellite's velocity is then adjusted to synchronize it with the earth's rotation, and its attitude is swung so that its antenna points in the right direction.

Understanding Satellite-Based Telecommunications Systems

During the launching of the first INTELSAT II satellite, the "apogee" motor, which should change the elliptical orbit into the circular one, terminated its 16-second thrust prematurely and left the satellite plunging through space on a large elliptical non-synchronous orbit. Comsat, however, managed to use the satellite. Following its unplanned journey through space with their big antennas, they succeeded in transmitting the first live color TV between Hawaii and the American mainland. It was also used for commercial telephone circuits during those periods when its wanderings brought it within line-of-sight of suitable earth stations.

Once a synchronous orbit is achieved, a satellite needs periodic adjustments to keep it where it is needed. Solar gravitation will pull it slowly out of the equatorial orbit, and irregularities in the earth's gravitational field will make it drift along the equator over a period of months. Solar radiation pressure tends to make the satellite drift. Similarly, its attitude will change slightly, and after a long period its antennas may not point correctly toward earth. To compensate for this, the satellite is equipped with small gas jets that can be fired on commands from earth to make minor adjustments to its velocity and attitude. The jets can be operated in short bursts by signals from an earth station. These bursts nudge the satellite into the position required to keep its orbit "stationary." The orbital position may be adjusted every few weeks. The attitude of some satellites is stabilized automatically with on-board control equipment.

ORBITAL POSITION

Satellites cannot be too closely spaced in orbit; otherwise the up-link microwave beams for adjacent satellites interfere with one another. Five-degree spacing between satellites has been accepted practice; however, it appears feasible to operate satellites using today's frequencies with 3° or 4° spacing. Higher frequency satellites can be spaced more closely.

Figure 6 shows the orbital slots that could serve North America with 3° spacing; 5° elevation of the earth station antenna is about the minimum that can be used. Figure 6 shows two arcs showing the positions of antennas with 5° elevation for satellites at the outer limits of North America. The satellite at 67° longitude cannot reach Newfoundland. There is more freedom in positioning the earth station if not less than 15° elevation of the antenna is needed; Figure 6 shows the locus of 15° elevation antennas for a satellite of 100° longitude. It is desirable that the elevation of the earth station antenna be greater than 0.25° if possible because for a smaller angle the beam passes through enough earth atmosphere and rain to cause much signal loss. Earth stations of higher elevation can be less expensive because they have lower signal gain requirements.

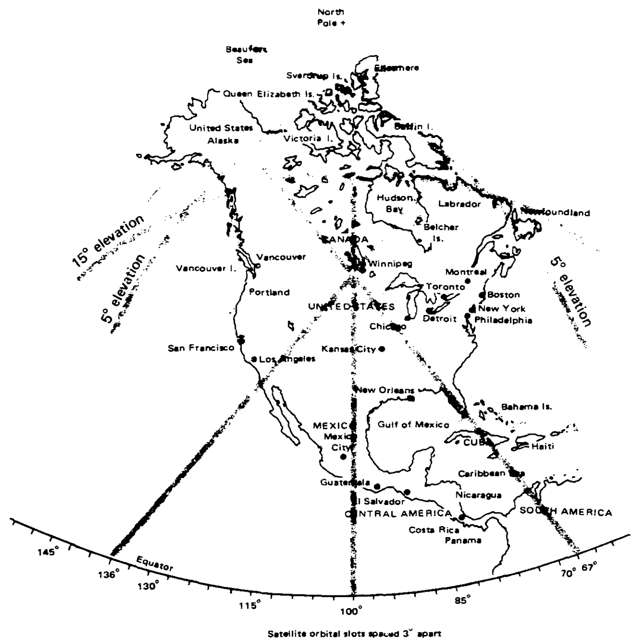


Figure 6. The number of available orbital slots for satellites covering North America is limited, but can be greatly increased by using satellites of higher frequencies

ANTENNAS IN SPACE

Satellite antennas, like all microwave antennas, are directional. Those on the early satellites were not highly directional and transmitted most of their signal into empty space. Later antennas pointed towards the earth as a whole because intercontinental transmissions were needed. At the geosynchronous orbit, the earth subtends an angle of 16°, and the antennas distribute the transmitted energy over this angle.

If the antenna covers a smaller angle, then the signal strength received on earth from a satellite of a given

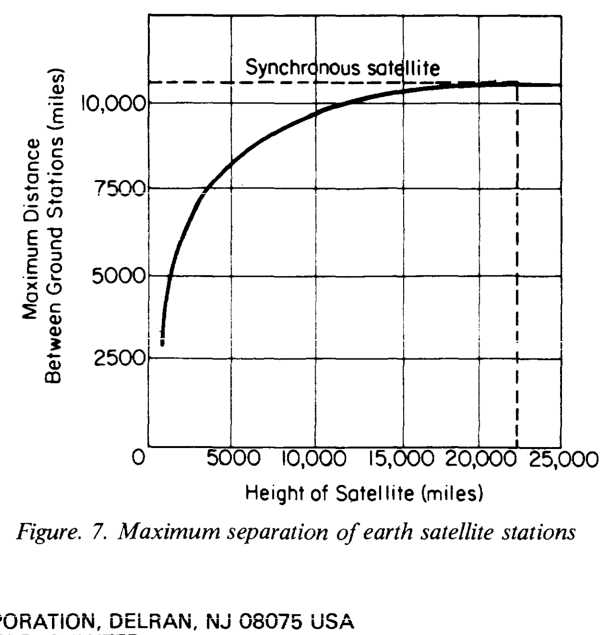


Figure. 7. Maximum separation of earth satellite stations

Understanding Satellite-Based Telecommunications Systems

power is greater, INTELSAT IV carries two narrower-angle antennas subtending 4.5° as well as its 17° earth-coverage antenna. A domestic satellite designed for a country such as Canada or Brazil requires an antenna that focuses on the country.

The NASA ATS 6 satellite has an antenna large enough to transmit narrow beams like search-lights 0.25° in width. This antenna is an umbrella-like structure, 30 feet in diameter, which unfolds in space. The beam, referred to as a spot beam can illuminate and area about the size of England. The satellite has multiple antenna feeds underneath the umbrella, each of which bounces its transmission off the umbrella to form a spot beam traveling in a given direction. Similarly, the umbrella focuses the signal from the area of the spot beam on to the antenna feed when receiving.

There is a trend toward satellites with multiple narrow-beam transmissions. Not only do they give a higher effective radiated power, but also the same frequency can be reused several times for different positions of the earth. The satellite can therefore contain more transponders for a given fixed bandwidth. Different satellites can also reuse the same frequency, thus conserving the valuable spectrum space. If a satellite transmits many spot beams, there should be on-board equipment that can switch signals from one beam to another.

The early satellites with earth-coverage antennas were stabilized about one axis. Satellites transmitting narrow beams require three-dimensional stabilization. To achieve stabilization, most satellites are spun so that they act as their own gyroscope. WESTAR, for example, spins at 100 revolutions per minute. If multiple directional antennas are used, the antenna array must be despun so that it remains stationary relative to the earth. Alternatively, the satellite can contain a gyroscope for stabilization. The use of narrow beams places severe requirements on the systems that stabilize satellites. Typical satellites have their axis held steady to ±0.1°. Their axis can be adjusted by commands from earth, but also on-board control equipment senses the edge of the earth and holds the satellite steady.

Also the satellite position in orbit must be maintained so that it does not drift by more than ±1.0°. This permits the use of cheaper nonsteerable earth station antennas.

DELAY

A disadvantage of satellite transmission is that a delay occurs because the signal has to travel far into space and back. The signal propagation time is about 270 milliseconds and varies slightly with the earth station locations.

The bad effects of this delay have been much exaggerated by organizations that operate long-distance terrestrial links. The claim is frequently heard that the delay is psychologically harmful in telephone conversations and renders satellite links useless for interactive data transmission.

A telephone user may wait for the reply of the person he is talking to for an extra 540 milliseconds if the call goes via satellite in both directions. He certainly notices this delay but very quickly becomes used to it if he makes many satellite calls. Assessment of the psychological effect should not be based on the first satellite call a person makes. The delay seems less harmful to a person who is accustomed to it than the effects of TASI on heavily loaded circuits, which sometimes deletes the first spoken syllables. Transcontinental callers sometimes confuse the effects of TASI with the effect of the satellite delay.

While a telephone user can learn to ignore one or two 270-millisecond delays in a conversational response, four such delays (1080 milliseconds) may strain his tolerance. It is therefore desirable that the switching of calls should be organized so that no connection contains two or more round trips by satellite. Where satellites supplement the terrestrial toll telephone network, the switching can usually be organized to limit the delay to 270 milliseconds. A transatlantic call, for example, often goes one way by satellite and the other way by submarine cable.

In interactive data transmission via satellite, a terminal user will experience a constant increase in response time of about 540 milliseconds. A systems designer has to take this into consideration in designing the overall system response time. In many interactive systems, it is desirable that the mean response time not be greater than 2 seconds.

This is achieved satisfactorily on many interactive systems using satellites today. The line-control procedures selected must be appropriate to satellite channels. It would be inappropriate, for example, for a computer to "poll" devices at the other end of a satellite link, asking them one by one if they have any data to send. Polling is efficient only when the propagation and turn-around time is low. In general, procedures that require protocol signals to travel to and fro before the data message is transmitted are to be avoided on satellite links.

The comment is sometimes made that satellite channels are inappropriate for fast interactive computing with short messages. A more correct comment would be that some commonly-used terrestrial protocols are inappropriate over satellite channels. Various control procedures in use are efficient for interactive systems using satellite channels.

FREQUENCY ALLOCATION

Figure 8 shows the frequencies allocated to satellite use. A different frequency band is allocated to the up-link and down-link. The frequencies are referred to with phrases such as the 4/6 GHz band, the 11/14 GHz band, and the 20/30 GHz band, the first number in each case referring to the down-link band and the second number to the up-link band.

Most of today's commercial satellites use the 4/6-GHz band—the frequencies shown in the left-most diagram. These are the main frequencies used by the terrestrial microwave common carriers. A super-group, mastergroup, or higher group can be taken directly from a terrestrial microwave or coaxial trunk to the satellite transmitter as shown in Figure 9. Similarly, a T1 carrier or other PCM bit stream can be placed directly on the satellite link.

MICROWAVE INTERFERENCE

Unfortunately, because satellite and terrestrial microwave links use the same frequencies, there is a serious problem of radio interference. Four types of interference are theoretically possible:

- Transmission from the earth station interferes with the terrestrial link receiver.
- Terrestrial link transmission interferes with reception from the satellite.
- Transmission from the satellite interferes with the terrestrial link receiver, or
- Terrestrial link transmission is received by satellite.

The first is by far the most serious. An earth station must transmit a powerful signal to compensate for the vast distance and the low gain of the satellite-receiving equipment. The dish-shaped antenna transmits a highly directional beam toward the satellite, but nevertheless some of the signal spills in other directions and may interfere with a microwave receiver. The earth station transmitter must therefore not be too close to a microwave antenna.

The second of the above types of interference is the next most serious. To avoid it, an earth station should not be located close to a terrestrial microwave path so that part of the terrestrial beam shines into the receiving antenna.

Because of the already serious microwave congestion in the cities, earth stations using the 4/6-GHz band cannot be located in many urban areas. In large cities they often have to be 50 or more miles away.

To avoid the interference problems, it seems highly desirable that satellites should use their own frequency bands. When this is done, small satellite earth stations can be on the rooftops of city buildings. The 11/14-GHz and 20/30-GHz bands were allocated with this intent. Unfortunately radiation at 11/14-GHz suffers more absorption by clouds, rain, and the atmosphere than radiation at 4/6 GHz, and 20/30 GHz is still worse. Because of the absorption, higher gain is needed at these frequencies. Several satellites have been designed for operation at 11/14 GHz, but 20/20 GHz, with its higher bandwidths, is still regarded as experimental.

The up-link of 14 to 14.5 GHz does not compete with terrestrial links, and hence a powerful earth station

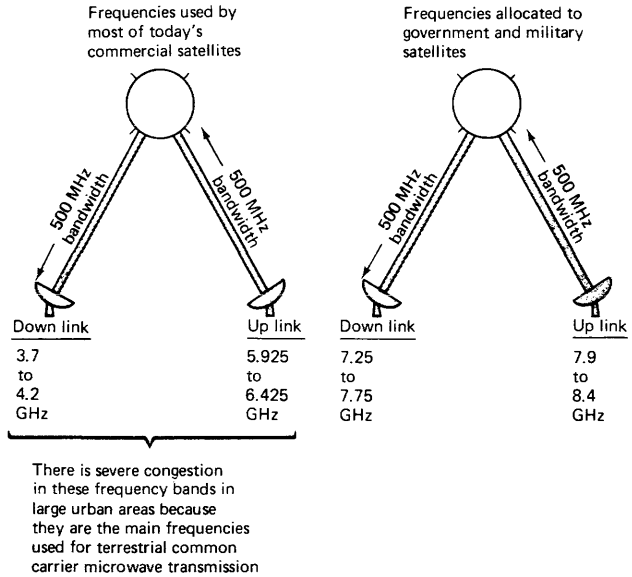


Figure 8. The communication satellite frequency allocations of 500 MHz bandwidth or more, below 40 GHz.

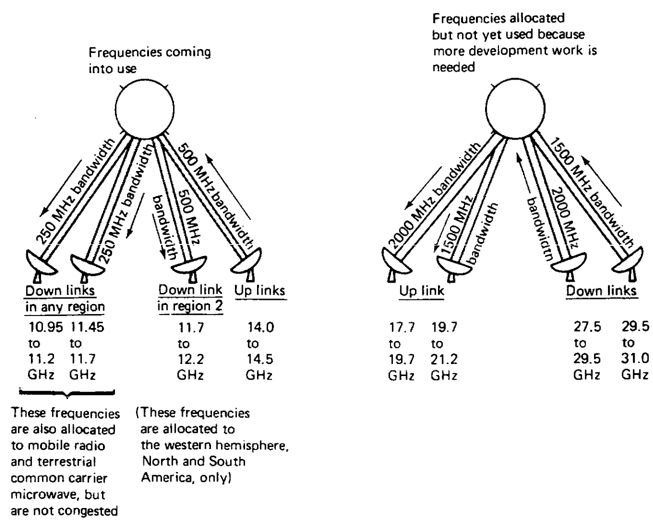


Figure 9. Frequencies allocated for communication satellites at the 1971 WARC (World Administrative Radio Conference)

transmitting from a city rooftop will cause no interference. The down-links at 10.95 to 11.7 GHz can be interfered with, and so the antenna must not be in the path of terrestrial beams at these frequencies.

The 11/14-GHz transmission has several other advantages over 4/6-GHz transmission:

- Antennas for the higher frequency can be smaller and hence cheaper and easy to install on buildings.
- The higher frequency tends to spread out less. Beams tend to be narrower, and so more such beams can crisscross a city without interference.
- Because the beam from the satellite is narrower, its energy is more concentrated, and greater energy tends to be received by an antenna of a given size.
- Because the beam from the satellite is narrower, multiple beams could use the same frequency to different areas on earth, i.e., a form of space-division multiplexing.
- Because the beam to the satellite is narrower, more satellites can occupy the geosynchrous orbit without interference, thereby extending this uniquely valuable earth resource.

ECLIPSES

A satellite is prone to two types of eclipses. First, when the earth's shadow passes across the satellite its solar batteries stop operating. These eclipses last from minutes to slightly more than an hour on 43 consecutive nights in spring and fall. Less commonly, the moon's shadow passes across the satellite, like a solar eclipse on earth. The satellite can carry backup batteries or storage cells to ensure continuous operation if the need for continuous operation justifies carrying the extra weight.

A more serious form of outage occurs when the sun passes directly behind the satellite. The sun, being of such a high temperature, is an extremely powerful noise source and so blots out transmission from the satellite. This outage lasts about 10 minutes on 5 consecutive days twice a year. The only way to achieve continuous transmission is to have two satellites and switch channels to the non-eclipsed one before the eclipse begins. Most satellites are duplicated in orbit, not only to provide protection from eclipses but also protection if one satellite should fail.

Another brief random form of outage is sometimes caused by airplanes flying through the satellite beam.

UNIQUE PROPERTIES OF SATELLITE LINKS

A satellite channel is often used simply as a substitute for a point-to-point terrestrial channel. However, it has certain properties that are quite different from conventional telecommunications. It should not be regarded as merely a cable in the sky. Different types of design are needed, especially in computer systems, to take advantage of satellite properties and avoid the potential disadvantages.

A satellite channel is unique in the following respects:

- There is a 270-millisecond propagation delay.
- Transmission cost is independent of distance. A link from Washington to Baltimore costs the same as a link from Washington to Vancouver. A computer center can be placed anywhere within range of a satellite without affecting transmission costs. It is becoming economical to centralize many computing operations. In an international organization worldwide links can be similar in cost to national links if the regulatory authorities so permit.
- Very high bandwidths or bit rates are available to the users if they can have an antenna at their premises, or radio-link to an antenna, thereby avoiding local loops.
- A signal sent to a satellite is transmitted to all receivers within range of the satellite antenna. The satellite broadcasts information unlike a terrestrial link.
- Because of the broadcast property, dynamic assignment of channels is possible between geographically dispersed users. This can give economies, especially with data transmission on a scale not possible with terrestrial links, but needs new forms of transmission control.
- Because of the broadcast property, security procedures must be taken seriously.
- A transmitting station can receive its own transmission and hence monitor whether the satellite has transmitted it correctly. This fact can be utilized in "contention" forms of transmission control in which two transmissions from different locations might coincide on the same channel and destroy each other.

DIGITAL TRANSMISSION

A 36-MHz transponder on the WESTAR satellite can be used to relay 50 million bps. With advanced modem design, 72 million bps have been relayed. At this bit rate, digital encoding and transmission of voice channels give a higher satellite throughput than analog voice. Most of the recently installed satellite equipment use PCM transmission of voice or digitized voice with some form of compaction differential PCM or delta modulation. Straight PCM uses 64,000 bps for one voice channel. Operating compaction schemes give 32,000 bps.

Understanding Satellite-Based Telecommunications Systems

Digital TV compression has been used to compress color television down to 30 Mbps, thus increasing a transponder's TV channel capacity from 1 to 2.

While such digital techniques can increase the capacity of a satellite, alternatively they can be used to decrease the cost of the earth stations. Earth stations transmitting a lower bit rate can employ a smaller antenna and less expensive equipment.

The future of satellites clearly lies in digital technology, and it emphasizes, once again, the intermingling of telecommunications and computer technologies. Satellite data rates begin to make computer data transmission look inexpensive, at least when compared with established common carrier costs. On the other hand, to take advantage of the unique properties of satellites, computer control of the assignment and switching of channels is essential. It is a complex problem to allocate the satellite capacity to all its potential geographically scattered users so as to maximize the usefulness of that capacity.

Sooner or later, the old established common carriers will have to come to grips with the fact that satellite trunks can be constructed at a fraction of the cost of terrestrial trunks. Several billion dollars per year are being spent on expanding the terrestrial trunk networks.□

A Summary Overview of Communications Network Control Architectures and Protocols

Problem:

We have always been somewhat disinclined to sanction the use of the word architecture as a descriptor for a pile of circuits and other oddments of hardware and software; but we must admit that the sound of the word does lend a certain grandeur and maybe even a bit of respectability to an otherwise dull recital of expressions like digital data communications message protocols, message-independent blocking, X.25 (cryptic), and so forth. We do, however, draw the line at architected and architecturized.

A network control architecture is simply a reasoned method of holding together all the pieces of a communications system by imposing some order on the way this equipment talks to that equipment and on the way messages are composed and routed throughout the network. An architecture consists of two basic parts—protocols and software—both of which consist, in turn, of several discrete levels. The network protocols are drawn from a vast body of standard and semi-standard techniques or are devised uniquely by each vendor depending largely on the level.. The more popular ones are explained in detail in Section CS93. There are no control software standards. Each major computer vendor and communications service purveyor has, of course, its own special insights into the perfect solution to the problems of network control, and each vendor packages these insights into a unique offering, which usually does not recognize the existence of any other offerings.

This report is a compilation of important and current protocols and control packages with brief descriptions of their purposes and general functions. Unless you are already locked into some vendor's architecture, like IBM's SNA (more about SNA later in this service), this report will be an invaluable guide through the morass of network control options.

A Summary of Overview of Communications Network Control Architectures and Protocols

Solution:

The network control architecture is a functional description of the components of a data communications system, including hardware, software, and communications links. It is useful to first describe these network architectures in terms of the data link control or protocol, and then describe some of the other ancillary network structures.

A protocol is essentially a set of rules for defining the communications system. There are several discrete "levels" of a protocol that must be defined:

- Physical interface
- Electrical interface
- Link control
- Message handling

The physical interface refers to the mechanical characteristics of the connection between two components in the data communications system. The number of signal lines and the shape and size of the connector are specified by the physical interface standard. An example of a physical interface standard is found in the Electrical Industries Association (EIA) RS-232C standard for connections between data terminal equipment (DTE) and data communication equipment (DCE).

The electrical interface refers to the electrical signals that are applied to the signal lines of the connection between two components of the data communications system. An example would be the 20-mA current loop in which the line current is switched on and off to represent bits. The RS-232C standard also defines electrical interface requirements. More recent EIA standards, RS-422 and RS-423, are expected to replace the RS-232C standard eventually.

In countries outside the United States, telecommunications standards are promulgated by the Consultative Committee on International Telephone and Telegraph (CCITT) and the International Organization for Standardization (ISO). The CCITT is part of the International Telecommunications Union (ITU), a United Nations agency. There are certain CCITT codes and recommendations that are important in data communications, and it would be worthwhile to list them here.

From *Distributed Processing and Data Communications* by Daniel R. McGlynn. Chpt 5, pp 111-164 © 1978 by John Wiley & Sons. Reprinted by permission.

The CCITT codes are designated by a "V" number, the recommendations by an "X" number. Some of the important CCITT codes concerned with electrical interface standards are:

- CCITT V.10 (X.26)—Electrical characteristics for unbalanced double-current interchange circuits.
- CCITT V.11 (X.27)—Electrical characteristics for balanced double-current interchange circuits.
- CCITT V.28—Electrical characteristics for unbalanced double-circuit interchange circuits.

The next protocol level is that of link control. Data-link control protocols are the rules for transferring data and control information over the data communications link to remote stations. It is useful to classify such protocols into two basic groups: asynchronous protocols and synchronous protocols.

An asynchronous protocol is one in which data is transferred at nonuniform rates. The beginning of each character is marked by a start bit, the end of each character by a stop bit. (Such protocols are, therefore, sometimes called start-stop protocols.) Of course, the use of start and stop bits associated with each character is not efficient, and such protocols are not useful for high data rate applications. The two asynchronous protocols we examine in this report are the Teletype and IBM 2740 protocols.

Synchronous protocols provide for the transfer of data at a fixed rate with the transmitter and receiver operating in synchronization. Synchronous modems are typically used, so that a clock signal is transmitted along with the data stream to ensure that the transmitter and receiver stay in synchronization. Synchronous protocols may be classified into three types, depending on the message-framing format used:

- Character-oriented protocols
- Bit-oriented protocols
- Byte-oriented protocols

In this report we consider the bi-sync character-oriented protocol; the SDLC, ADCCP, HDLC, X.25, BDLC, BOLD, and CDCCP bit-oriented protocols; and the DDCMP byte-oriented protocol.

In analyzing and comparing data-link protocols, the following features must be noted:

A Summary of Overview of Communications Network Control Architectures and Protocols

- Framing
- Line control
- Error control
- Sequence control
- Transparency
- Synchronization
- Time-out and start-up

Framing refers to the rules used for determining (1) which bits constitute characters and (2) the message portion of the data transmitted. As pointed out above, protocols may be either character, bit, or byte oriented, depending on the framing technique.

Line control refers to the rules for determining which station transmits and which station receives on a half-duplex or multipoint line.

Error control refers to the error-checking code used and the ARQ retransmission procedures.

Sequence control refers to the numbering of messages in sequence as they are transmitted in the network so that the entire message may be properly reconstructed. In this connection two different types of packetization should be noted: datagram service and virtual call service. In datagram service, packets are transmitted independently through the network and arrive at the destination in arbitrary order. Sequence control is thus necessary for such service. In virtual call service, the stream of packets is delivered to the destination in the order sent. Such a circuit therefore resembles a physical circuit.

Transparency refers to the ability to transmit data in a message that has the same bit pattern as the control characters. Character-oriented protocols use a special character to designate the beginning and end of the message portion. Bit-oriented protocols utilize a special flag character to identify the message. If the bit pattern of the flag character appears in the message itself, special "bit-stuffing" routines are used to change that bit pattern in a predetermined manner. After receipt of the message, bit-deleting routines are used to reconstruct the original message. Finally, bit-oriented protocols utilize a header that includes a "count" parameter to indicate the number of data characters in the message.

Synchronization refers to the technique used so that the receiving and transmitting stations can maintain synchronous clocks. The transmission of a special synchronization idle character (SYN) is one such technique.

Time-out and start-up refer to a procedure for providing a fixed time during which a given message must be received, and after which alternative action is taken, such as starting up for new transmissions.

Data-link control protocols may also be analyzed in terms of at least three different "levels," corresponding to the levels or layers in the network architecture. The level-0 data-link control protocol refers to controlling transmission between adjacent communication nodes, that is, at the level of the "transmission subsystem" layer. In terms of the format of the transmitted packet, the level-0 protocol refers to the initial or outermost data fields in the packet, including, for example, the SYN characters. The level-1 data-link control protocol refers to transmission control between the source and destination nodes and includes the packet header and other control information. The level-1 protocol thus operates at the "function management" layer and is defined in terms of those fields in the packet adjacent to the level-0 fields. Finally, the level-2 data-link control protocol refers to host/host message handling and operates at the "application layer" of the network architecture. In terms of the format of the transmitted packet, the level-2 protocol refers to the innermost fields of the packet, including the message text itself. When used generically, the term "protocol" is typically understood to mean all three levels, although it must be realized that such a single protocol is in fact composed of several subprotocols.

The following protocols for data communication are considered in this report:

- Asynchronous Protocols
  —Teletype
  —IBM 2740
- Synchronous Character-oriented Protocol
  —Binary Synchronous
- Synchronous Bit-oriented Protocols
  —SDLC
  —ADCCP
  —HDLC
  —X.25
  —BDLC
  —BOLD
  —CDCCP
- Synchronous/Asynchronous Byte-oriented Protocol
  —DDCMP

ASYNCHRONOUS PROTOCOLS

Teletype

The teletype interface is typically a 20-mA current loop represented by the block diagram in Figure 1.

A Summary of Overview of Communications Network Control Architectures and Protocols

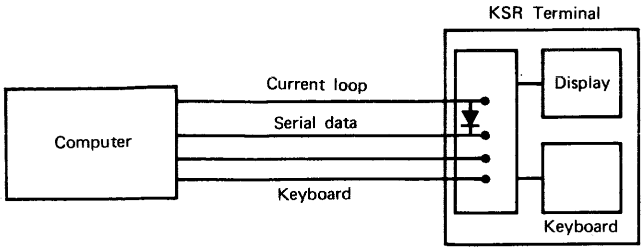


Figure 1. Teletype interface: current loop

The RS-232C serial interface (Figure 2) may also be provided. Actual communication is performed asynchronously, with a message "header" defined by a START bit (space), followed by one or more STOP bits (mark).

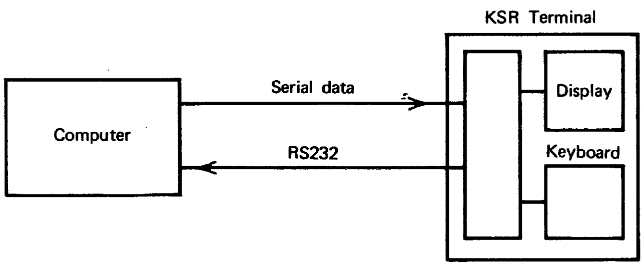


Figure 2. Teletype interface: RS-232C

A teletype-to-telephone communications link may be suitable for the transmission of text, but its lack of polling capability or error-detection makes its application for data transmission less useful or reliable than other techniques.

Teletypewriters fall into one of three general types: automatic send receive (ASR), keyboard send receive (KSR), and read only (RO). The type of lines that may be used with teletypewriters include simplex (send only or receive only), half-duplex (send and receive alternately), or full-duplex (send and receive simultaneously). Data rates generally range from 45 to 150 bits/second.

IBM 2740

The 2740-type protocol is a start-stop half-duplex technique designed for use with the IBM 2740/2741 communication terminals. The sophistication of a protocol can be initially assessed by examining its error-control procedures. The 2740 protocol offers the more basic error-checking procedure, which is a single-bit check on each character. Such single-bit parity checking, called vertical redundancy checking (VRC), is generally adequate for simple applications where high reliability is not essential.

The error-recovery technique used in the 2740 protocol is a simple retransmission based on the reception of a negative acknowledgment (NAK) signal. Since such a technique is basic to many other protocols, it would be worthwhile to review the error-detection and correction technique at this point.

Once a message is received by a receiving station, it is checked for errors using the error-checking technique designed into the protocol. If an error is detected in the message, the receiving station sends a signal back to the sending station requesting a retransmission. Such a retransmission request consists simply of the characters NAK. In order to further increase the reliability of the system, if a message is received correctly by the receiving station, the receiving station also sends a signal back to the sending station requesting transmission of subsequent messages. Such a transmission request consists of the characters ACK (acknowledgment). Once the sending station receives an ACK signal, it will transmit the next message in sequence. If a NAK signal is received instead, it will transmit the previous message, corresponding to the NAK signal, and await an ACK signal. Of course, if neither an ACK nor NAK signal is received, the system may become deadlocked, and other techniques must be used to get the system started again.

The limitations and complexity of data link control operations are apparent even from the simple 2740 protocol. In examining the network control architecture of a data communications system, it is therefore expedient to first examine the characteristics of the protocol before describing the other network structures.

As we suggested above, the 2740 protocol is associated with the IBM 2740 communication terminal, which resembles a SELECTRIC typewriter. The IBM 2740 is a general-purpose terminal effective for internal communication among company departments, as well as remote-batch processing and inquiry type use. The 2740 is part of the IBM 2770 communication system, which is a media-oriented terminal system. The various types of media available with the 2770 communication system include card readers/punches, paper tape readers/punches, magnetic card readers/inscribers, and printers.

It is also worthwhile to point out that certain common carrier teletype systems use essentially the same type of data link protocol as the 2740. For example, in the AT&T 83B-type teletypewriter equipment, a positive acknowledgment signal consists of the letter V rather than the ACK signal.

SYNCHRONOUS CHARACTER-ORIENTED PROTOCOL

Binary Synchronous (Bi-Sync)

The IBM bi-sync (or BSC) protocol is a somewhat more sophisticated data link control for half-duplex

A Summary of Overview of Communications Network Control Architectures and Protocols

operations. It uses a larger number of control characters and permits synchronous operation. Furthermore, more reliability is achieved through better error-checking techniques.

Cyclic redundancy checking (CRC) using 16 bits (CRC-16) is the method of error detection. CRC uses a mathematical algorithm to calculate a block check character (bcc). The value of this bcc is sent from the sending station to the receiving station along with the message. The receiving station also performs the same mathematical algorithm on the message and computes its own bcc. The calculated bcc and the received bcc are then compared to determine if there has been an error in transmission.

Some of the additional control characters found in BSC are as follows:

SYN: synchronous idle. A control character used on synchronous channels for the purpose of initiating or maintaining synchronism between stations. Some systems perform automatic hardware insertion/deletion of SYN characters.

EOT: end of transmission. A character that indicates the end of a particular transmission.

ENQ: enquiry. A control character used to solicit some type of response from the receiving station.

STX: start-of-text. A character that indicates the beginning of the text or message.

DLE: data link escape. A control character used to extend the set of control characters.

The various types of control characters are concerned with the following operational facilities:

- Transmission codes
- Synchronization
- Initialization
- Framing and blocking of text
- Error detection and correction
- Acknowledgment
- End-of-transmission signaling

Such facilities may be applied on point-to-point as well as multipoint operations.

The transmission format for sending a message in BSC thus might appear as follows:

```
S S S S             E  B B
Y Y Y T T  E  X  T  T  C C
N N N X             X  C C
```

It is generally customary to write the control characters vertically. The B C C characters refer to the block check characters.

SNA—SDLC

SDLC (synchronous data link control) is the data link control protocol implemented in IBM's systems network architecture (SNA). Because of the importance of IBM's installed base, and therefore the expected widespread implementation of these concepts and facilities, we will review the basic features of SNA and SDLC in this segment and will examine SNA in more detail in later reports in this Section. The following basic concepts will be considered:

- Systems Network Architecture (SNA)
  —Layered structure
  —Network addressable units
  —Transmission subsystem
  —Path control element
  —Data link control element
- Synchronous Data Link Control (SDLC)
  —SDLC format
  —Primary station operation
  —Secondary station modes
  —Command codes
  —Function management services
  —System services control point
  —Network configuration
  —Communications system protocol

Systems Network Architecture

System network architecture (SNA) is a functional description and definition of all components in a data communications system, including hardware, software, and communications links. Previous data communications systems and networks used a combination of several different computer access methods, different data-link communications protocols, and specialized terminals and lines dedicated to particular applications. SNA aims at providing an integrating network structure that is broad enough to satisfy the diverse requirements of various customer communications system configurations and that is flexible enough to be adaptable to particular dedicated applications.

From the point of view of software or program control, there are three basic structures in SNA:

A Summary of Overview of Communications Network Control Architectures and Protocols

- Access method interface
- Communications protocol
- Communications control program

The access method interface is VTAM (virtual telecommunications access method), which replaces a variety of previously used access method interfaces, including IMS/VS, CIC/VS, TCAM/VS, BTAM, and so on. The access method interface couples the communication system to the host computer system.

The new communications protocol is SDLC (synchronous data link control), which replaces bisynchronous start/stop (bisync) or other protocol techniques. The protocol is the procedure for defining connection, synchronization, control, and message information over a data link.

A new communications control program, network control program (NCP), is also provided. The NCP interfaces with VTAM on one side and with other SNA components, including other network control programs, on the other.

LAYERED STRUCTURE. One of the essential features of SNA, as well as many other similar network architectures, is its "layered" structure. Each network node, regardless of whether it is a host processor or a simple terminal, has the same layered structure, described in greater detail below (see Figure 3).

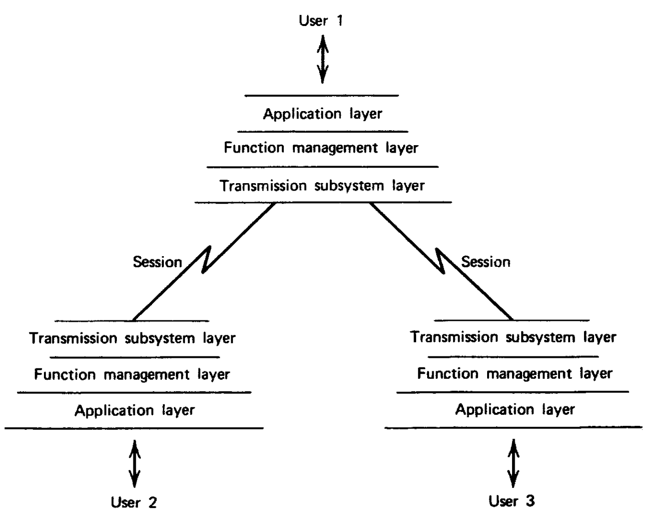


Figure 3. Systems network architecture (SNA): layered structure

SNA may be analyzed, on a functional basis, in three fundamental "layers":

- Application layer
- Function management layer
- Transmission subsystem layer

The application layer defines the customer's particular application, and includes the hardware and software functions for implementing the application programs.

The function management layer is one level higher than the application layer and manages the transfer of information between the layers defined by discrete devices distributed throughout the system network.

The transmission system layer is one level higher than the function management layer and describes the generalized routing and transfer of information between systems nodal points. The transmission subsystem layer comprises three types of hardware elements (Figure 4):

- Data-link control elements
- Path control elements
- Transmission control elements

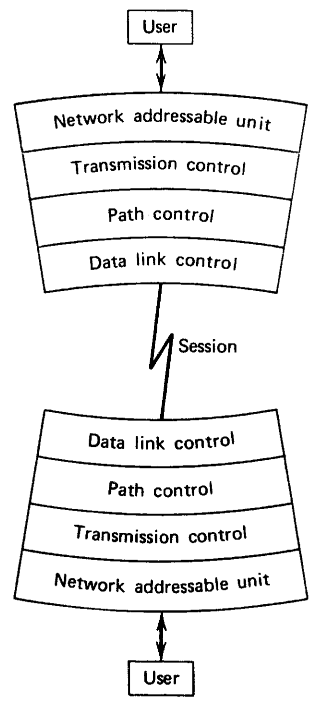


Figure 4. SNA: transmission subsystem

As noted above, the system network consists of a number of transmission links extending between discrete points. The purpose of the data-link control element is to manage these data transmissions links. Path control elements are concerned with the routing of data over particular interconnecting network addresses. Finally, transmission control elements control the linkages or sessions and manage the data flow in the network.

A Summary of Overview of Communications Network Control Architectures and Protocols

An important feature of the layered architecture is that each layer is able to communicate with adjoining layers as well as with corresponding layers in another node. The layers basically serve as a hierarchical interface to the user at the node, which is referred to as the network addressable unit.

NETWORK ADDRESSABLE UNITS (NAU). Each device in the data network that provides the user input or output linkage is known as a network addressable unit (NAU). Each NAU is provided with both a name and an address. The network name is a string of characters by which the users specify the particular NAU in question. The network address is a string of bits by which the data communications system addresses the NAU. A linkage between two discrete NAUs is called a session. A session is defined as a particular communications service over a specified link between two NAUs. A session is implemented by means of a presentation service (PS), a function management element which is part of the NAU. It is possible that a NAU may connect with several discrete NAUs (Figure 5). There are three types of NAUs defined by SNA:

- System services control point (SSCP)
- Physical unit (PU)
- Logical unit (LU)

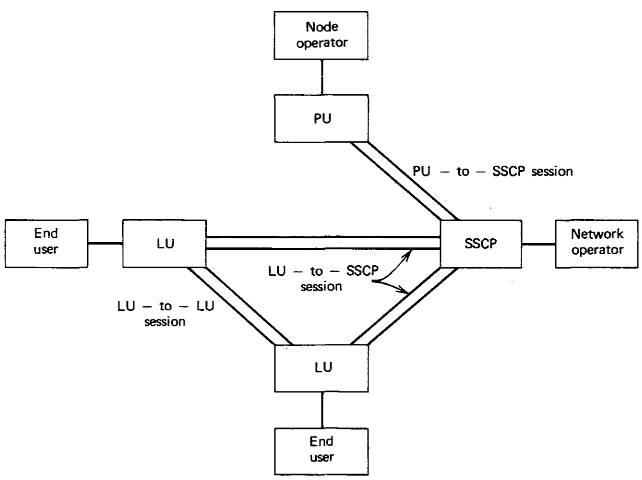


Figure 5. SNA: sessions between network addressable units (NAU)

There are three possible kinds of sessions between NAUs:

1. LU to LU
2. LU to SSCP
3. PU to SSCP

Examples of these sessions are illustrated in Figure 5.

TRANSMISSION SUBSYSTEM: The transmission subsystem, as noted above, consists of the following elements: transmission control element, path control element, and data link control element.

The transmission control element consists of:

- Connection point manager
- Session control
- Network control component

The connection point manager provides the means of communication between NAUs with corresponding elements throughout the common network by means of an information unit known as a request response unit (RU), symbolized by the data field or block shown in Figure 6. A message is defined in SNA as a RU.

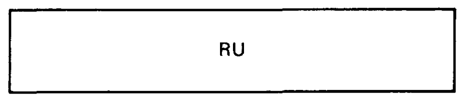


Figure 6. Request response unit (RU)

The actual data frame transmitted over the communication link is built up from the RU by adding a sequence of header fields and finally a trailer field. The header fields describe the destination as well as the type of information to which the header is prefixed. The trailer fields indicate the end of the frame and possibly also contain error control information.

The sequence of header fields appended to the RU correspond to the basic layers of the network architecture:

- Transmission control
- Path control
- Data link control

As will be described shortly, the transmission control element appends a function management header and a request/response header to the RU. The patch control element appends a transmission header, and the data link control element appends the data link control header (e.g., the SDLC header).

As the transmitted data frame travels from node to node, the frame penetrates only those layers of each node as required. Each layer strips the corresponding header from the frame and processes the information contained therein. Thus, for example, the path control element in a receiving node will check the address to see if the data was originally intended for that node. If not, the frame will not penetrate any further layers of that node but will be reconstructed and routed to another node along a path toward its ultimate destination.

The processor adds a request response header (RH) to the beginning of the request unit of Figure 6. for identification. The RH-RU combination is defined as a basic information unit (BIU), as shown in Figure 7.

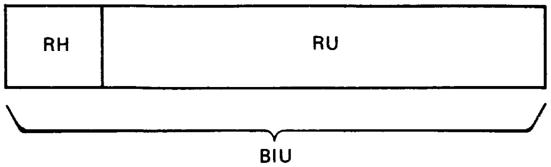


Figure 7. Basic information unit (BIU)

The session control component provides a means of implementing a session between NAUs and coordinating the system resources required to maintain the session. The component also includes means for resynchronizing the information flow if a transmission error occurs.

The network control component provides a control communication path through the systems network using the same sessions that have been already established between NAUs.

The path control element routes the BIUs through the system network. The path control unit thus formats a BIU for transmission of an information unit, which is defined as a path information unit (PIU). The first step in this first implementation is the fragmentation of a large BIU into shorter uniform segments (Figure 8). Each PIU segment is then supplied with another header, a transmission header (TH) (Figure 9). The TH includes data bits that specify the destination address, mapping or segment indicator, sequencing, and related transmission information.

A sequence of these PIUs, each having a possibly different transmission header, may then be grouped together by the path control element into a single basic transmission unit (BTU), as shown in Figure 10.

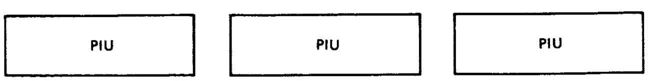


Figure 8. Path information units (PIU)

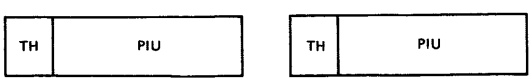


Figure 9. PIU with transmission header (TH)

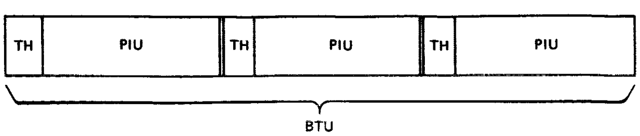


Figure 10. Basic transmission unit (BTU)

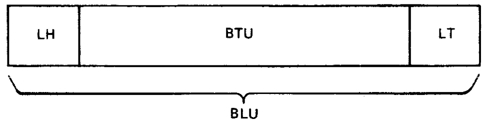


Figure 11. Basic link unit (BLU)

The data link control (DLC) element manages a particular data link. A header (LH) and a trailer (LT) is added to the BTU to form yet another new information unit defined as a basic link unit (BLU), as shown in Figure 11. It is the BLU that is actually transmitted over the data link transmission facility. The BLU is transmitted to the next node, an NAU, where it is decoded and utilized at that node or transferred to another node. The first element in the NAU that is encountered at the node is the DLC element. The DLC element strips the link control information (the LH header and LU trailer) from the BLU and transfers the remaining information unit, a BTU to the adjacent path control element. The path control element breaks the BTU into individual PIUs and examines the particular destination transmission control bits contained in the TH. If, after examining the TH's, it is determined that the BTUs are destined for a different node, the BTU will be reformed and passed on to an adjacent DLC element, which then reassembles the BLU and retransmits the BLU along the transmission facility to the next node.

BIT-ORIENTED PROTOCOLS

Synchronous Data Link Control (SDLC)

Synchronous data link control (SDLC) is a data link control protocol particularly adapted for implementation on SNA. SDLC is characterized by a number of basic architectural features:

- The use of a common grammar
- Increased reliance on the data link facility for error detection and recovery
- Two-level hierarchy consisting of primary stations and secondary stations
- Specific format for each data transmission block called a frame

SDLC FORMAT. The basic structure of an SDLC frame is shown in Figure 12. The frame is divided into

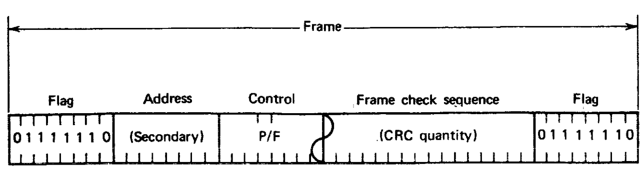


Figure 12. Synchronous data link control (SDLC): frame

six fields: an eight-bit flag field (F), an eight-bit address field (A), an eight-bit control field (C), a variable-length data or information field, a 16-bit block check sequence (BC), and an eight-bit ending flag field (F).

The flag field is a unique bit sequence used to designate the beginning and end of each frame. The actual eight-bit sequence used is 01111110. Provisions are made in the system architecture so that this particular sequence is prevented from appearing in any other position other than the flag field. The actual technique used is a test for five contiguous ones, after which a binary zero-bit is automatically inserted. In a received bit stream, after the flag has been detected, the inserted zeros are automatically deleted.

The address field serves to designate the particular secondary station to which the frame is addressed. Of course, one address may designate more than one station.

The control field is shown in Figure 13. Three basic formats may be utilized in the control field for different purposes, as shown in the figure: an information transfer format, a supervisory format, and a non-sequenced format. The control field contains frame sequence information and a poll/final (P/F) bit which acts as a send/receiver control signal.

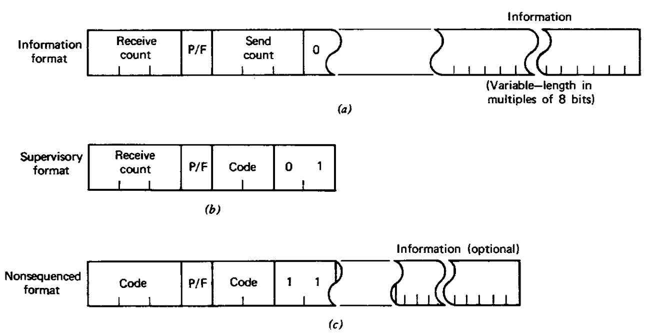


Figure 13. SDLC: control field. (a) information format; (b) supervisory format; (c) nonsequenced format

Frame sequence information is defined by two counts, which may be incorporated in the control field. The transmitting station counts each frame transmitted and numbers them sequentially. These numbers are designated Ns and are inserted in the fourth, fifth, and sixth bit-positions in the control field in the information transfer format (Figure 13a). The receiving station also counts the frames it receives and numbers them sequentially. Then numbers are designated Nr and are inserted in the zero-, one-, and two-bit positions of the control field.

The frame sequence information is used as a means of identifying the occurrence of an error, as well as acknowledging correctly received frames to the transmitting station. An important aspect of this procedure is that up to seven frames may be received before acknowledgment is necessary. A single frame is then sent back to the transmitting station in which the Nr field specifies the sequence number of the next frame that is expected to be received by the receiving station. The transmitting station then checks the Nr and, if incorrect, concludes that an error must have been made at the receiving station. The transmitting station then terminates its current transmission, and returns to the frame sequence number specified by Nr. This specified frame and subsequent frames are then transmitted again to the receiving station.

The P/F bit is used to signify whether a "sending" or "receiving" operation is taking place. A poll bit is sent to a secondary station to indicate that a transmission is requested. A final bit is sent from a secondary station in response to a poll bit-containing frame.

A distinction must be made here between half-duplex and full-duplex communications networks. In a half-duplex system, the line must "turn around" and let the secondary station respond to the poll. In a full-duplex system, the response may be transmitted simultaneously with the polling signals.

The supervisory format (Figure 13b) is used to designate ready or busy conditions or for special purposes. For example, it may be used to check the operability of a secondary station even though no data is to be transmitted to the secondary station.

The nonsequenced format (Figure 13c) is used for data link management, such as initialization of secondary stations, and various control functions. This particular format is called "nonsequenced" since frames including a nonsequenced format control field are not counted in the Nr or Ns counts.

The actual designation of the information transfer, supervisory, and nonsequenced format is made in the six- and seven-bit positions of the control field, as shown in Figure 13.

The information field is any multiple of eight bits in length and contains the data for a message that is to be transferred from station to station in the communications network.

The frame check sequence (FCS) field, or the block check field, is 16 bits in length. The FCS field serves as an error check field by representing a mathematical transformation of all the bits in a given frame. This representation is inserted into the FCS field by the transmitter, while the receiver performs a similar computation on all of the bits of the frame at the secondary station and compares the computed value with the value found in the FCS field. If the computed value and the value of the FCS field do not match, the secondary station rejects the frame as being erroneous.

A Summary of Overview of Communications Network Control Architectures and Protocols

The particular mathematical transformation performed on the bits of the frame is known as cyclic redundancy checking.

PRIMARY STATION OPERATION. The primary station includes two time-out functions used as control operations: idle detect and non-productive receive.

Idle detect gives a means of detecting a nonresponse condition from a secondary station when a response should have been received. A number of factors must be considered to determine an appropriate time period in which a response should be received: the propagation time to the secondary station, the processing time at the secondary station, the operational characteristics of the secondary station modem, and the propagation time from the secondary station. The timing factor required for various modes of transmission is shown in Figure 14. Once the set time period is exceeded, the idle detect function in the primary station assumes that a transmission failure has occurred and initiates appropriate recovery or retransmission by the primary station.

Communications Channel	*Secondary Station Modem Clear-to-Send*	*Approx. Two-Way Propagation Time*
Switched (through local exchange only) or very short (distance) private line	0 msec to 25 msec	2 msec/15 miles (X)
Long (distance) duplex private line	0 msec to 25 msec	2 msec/150 miles + 24 msec (Y)
Long (distance) half-duplex (switched or nonswitched)	75 msec to 250 msec	2 msec/150 miles + 24 msec (Y)
Satellite duplex (switched or nonswitched)	0 msec to 250 msec	600 msec + 24 msec (Z)

Figure 14. Transmission timing factors

The second time-out function is the nonproductive receive function. In this case, a return signal may be received from the secondary station, but that signal may not be intelligible. Such a situation is known as a nonproductive receive, which is again detected by the primary station, and recovery or retransmission measure is initiated.

An abort condition is also a function of the primary station. An abort consists of a transmission of eight consecutive binary ones by the transmitting primary or secondary station. The data link is shut down and returns to the idle state.

SECONDARY STATION MODES. A secondary station is characterized by one of three possible modes: a normal response mode (NRM), a normal disconnect mode (NDM), an an initialization mode.

A normal response mode is one in which the secondary station does not initiate any unsolicited transmissions and transmits only in response to a poll from a primary station. The normal disconnect mode is one in which the secondary station is off-line and only responds to a test or other supervisory command from the primary station.

The initialization mode is one in which transmission to the primary or secondary station is initiated by a specific routine in the respective stations.

COMMAND CODES. Specific supervisory commands that are possible using the nonsequenced format of the control field are defined by specific bit patterns placed within the control field. These bit patterns and their corresponding control functions are shown in Figure 15.

FUNCTION MANAGEMENT SERVICES. As noted above, four types of function management services are provided within the different NAU types. These are presentation, logical-unit, physical-unit, and network services.

The first important element of the function management service is the data flow control protocol support. The protocol is implemented by means of a discrete set of encoded requests, called data flow control (DFC) requests. These data flow requests are exchanged between DFC elements for managing the structures and states of the data flow.

The presentation services provide communication support between users communicating along a session.

SYSTEM SERVICES CONTROL POINT. The system services control point (SSCP) is operable with logical-unit sessions, physical-unit sessions, and the network itself.

The SSCP logical-unit sessions are engaged in supporting the logical-unit control and use of the communications network. The logical-unit services enable the user to engage in SSCP control functions. SSCP commands and replies may be either field formatted ("formatted") or character coded ("unformatted"). Character coded commands are translated by the SSCP into field formatted commands for execution. The SSCP then routes the instruction to an associated command preprocessor for transmission to the designated network services command processor for execution.

For example, the user may desire to create or destroy sessions between logical units. The user may initiate a program—suppose it is designated LOG ON. The LOG ON character string is sent to the LU—SSCP session. The LOG ON character string is then decoded by a syntax scanner to a field formatted "initiate" request. The initiate request contains the name of the logical unit to which a communications link is desired. Other information is also provided within the request

Format*	Sent Last ↓ Binary Configuration ↓ Sent First	Acronym	Command	Response	I-Field Prohibited	Resets Nr and Ns	Confirms Frames through Nr-1	Defining Characteristics
NS	000 P/F 0011	NSI	X	X				command or response that requires nonsequenced information
	000 F 0111	RQI		X	X			initialization needed; expect SIM
	000 P 0111	SIM	X		X	X		set initialization mode; the using system prescribes the procedures
	100 P 0011	SNRM	X		X	X		set normal response mode; transmit on command
	000 F 1111	ROL		X	X			this station is off-line
	010 P 0011	DISC	X		X			do not transmit or receive information
	011 F 0011	NSA		X	X			Acknowledge NS commands
	100 F 0111	CMDR		X				nonvalid command received; must receive SNRM, DISC, or SIM
	101 P/F 1111	XID	X	X				system identification in I field
	001 0/1 0011	NSP	X		X			response optional if no P bit.
	111 P/F 0011	TEST	X	X				check pattern in I field.
S	Nr P/F 0001	RR	X	X	X		X	ready to receive
	Nr P/F 0101	RNR	X	X	X		X	not ready to receive
	Nr P/F 1001	REJ	X	X	X		X	transmit or retransmit, starting with frame Nr
I	Nr P/F Ns 0	I	X	X			X	sequenced I-frame

*NS: nonsequenced, S: supervisory, I: information.

Figure 15. SDLC: control field command codes

format. The SSCP functions to transform the logical unit into a network address. A control initiate command is then generated by the SSCP and is transferred to the primary logical unit in control of the particular network position desired by the user.

The primary logical unit then either accepts or rejects the session request. If the request is accepted, the logical unit transfers the control initiate command into a "bind" command and transfers it to the secondary logical unit in the terminal. The bind command serves to establish a session between the two logical units.

Once indication is received that the session circuit is established, the primary logic unit sends a "clear" signal, followed by an "unbind" signal.

The SSCP physical unit (PU) provides services for each physical unit in the system network configuration. Sessions exist between the SSCP and the PU.

The SSCP essentially performs a network administrative function in processing commands and acknowledgements used in the creation and destruction of a particular physical network linkage, as well as for recovery and resynchronization after a network failure.

The network services provided by the SSCP include:

- Configuration services for the activation and the deactivation of logical units and data links.
- Maintenance services for testing of network facilities.
- Session services by providing means for establishing or terminating sessions between logical units.

NETWORK CONFIGURATION. The physical configuration of the network is defined in terms of four specific node types:

- Host
- Communication controller
- Cluster controller
- Terminal

A Summary of Overview of Communications Network Control Architectures and Protocols

A host node is a multipurpose facility that is engaged in general systems operation, such as the execution of application programs or management of data bases. An example of a host node is a System/370 computer operating under VTAM.

A communications controller node is concerned with the control of the communications lines. An example of a communications controller node is an IBM 3704 or 3705 operating under NCP/VS.

A cluster controller is a node that services a wide variety of peripheral devices operated by specific users. Examples of cluster controller nodes are the IBM 3601 and 3791.

A terminal node is a specific user device, such as an IBM 3767 or other data terminal.

The 370X communications controllers (e.g., the 3704 and 3705) perform both the basic and more complex network management functions, including:

- Data link control
- Dynamic buffering
- Control character insertion and deletion
- Character codes translation
- Error recording and line statistics
- Line control
- On-line diagnostics

The controllers operate under the following operating systems and access methods: OS/TCAM (except remote), OS/VS TCAM (except remote), OS/VS VTAM, and DOS/VS VTAM. A network control program (NCP) resides in the 370X system and performs in software the hardware function previously performed by the IBM 2701 data adapter, and 2702-2703 transmission control units.

The IBM 3704 communications controller is a programmable unit that improves CPU processing throughput by performing various message control program functions previously performed by the CPU, including: addressing, polling, interrupt serving, error recovery, editing, and code translation. The 3704 is able to handle up to 32 low-speed start-stop type lines, eight synchronous type lines, and up to two wideband lines.

The IBM 3705 is similar in function and application to the 3704, but is able to handle up to 352 low-speed start-stop lines and to perform some of the teleprocessing functions previously executed by the processor.

COMMUNICATIONS SYSTEM PROTOCOL. Information transferred between users and defined physical facilities makes use of requests and responses that take place between paired function interpreters. Four particular types of function interpreter pairs are recognized and processed:

- Function management pair
- Data flow control pair
- Session control pair
- Network control pair

The function management protocol defines a number of different operational modes. These include delayed-control modes, and immediate- and delayed-response modes. An "immediate" mode means that the issuer will transmit a single RU and wait for a response before sending another RU. A "delayed"-control mode means that the user may send many requests before waiting for a response.

Within the delayed-control mode there are two options: immediate request and delayed request. The immediate request indicates that the user may send a number of requests, but only the last such request may indicate response. The delayed request allows the issuer multiple requests without waiting for an intervening response. Each such request may require any particular form of response.

The immediate and response modes indicate the manner by which the receiver of a request returns a response. The immediate-response mode specifies that the responses must be returned in the same order in which requests are received. The delayed-response mode enables the receiver to accept a number of requests before responding, and the responses may be returned in an order different from that in which the requests were received.

The function management protocol also enables the type of response to be specified. A definite response in which no more requests will be issued until an acknowledgment of the definite response has been received is one such specification. Another specification is that of an exception response, which requests that the responses be acknowledged only on detection of an error condition. Such capabilities are implemented by means of control bits in the RH portion of a basic information unit.

Another type of control that may be implemented is chaining consecutive RUs. A chain consists of a pre-

A Summary of Overview of Communications Network Control Architectures and Protocols

Type of System	Salient Features	Subcategory Designation
One-Way-Only Systems		
Nonswitched multipoint	Master status permanently assigned Single or group selection without replies	1.1
Point-to-Point Two-Way Alternate Systems		
Switched	No identification Calling station has master status initially Terminate Mandatory disconnect	2.1
Switched	Station identification Calling station has master status initially Terminate Mandatory disconnect	2.2
Nonswitched	Contention Replies Terminate	2.3
Multipoint Two-Way Alternate Systems		
Centralized operation	Polling Selection (single slave) Control-tributary communication only Return to control on termination	2.4
Centralized operation	Polling Selection or fast selection (single slave) Control-tributary communication only Return to control on termination	2.5
Noncentralized operation	Polling Selection (single slave) Tributary-tributary communication permitted Return to control on termination	2.6
Centralized operation	Polling Selection (multiple slave) Control-tributary communication only Return to control on termination Delivery verification	2.7
Noncentralized operation	Polling Selection (multiple slave) Tributary-tributary communication permitted Return to control on termination Delivery verification	2.8
Point-to-Point Two-Way Simultaneous Systems		
Switched	Station identification Both stations have concurrent master and slave status Mandatory disconnect	3.1

Source: American National Standard X3.28-1976.

Table 1. Establishment and termination subcategories

Type of System	Salient Features	Subcategory Designation
Message-oriented	Without replies Without longitudinal checking	A1
Message-oriented	Without replies With longitudinal checking	A2
Message-oriented	With replies Without longitudinal checking	A3
Message-oriented	With replies With longitudinal checking	A4
Message-associated blocking	With longitudinal checking Retransmission of unacceptable blocks Single-character acknowledgment	B1
Message-associated blocking	With longitudinal checking Retransmission of unacceptable blocks Alternating acknowledgments	B2
Message-independent blocking	With longitudinal checking Retransmission of unacceptable blocks Alternating acknowledgments Noncontinuous operation Nontransparent heading and text	C1
Message-independent blocking	With longitudinal checking Retransmission of unacceptable blocks Modulo-8 numbering of blocks and acknowledgments Continuous operation Nontransparent heading and text	C2
Message-independent blocking	With cyclic redundancy checking Retransmission of unacceptable blocks Alternating acknowledgments Noncontinuous operation Transparent heading and text	D1
Conversational	Without blocking Without longitudinal checking	E1
Conversational	With blocking With longitudinal checking	E2
Conversational	With blocking With longitudinal checking With batch transmission capabilities	E3
Message-associated blocking for two-way simultaneous transmission	With longitudinal checking Retransmission of unacceptable blocks Alternating acknowledgments Embedded responses	F1
Message-independent blocking for two-way simultaneous transmission	With longitudinal checking Retransmission of unacceptable blocks Modulo-8 numbering of blocks and acknowledgments Continuous operation Nontransparent heading and text Embedded response	F2

Source: American National Standard X3.28-1976.

Table 2. Message transfer subcategories

determined number of RUs with a well-defined beginning and end defined by specific control bits in each RH portion of the basic information units in the RU chain.

Advanced Data Communications Control Protocol (ADCCP)—ANSI

The advanced data communications control protocol (ADCCP) is the popular name for the protocol established by the American National Standards Institute, Inc. (ANSI), and initially defined in the X3.28-1976 standard entitled "Procedures for the Use of the Communication Control Characters of American National Standard Code for Information Interchange in Specified Data Communication Links."

The ASCII control characters referred to in the title are:

SOH: start-of-heading
STX: start-of-text
ETX: end-of-text
ETB: end-of-transmission block
ENQ: enquiry
ACK: acknowledgment
NAK: negative acknowledgment
SYN: synchronous idle
DLE: data link escape

The standard defines a number of generalized data communication control procedures based upon the combination of two subcategories: (1) establishment and termination (see Table 1) and (2) message transfer (see Table 2).

A particular category of data communication is specified by one type of establishment and termination procedure (selected from Table 1) with one type of message transfer procedure (selected from Table 2)—for example, category 2.2/B1 consists of the establishment and termination subcategory 2.2 together with message transfer subcategory B1.

A Summary of Overview of Communications Network Control Architectures and Protocols

Standards that prescribe the signalling speed and message format are being developed by Task Group X3S34.

The frame structure of the ADCCP data link control protocol is the same as that of SDLC. The command/response codes of the control field are summarized in Table 5.

High Level Data Link Control (HDLC)—ISO

High Level Data Link Control (HDLC) is the name of the protocol being formulated by the International Standards Organization. The frame structure standard, shown in Figure 16 has been approved and published as standard IS 3309. The procedure standard, which involves the definition and function of the bits in the control field of the frame, is being considered by member organizations and representatives at this writing. Details on some preliminary recommendations of the ISO are presented in Table 5.

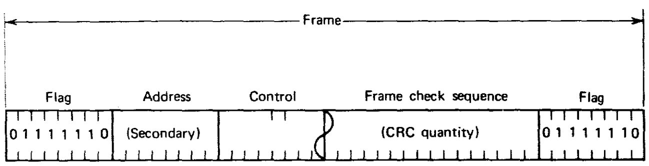


Figure 16. X.25: frame format

X.25—CCITT/150

The X.25 protocol, proposed by the CCITT and adopted by the International Standards Organization (ISO), is a recognized standard for international data communications. Both the European PTT's (Post, Telephone, and Telegraph authorities) and the ITU (Internation Telecommunications Union) have accepted X.25 as the data communications standard.

Because of the importance of X.25, it is worthwhile to summarize some of its significant features. All messages are assembled into frames, which take one of two possible formats shown in Figure 16. There is an eight-bit opening flag, followed by an eight-bit address field, an eight-bit control field, an optional information field of n-bits, a 16-bit frame-checking sequence, followed by a final eight-bit flag.

The flags, address field, information field, and frame check sequence are standard in HDLC and other protocols and need not be described further here. The control field does, however, merit our attention. There are three types of control field formats:

- Information transfer (I frames)
- Numbered supervisory functions (S frames)
- Unnumbered control functions (U frames)

Control Field Bits	1*	2	3	4	5	6	7	8
I frame	0		N(S)		P/F		N(R)	
S frame	1	0	S		P/F		N(R)	
U frame	1	1	M		P/F		M	

N(S): transmitter send sequence count; N(R): transmitter receive sequence count; S: supervisory function bits; M: modifier function bits; P/F: poll bit when used by primary final bit when used by secondary (one when poll or final).
*: The least significant bit, or first bit transmitted.

Table 3. Control field format

The contents of the specific bit locations of the control field for each of these frame types are shown in Table 3, similar to that for SDLC.

The similarity between the X.25 and SDLC control fields is readily apparent, but differences will be noted later in this report. Furthermore, the information field in SDLC is a multiple of eight bits long, while the information field in X.25 is an arbitrary N bits long.

Unlike SDLC, X.25 permits two different modes of operation to be implemented in the system: primary/primary and primary/secondary transmission. In primary/primary operation, each of two stations connected by the data link can act either as the primary station (i.e., initiating command and control functions), or the secondary station (i.e., executing the commands of another station). In primary/secondary operation, one station is designated the "primary" station, and the other the "secondary" station. Primary/secondary operation is implemented by SDLC at the present time but not primary/primary operation.

Burroughs Data Link Control (BDLC)

Burroughs data link control (BDLC) protocol is a bit-oriented data link protocol designed for use with Burroughs equipment, including the TC 3500 intelligent terminal, the DC 140 intelligent communications controllers, and the TC 1700 and TC 5100 terminals. A communications controller, the B 776, is also utilized in the data communications networks.

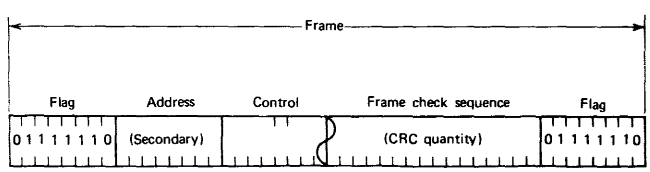


Figure 17. BDLC: frame format

The basic structure of the unit of transmission in BDLC is shown in Figure 17. This transmission frame consists of the following fields:

- Flag-bit sequence
- Address field

- Control field
- Information field
- Frame-check field
- Flag-bit field.

The flag-bit field is placed at the beginning and end of each frame in order to provide frame synchronization. The address field designates the address of the secondary station to which the message is destined. The address field is normally eight bits long but may be extended in eight-bit increments to accommodate additional secondary stations. This extended address field is the first distinction of BDLC from SDLC or X.25. (ADCCP and HDLC also allow an extended address field).

The control field is an eight-bit field that is used to transmit commands from the primary station, responses from a secondary station, and sequence numbers of transmissions. Three bits are normally used for the sequence numbers, permitting up to seven unacknowledged frames to be outstanding at any given time. The control field is, however, expandable to 16 bits, with the sequence number field being expanded to seven bits, so that the potential number of unacknowledged frames may be as high as 127.

The control field also includes a poll bit that serves to solicit a response from a secondary station.

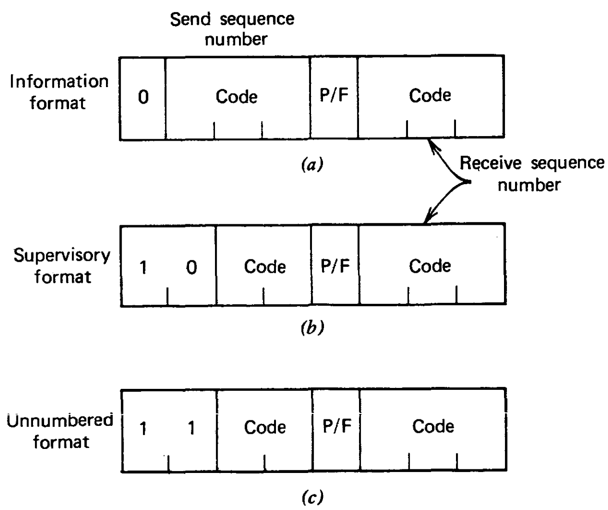


Figure 18. BDLC: control field. (a) information format; (b) supervisory format; (c) unnumbered format

The particular format for the control field is shown in greater detail in Figure 18. There are three possible formats for the control field:

- Information format
- Supervisory format
- Unnumbered format

The information format (Figure 18a) is basically used to keep track of the numbers of the last frames sent to and received by the secondary station. The information format includes the send sequence number of the frame being transmitted, and the received sequence number that is expected to be received in the next information frame coming from the addressed secondary station. The received sequence number thus informs the other station of the number of the last frame actually received. If the number is lower than the next send sequence number, then at the secondary station it is presumed that one or more frames were lost in transmission. The lost frames are then uniquely specified, and the transmitting station may retransmit the frames that have been lost.

The format identifier of the control field is specified by the left-most bit in the field. If the left-most bit is a zero, the control field has the information format. If the left most bits are one and zero, the control field is supervisory format, and if the left most bits are one and one, the control field is in the unnumbered format.

Following the format identifier bit or bits are the specific control bits used in the type of control in question.

In the information format, the format identifier is followed by three bits of send sequence number, one bit designating a poll or final bit, and three bits specifying the received sequence number.

The supervisory format of the control field, as shown in Figure 18b, consists of the two-bit format identifier, followed by a two-bit command response field, a one-bit poll or final bit, and a three-bit sequence number. The supervisory format, as the name implies, is used to transmit numbered supervisory frames which indicate a particular link or device status: situations such as a readiness to receive information, a request for retransmission of information frames, and a temporary input of the receiving capability. Such supervisory commands permit the primary station to inform a particular secondary station of the type of transfer being performed.

The unnumbered format (Figure 18c) is utilized for transmitting supervisory commands or responses for different data link control functions. The particular format of the unnumbered control field format consists of the two-bit format identifier, followed by two modifier bits, the poll or final bit, and three modifier bits in the lower-most bit position.

A Summary of Overview of Communications Network Control Architectures and Protocols

Format	Control Field Bit Encoding 1	2	3	4	5	6	7	8	Commands		Responses	
Information	0	—	N(S)	—	*	—	N(R)	—	I	— Information	I	— Information
Supervisory	1	0	0	0	*		N(R)		RR	— Receive Ready	RR	— Receive Ready
	1	0	0	1	*		N(R)		REJ	— Reject	REJ	— Reject
	1	0	1	0	*		N(R)		RNR	— Receive Not Ready	RNR	— Receive Not Ready
	1	0	1	1	*		N(R)		SREJ	— Selective Reject	SREJ	— Selective Reject
Unnumbered	1	1	0	0	*	0	0	0	UI	— Unnumbered Information	UI	— Unnumbered Information
	1	1	0	0	*	0	0	1	SNRM	— Set Normal Response Mode		
	1	1	0	0	*	0	1	0	DISC	— Disconnect	RD	— Request Disconnect
	1	1	0	0	*	1	0	0	UP	— Unnumbered Poll		
	1	1	0	0	*	1	1	0			UA	— Unnumbered Acknowledge
	1	1	0	1	*	0	0	0	USER 0	—	USER 0	
	1	1	0	1	*	0	0	1	USER 1		USER 1	
	1	1	0	1	*	0	1	0	USER 2		USER 2	
	1	1	0	1	*	0	1	1	USER 3		USER 3	
	1	1	1	0	*	0	0	0	SIM	— Set Initialization Mode	RIM	— Request Initialization Mode
	1	1	1	0	*	0	0	1	RSPR	— Response Reject	CMDR	— Command Reject
	1	1	1	1	*	0	0	0	SARM	— Set ASYNC Response Mode	DM	— Disconnect Mode
	1	1	1	1	*	0	1	0	SARME	— Set ARM Extended Mode		
	1	1	1	1	*	0	1	1	SNRME	— Set NRM Extended Mode		
	1	1	1	1	*	1	0	0	SABM	— Set ASYNC Balanced Mode		
	1	1	1	1	*	1	0	1	XID	— Exchange Identification	XID	— Exchange Identification
	1	1	1	1	*	1	1	0	SABME	— Set ABM Extended Mode		

* = P/F

Table 4. CDCCP command/response repertoire

Examples of the data link control functions specified by the format identifier are the disconnect command (DISC in Figure 15), which causes the address secondary to go off-hook; the setting of normal and asynchronous response nodes; and the acknowledgment of unnumbered frames. The normal response node command is used to set the addressed secondary to the "normal response," and thereby reset the send-and-receive sequence to zero to initiate operation in this mode. The asynchronous extended command is used for the same function while in asynchronous mode. An unnumbered acknowledgment message is used to notify the sending station of the receipt of an unnumbered frame.

Following the control field is the information field, which may be of variable length. Following the variable length information field is a 16-bit frame-check sequence used for the cyclic redundancy check. The check sequence is then followed by the eight-bit final flag.

BOLD—NCR

BOLD is the bit-oriented data link control protocol announced by NCR Corporation. BOLD is a subset of ADCCP and need not be further described at this point. Details concerning the command and response codes defined in the control field are presented in Table 5.

Control Data Communications Control Diagram (CDCCP)

Control Data has announced a comprehensive network architecture in which multivendor host computers can be connected together in the network through Local Network Processors (LNP). The Control Data Communications Control Program (CDCCP) is expected to be implemented on the 2550-series network processors, which interface with the CDC Cyber 170, Cyber 70, 3000, and 6000 series. Communications software include the CDC network operating system (NOS), together with a network access method (NAM) that permits large multihost networks to be implemented.

CDCCP is a bit-oriented protocol that is a subset of ADCCP. Since some of the key differences between CDCCP and the other bit-oriented protocols are the commands and responses defined by the control field bit encoding, Table 4 lists the CDCCP command/response repertoire.

Comparison of Bit-Oriented Data Link Protocols

One important formal means of comparison between the bit-oriented data link protocols is the definition of the bit encoding of the control field. Table 5 summarizes the command (CMD) and response (RES) repertoire of the key protocols. (The control

A Summary of Overview of Communications Network Control Architectures and Protocols

ADCCP		HDLC		SDLC		CDCCP		BOLD		BDLC		SNAP/X25	
CMD	RES	CMD	RES	CMD	RES	CMD	RES	CMD	RES	CMD	RES	CMD	RES
I	I	I	I	I	I	I	I	I	I	I	I	I	
RR	RR	RR	RR	RR	RR	RR	RR	RR	RR	RR	RR		RR
REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ	REJ		REJ
RNR	RNR	RNR	RNR	RNR	RNR	RNR	RNR	RNR	RNR	RNR	RNR		RNR
SREJ	SREJ	SREJ	SREJ			SREJ	SREJ	SREJ	SREJ	SREJ	SREJ		
UI	UI	*	*	NSI	NSI	UI	UI	NSI	NSI				
SNRM		SNRM		SNRM		SNRM		SNRM		SNRM			
DISC	RD	DISC	*	DISC		DISC	RD	DISC		DISC		DISC	
UP		*		ORP		UP		ORP					
	UA		UA		NSA		UA		NSA		UA		UA
USR (4)	USR (4)					USR (4)	USR (4)						
SIM	RIM	*	*	SIM	RIM	SIM	RIM	SIM	RIM				
FRMR	FRMR	*	CMDR		CMDR	RSPR	CMDR	RSPR	CMDR	RSPR	CMDR		CMDR
SARM	DM	SARM	*		ROL	SARM	DM	SARM	ROL	SARM		SARM	
SARME		SARME				SARME		SARME		SARME			
SNRME		SNRME				SNRME		SNRME		SNRME			
XID	XID	*	*			XID	XID						
SABM		*				SABM							
SABME		*				SABME							

* = Pending approval

Table 5. Command/response repertoire of selected standards

field bit-encoding order and command/response definitions may be found in Table 4.)

BYTE-ORIENTED PROTOCOLS

Digital Network Architecture (DNA)—DEC

Digital network architecture (DNA), the data communications architecture developed for Digital Equipment Corporation (DEC) products, utilizes three major protocols:

- Digital data communications message protocol (DDCMP)
- Network services protocol (NSP)
- Data access protocol (DAP)

DDCMP is concerned with physical link control and error recovery practices within the digital communications network. DDCMP operates utilizing existing hardware interfaces, full- or half-duplex transmissions facilities, and either synchronous-asynchronous or parallel lined circuits.

NSP is concerned with management of network functions, such as message routing between systems and processor-to-processor communication within the network.

DAP is a specialized protocol for enabling programs or service routines on one particular node of the network to utilize I/O services available on different network nodes.

These DNA protocols are arranged in a hierarchical order so that various changes may be made in one or more of the protocols at each point or node a message is received. Such an arrangement is referred to as "layered" protocols. The particular layers are:

- Dialog layer
- Logical link layer
- Physical link layer
- Hardware layer

These are illustrated in Figure 19.

The layering or envelopment of a message within two or three protocols is a key feature of DNA. By defining a hierarchical layer for each of the individual protocols,

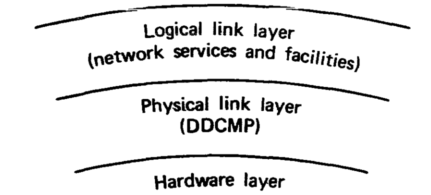


Figure 19. Digital network architecture (DNA): layered structure

A Summary of Overview of Communications Network Control Architectures and Protocols

the message may be transferred from node to node, and each node will access the particular layer that the node itself is associated with within the network hierarchy. Thus, there is a one-to-one correspondence between software hierarchy in terms of protocols or layers and a hardware hierarchy that distinguishes physical nodes or stations in the communications network.

As the message is sent from node to node, various parameters or other characteristics are added to or deleted from the message structure. The highest hierarchical layer, or dialog layer, represents the user-oriented functions, as shown in Figure 19. This layer includes user messages, programs, or data specifically coded by the user for his own use.

The next highest layer acts as a means for multiplexing various user messages into a single data stream for eventual transmission over a data link. This logical link layer therefore provides the network services and facilities required by the user in his interface with the communications network. This layer is defined by the network services protocol (NSP).

The next highest layer is concerned with physical link management—that is, the actual control of the data stream along the communications network. This physical link management layer, as typified by DDCMP, is concerned with message sequencing, synchronization, and error detection and recovery.

The last layer, referred to as the hardware interface layer, is concerned with physical hardware effects of transmission and reception of data bits over a physical link. This layer is concerned with the type of transmission mode (synchronous, asynchronous, or parallel), as well as physical device operation, including character synchronization and modem operation.

The advantage of the layered system implementation is to provide well-defined interfaces between network nodes, simple modification or replacement of layers by particular nodes, simplified error control and debugging, and consistent network integration procedures.

Figure 20 is a representation of how data flows through the layered network structure from a software or data structure viewpoint. The first block at the top of Figure 20 shows the user-created data labeled "user task data." The network services hardware and protocol then provides a routing header to the user task data log.

The physical link protocol hardware, in turn, provides a line protocol header and a block check trailer to the data block created by the network services hardware. This data block is sent over the data communications link to another node, where a corresponding physical link protocol hardware unit strips the link protocol header and the block check trailer from the

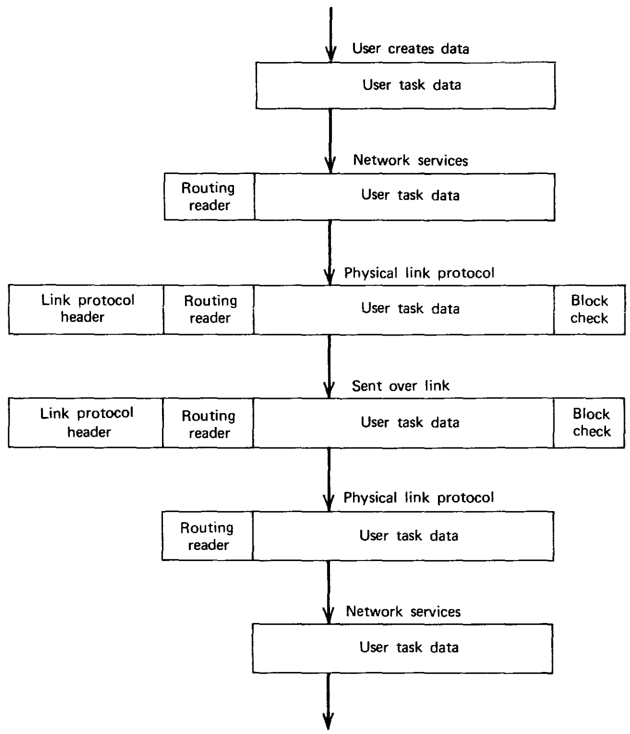


Figure 20. DNA: data flow

data block. Following the physical protocol interface hardware is the network services hardware, which strips the routing header from the data block.

Figure 21 is a simplified block diagram of the hardware used to implement these layered protocol functions in the data communications network.

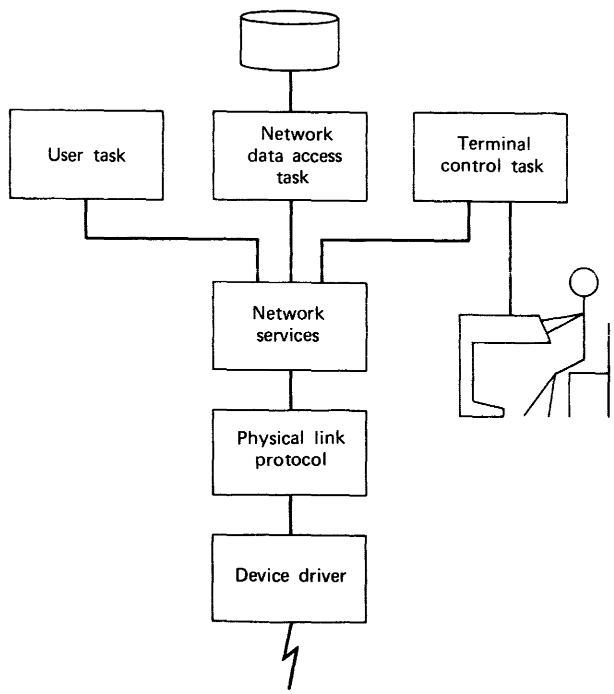


Figure 21. DNA: hardware implementation

A Summary of Overview of Communications Network Control Architectures and Protocols

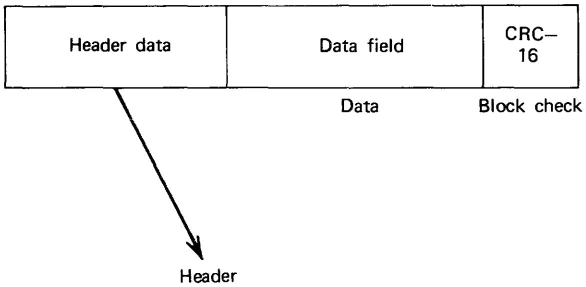


Figure 22. Digital data communications message protocol (DDCMP): frame format

DDCMP. The basic format of the DDCMP unit of transmission is shown in Figure 22. The unit of transmission consists of a header followed by a data field and a block check unit.

Figure 23 defines more explicitly the function of the particular bits in the header field. Table 6 more explicitly defines the functions of the various subfields within the header shown in Figure 23.

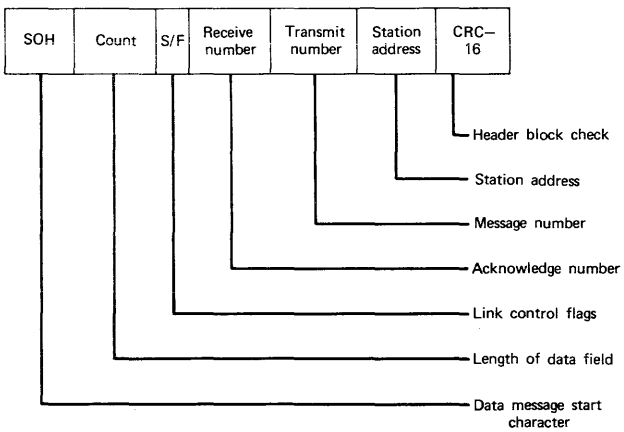


Figure 23. DDCMP: header field, data message

There are basically two types of DDCMP messages: (1) data messages and (2) control messages. The data messages transmit user information over DDCMP links between a source and a sink station.

Control information is sent by means of unnumbered control messages, or is incorporated within the headers of data messages going in the opposite direction. The particular format of the unnumbered control message block in DDCMP is shown in Figure 24. The interpretation of the various subfields in the header is in the unnumbered control message format of DDCMP as shown in Table 7.

DDCMP OPERATION. Transmission operation may be initiated by a start up (STRT) message by any one station. The message is transmitted to a receiving station, which responds to the start message by means of a start acknowledge (STACK) message. This

SOH	The numbered data message identifier.
Count	The byte count field, specifies the number of 8-bit bytes in the DATA field.
S/F flags	The link control flags used to control link ownership and permission to send. These flags are final flag: denotes end of current transmission stream; select flag: requests receiving station to transmit.
Receive number	The response number used to acknowledge correctly received messages.
Transmit number	The transmit number, which denotes the number of this data message.
Station address	The station address field used to address the destination station on multipoint links. Control stations and stations on point-to-point links use the address value 1.
CRC-16 (first)	The block check on the numbered message header.
Data field	The numbered message data field consisting of COUNT eight-bit quantities. This field is totally transparent to the protocol and has no restrictions on any bit patterns. The only requirement is that it be a multiple of eight-bits.
CRC-16 (second)	The block check on the data field. Computed on DATA only.

Table 6. Functions of the DDMCP subfields

technique of sending an inquiry and receiving an acknowledgment is known as a "handshaking" procedure. Once the sending station receives the STACK response and acknowledgment, it may begin transmitting data messages.

The data is transmitted in the form of numbered blocks or sequentially numbered messages. As the messages are received by the receiving station, the block check field is used to check for errors; if no errors

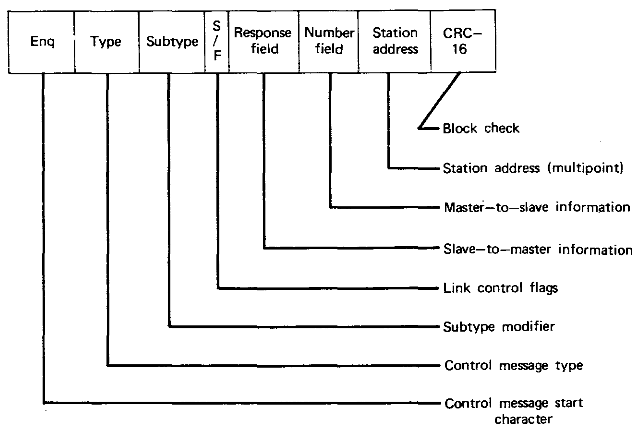


Figure 24. DDCMP: header field, control message

ENQ	The unnumbered control message identifier.
Type	The control message type.
Subtype	The SUBTYPE or TYPE MODIFIER field provides additional information for some message types.
S/F	The link control flags described for numbered messages.
Response field	The control message response field used to pass information from the slave (numbered message receiver) to the master (numbered message sender).
Number field	The control message number field used to pass information from the master to the slave.
Station address	The station address field for multipoint use.
CRC-16	The block check on the control message.

Table 7. Unnumbered control message format

A Summary of Overview of Communications Network Control Architectures and Protocols

are detected, it accepts the message as correct. The station also checks for the correct numbering and sequence of message number. If both are received correctly sequenced, the receiving station acknowledges the receipt of the messages with an "acknowledgment" (ACK) statement. A single acknowledgment statement may imply the acknowledgment of up to 225 previous message numbers. In the case of error—for example, if the block check indicates a description error—the receiving station sends a "no acknowledgment" (NAK) statement with the number of the last good message received. A number of other conditions may also generate an NAK message, such as a failure to receive the correct header.

If the primary station receives an acknowledgment, it frees the message first up to the data message number that has been acknowledged and continues to send data numbers in sequence.

It it receives a "no acknowledgment," it terminates the present transmission, freezes all messages from the message buffer through the message number not received, and retransmits the messages from the message buffer with the message following number r, preceded by a standard synchronization sequence.

If the sending station receives neither an acknowledgment, or a no-acknowledgment message from the receiving station, the sending station may send a REP message. Upon receiving the REP message, the receiving station compares the number R with the number of the last good message received. If the two numbers are equal, the receiving station sends back an ACKR message. If the numbers are not equal, it sends back a NAK message with the number of the last good message received and indicates it is making a REP response.

DDCMP may be utilized both on duplex point-to-point communications and half-duplex point-to-point systems. In a half-duplex channel, transmission from either station operates alternatively. The select and final bits are used to give up the channel, and indicate end of transmission.

NSP provides the following functions:

- Maintenance of a logical link
- Management of the logical link—physical link interface
- Error detection and recovery
- Error logging and administrative maintenance

The basic format of the NSP unit of transmission is shown in Figure 25. The unit of transmission consists of a NSP header followed by a user data field.

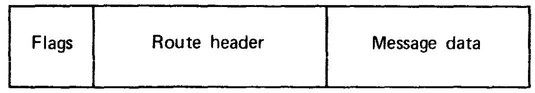


Figure 25. Network services protocol (NSP): frame format

There are also two basic types of NSP messages: (1) data messages and (2) control messages. Data messages are used to transfer dialog-level information between processes, while control messages transfer information between NSP modules.

The particular format of the NSP data message is shown in Figure 26. The function of the various subfields in the header in the NSP data message is shown in Table 8:

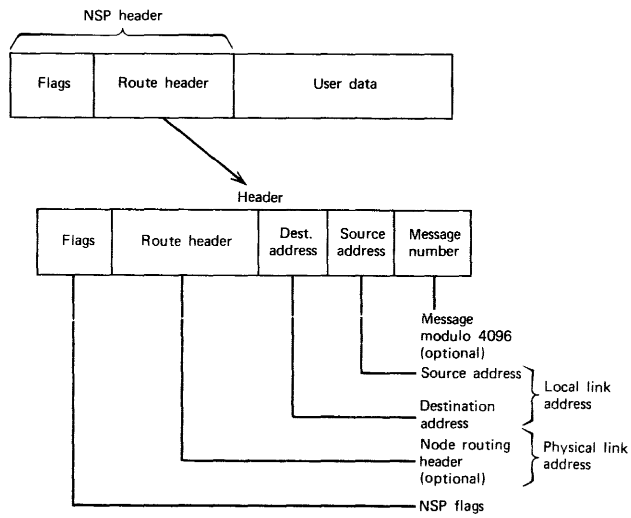


Figure 26. NSP: header field, data message

DESTINATION ADDRESS. The logical link destination address for this message. This is the address for a conversation and is usually dynamically assigned via the connection procedure.

SOURCE ADDRESS. The logical link source address for this message.

MESSAGE NUMBER. The message number incremented modulo 4096 by one for each message. This field is optional.

USER DATA. The data the user process wishes to send or receive over a logical link. This field is totally transparent and may use all eight bits of each data byte.

Table 8. Definition of NSP control message header fields

The particular format for the NSP control message is shown in Figure 27. The function of the various subfields in the header is shown in Table 9.

NSP basically performs logical link functions which find the end-points of a network connection.

The logical link operation consists of:

A Summary of Overview of Communications Network Control Architectures and Protocols

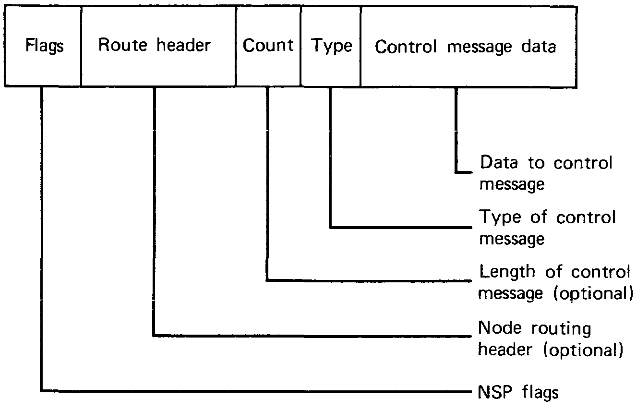


Figure 27. NSP: header field, control message

COUNT. The number of bytes in this Control Message—allows blocking of Control Messages. This field is optional.

TYPE. The number representing the type of Control Message.

CONTROL MESSAGE DATA. The data specific to each Control Message.

The following messages control logical link operation:

CONNECT. The connect message is used to establish a logical link (communication path) between processes in the network.

DISCONNECT. The disconnect message is used to destroy a logical link previously established and/or confirm that a connection rejection has been completed.

LINK STATUS. The link status message is used for message requesting.

ERROR. The error message is used to return error status to the sender of a message for a syntactical or semantic reason.

Table 9. Definition of NSP control message header fields

- Creation of logical links
- Message transfer over logical links
- Interruption mechanism over logical links
- Destruction of logical links

The creation of logical links consists of specifying two particular processes between which a logical link is desired to be created. The logical link is created by a message sent to the destination NSP including the name of the origin process with which the destination process is to be connected.

The destination process interprets this information and makes a decision whether to complete the connection depending on traffic or other conditions. If the message is desired to be completed, a connect command is sent back to the originating process, and the logical link is created. If the destination process desires to reject the connection, it issues a connect reject command that causes the logical link to be broken. Other conditions include an error detection or a disconnect message.

Message transfer over logical link is the transfer of data of common message. The same type of counting is used for message verification, so that when the sending station receives a confirm request count message, it may release storage buffers allocated to storing the messages transmitted over the link.

NSP transmission is different from DDCMP transmission in that no acknowledgment occurs on unnumbered links. The transmission is completed by giving control over to the physical link level or DDCMP on unnumbered links.

An additional feature of the message transfer over logical links is the use of instrumental request counts. Furthermore, the detection of an error in the message causes the logical links to be disconnected.

NSP has a mechanism for permitting interruption to occur on a logical link. This feature permits interruption by means of a interrupt signal for initiating data transfer.

Logical links may also be terminated by means of three occurrences:

- Request by the user
- Failure of the user process
- Failure of a communication link

If one of these occurred, the processors connected by the logical link are notified of the reason for disconnection. Messages received between the sending of a disconnect confirm are discarded.

The DAP provides the following functions in the data communications network:

- File management operations
- Input-output device operations
- Format operations
- Terminal control operations

The basic format of the DAP unit of transmission is shown in Figure 28. The unit of transmission consists of a DAP flag followed by a message operator, an optional channel number, an optional message length field, and finally the message data. Table 10 defines the functions of these various subfields more precisely.

The information field contained in the DAP message may contain various types of messages dependent upon the definition in the TYPE field. These defined fields are:

A Summary of Overview of Communications Network Control Architectures and Protocols

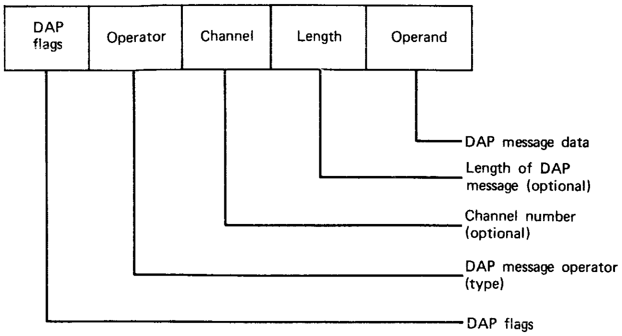


Figure 28. Data access protocol (DAP): frame format

FLAGS—DAP Characteristic FLAGS.

CHANNEL. The channel number field—optional. This is an address field used to allow up to 256 simultaneous transfers over a single NSP logical link (used primarily with terminal-type devices).

LENGTH. Length of the OPERAND field in number of eight-bit bytes—optional.

OPERAND. The information field for DAP messages. Dependent on the TYPE field. Described later.

Table 10. Definition of DAP Fields

- Data without end of record
- Data with end of record
- Status
- Continue transfer
- Control device file
- User identification
- Access
- Attributes
- Error

The basic operation of the DAP depends on the I/O structure of the system in operation. There are two levels of commands or messages in DAP. The first level is a means for setting up the connection path and access verification and handshaking. The second level is device oriented, which provides means for controlling a particular feature of the specific I/O devices.

The specific process consists of the issuing of a command to a NSP for requesting specific logical link. Such a request is specified in actual process name, depending on the particular facilities supported at the destination node. The connection is completed by the return of a clarification command, thereby establishing a link.

PARAMETER	Bi-Sync	SDLC,BDLC	ADCCP	HDLC	DDCMP
Character length (bits)	8	any	any	any	8-bit multiple
True full-duplex transmission capability	no	yes	yes	yes	yes
Control overhead	112 bits	24 bits	24 bits minimum	24 bits minimum	96 bits
Control sequence error checking	no	yes	yes	yes	yes
Allowed unacknowledged transmit frames (blocks)	2	8	8 minimum	8 minimum	256
Bit parallel capability	yes	no	no	no	yes

Table 11. Line protocol comparisons

Identification information is then sent to the node where the file or device is accessed. Various types of messages may then be interchanged between processes. These include a user identification message, an attributes message that indicates the desired mode and format of the data, an access message that indicates the desired operation, and control messages such as status messages or error messages.

Comparison of Data Link Protocols

Some of the essential characteristics of the key data link protocols are summarized in Table 11. The table highlights basic characteristics such as character length, control overhead, and the number of allowed unacknowledged transmission frames.

A more detailed comparison of data link protocols is a determination of throughput based on typical communication parameters. One measure of throughput or data communication efficiency is called the transfer rate of information bits (TRIB). TRIB is defined as:

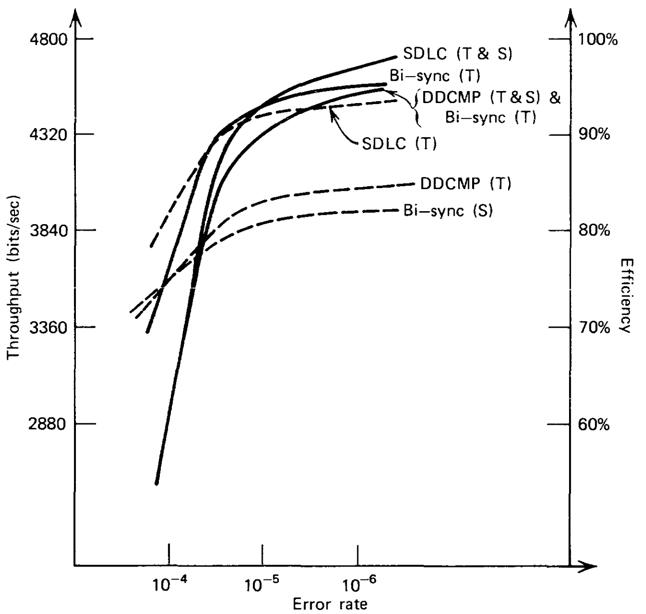
$$\text{TRIB} = \frac{K\,(M - C)\,(1 - P)}{(M/R) + T}$$

where K is the number of information bits per character, M is the total number of information characters in a message block or packet, C is the number of noninformation characters in a block or packet, P is the probability of one or more bit errors in a block or packet, R is the raw line speed in characters per second, and T is the time between blocks in seconds.

Using this formula, or similar formulas, one can then compare the relative efficiency or throughput of data link protocols and make certain conclusions.

Figure 29 is a graph of the relative throughput of bisync, SDLC, and DDCMP for terrestrial and satellite

A Summary of Overview of Communications Network Control Architectures and Protocols



Notes: 1. Solid curves are for 2000-bit frames (more suitable for satellite circuits). Dashed curves are for 500-bit frames (less suitable for satellite circuits).
2. Terms in parenthesis:
(T) terrestrial circuits; assumed 50-msec one-way delay. (S) satellite circuits; assumed 400-msec one-way delay.
3. Curves are for 4800-baud full-duplex communication channel.

Figure 29. Data communications protocol throughput comparison

circuits, using both 2000-bit and 500-bit frames. The calculated data is based upon the assumption of 4800-baud full-duplex communication channels and is plotted as a function of the error rate.

Of course, the formula and graph are not meant to demonstrate that any one data link protocol is "better" than any other, but to present a method for quantitatively comparing the features and throughput of such protocols as a function of the characteristic communication parameters of a given system.

OTHER NETWORK ARCHITECTURES AND PROTOCOLS

Other mainframe manufacturers have also announced data link protocols. Control Data Corporation has developed the Control Data Communications Control Program (CDCCP), while Sperry Rand's Univac Division has its Univac Data Link Control (UDLC).

CDCCP is implemented on the 2550-series network processing unit (NPU), which interfaces with Control Data's Cyber 170, Cyber 70, 3000, and 6000 series models. The protocol operates with the CDC network operating system (NOS), or network operating system/batch environment (NOS/BE), together with a network access method (NAM) that enables large multihost networks with many communications nodes to be implemented.

UDLC is implemented on Univac computers and intelligent terminals. An end-to-end control scheme is also being developed to perform certain functions now performed by telecommunications access methods. Furthermore, the network operating system is expected to support SDLC protocol as well.

RELATED READINGS

2740/2741—"IBM 2740/2741 Communication Terminal, Original Equipment Manufacturers Information," IBM Systems Reference Library (SRL) GA27-3002-0.

BSC—"General Information—Binary Synchronous Communications," IBM SRL GA27-3004.

SDLC—SNA—"IBM Synchronous Data Link Control, General Information," IBM SRL GA27-3093-1. "Systems Network Architecture, General Information," IBM SRL GA27-3102-0. R.A. Donnan and J.R. Kersey, "Synchronous Data Link Control: A Perspective," *IBM Systems Journal,* May 1974, pp. 140-162.
"System Network Architecture Format and Protocol Reference Manual: Architecture Logic," IBM SC30-3112-0.

ADCCP—ANSI Doc. X3.28—1976, Carlson, D.E., "ADCCP—A Computer-Oriented Data Link Control," *IEEE Compcon,* 1975, pp. 110-113.

HDLC—International Organization for Standardization: ISO/TC 97/SC 6N, "HDLC Proposed Balanced Class of Procedures" and "Proposed Enhancement to DIS 4335."

X.25/SNAP—CCITT Doc. AP VI-No. 55-E, May 1976.

BDLC—Bedford, M.J. *Data Communications 4*(6). November/December 1975, pp. 41-47.

CDCCP—J.W. Conrad, "Control Data's CDCCP," Control Data Corporation, Santa Ana, Calif.

DDCMP—Digital Equipment Corporation, "Decnet," Maynard, Mass.□

Topic Index—Section CS15

General Topic	Report Title	Report No. CS15-	See Also Report
System Component Overview	—A Brief Classification Guide to Data Communications Hardware	-050-101	
	—The Integrated Uses of Hardware in Network Configurations	-060-101	
Terminals	—The Categories and Functions of Data Terminals	-100-101	CS10-210-101; CS20-205-101; CS20-210-101; CS20-215-101; CS20-235-101; CS25-220-101; CS25-221-101; CS60-220-101; CS15-160-101
	—Typing and Printing at a Distance	-120-101	
	—Display Terminals—The Soft Copy Alternative to Teleprinters	-150-101	
	—User Programmable Terminals	-160-101	
	—Remote Batch Terminals	-170-101	
	—Integrated Point of Sale (POS) Systems	-180-201	CS60-160-101
	—Voice Response Data Terminals	-190-101	
Interface Equipment	—Modems: Terminal—Line Linking Devices	-220-101	CS10-210-101
Transmission Line Equipment	—Line Use Optimization with Multiplexing Equipment	-320-101	CS40-410-101; CS50-220-101;
Concentration and Line Management	—Communications Processors—A Special Kinds of Computer-Helping Computers	-420-101	
Telephone Equipment	—Private Telephone Systems—PBX and PABX	-610-101	CS10-610-101
	—The Equipment and Techniques of Digital Telephony	-612-101	
Transmission Facilities	—Current Transmission Facility Characteristics	-710-101	CS10-610-101; CS60-510-101;
	—Users' Views of Transmission Facilities	-711-101	

A Brief Classification Guide to Data Communications Hardware

Problem:

Anyone about to get seriously involved with planning and organizing a communications facility is immediately confronted with a bewildering alphabetic wall of RBT's, RJE's, TTY's, etc., etc., plus an equally bewildering array of unintelligible terms like modem, concentrator, multiplexer, etc., etc. This is jargon, which unfortunately is unavoidable because each jargon word is a shorthand-like compression of perhaps thousands of words that would otherwise be needed to express ideas between specialists . . . very handy for those who know the concepts that underlie the jargon but practically impenetrable for an unknowledgeable outsider. There is no easy solution to this problem, especially in the communications discipline, because you must understand the underlying concepts of communications technology before you can even begin to analyze, plan, buy, assemble, and manage a communications system.

All of Section CS10 (which you have read, of course) is basically tutorial and deals more with the fundamental ideas and problems of communications rather than with specific hardware/software implementations. Now you can safely enter the realm of real hardware/software and begin to learn some of the nitty-gritty details of transmission facilities to gain the tools you will need for analysis, planning, and acquisition. This report introduces the hardware segment of this section. The hardware coverage is restricted to data communications equipment. Voice and video equipment is discussed elsewhere in this service.

The author of this report neatly divides data communications hardware into five partitions beginning with, logically enough, the man-machine interface to the terminal. Data composed at the terminal is converted into transmittable form by another class of devices called modulators/demodulators (modems). The transmission channel is the next in line, and it is complex enough to warrant separate coverage in the CS15-700 segment of this section. The final two equipment classes deal with ways to gain maximum use of a transmission channel by techniques such as multiplexing and concentrating and with computer-like equipment that controls all the other hardware.

A Brief Classification Guide to Data Communications Hardware

Solution:

All communications data processing systems exhibit a very similar architecture with respect to their associated communications subsystems. The communications subsystem of an automatic data processing system consists of the functional devices that permit the reliable exchange of data between remote locations and a central processor or computer. While each specific system has a unique configuration of these functional building blocks, their existence is common to all such information systems.

The components of a generalized communications subsystem are as follows:

1. Data communications terminals.
2. Modulating-demodulating units.
3. Transmission channels.
4. Network multiplexers.
5. Computer communications controllers.

Of these five major component classifications, only the network multiplexers may be optionally applied, depending upon the relative economics of the particular system. The other four classifications are mandatory for a computer communications system. This report introduces and discusses the role of each of these components in today's information system applications, so that the reader can obtain an accurate overview of the interrelationships of the components and can assess the system economic factors with respect to each component's scope of performance.

COMMUNICATIONS TERMINAL SYSTEMS

Terminal devices are used in all data communications systems. Even when two computers exchange data through a communications channel, one computer is considered a terminal to the other computer. And even the normal I/O peripheral devices associated with a computer can be properly considered a level of terminal devices. Following is a classification of the various terminal devices that can be used to satisfy a variety of applications:

LEVEL 5: Applications-Sensitive Terminals
LEVEL 4: Conversational and Interactive Terminals
LEVEL 3: Batch Transmission Terminals
LEVEL 2: Satellite Processor Terminals
LEVEL 1: Central Computer I/O Peripherals (Central computer and files)

Any application may be satisfied by terminals at any of the specified levels. The single factor that is normally influenced by a replacement of specific terminal level by another terminal level is the responsiveness or timeliness of the system's resulting data. The relative factors illustrated in Figure 1 compare some of the other more pertinent aspects of each terminal-level classification in a generalized application.

Other schemes of classifying data terminal devices relate to the inherent transmission speed of the terminal. Figure 2 presents a terminal speed spectrum with illustrative candidates.

A push-button telephone instrument can be used as a terminal device. The pushbuttons on the instrument each generate discrete tones. These tones can be decoded into binary-encoded signals at the computer end. Response, however, must be oral. Several available voice-response devices can be programmed to generate a number of spoken words over the return path to the Touch-Tone telephone terminal user. (See Report CS15-190-101).

Probably the most common terminal in use today is the typewriterlike teletypewriter. These devices are characterized by a hard-copy printing mechanism and a keyboard. In some configurations the keyboard can be left off for receive-only applications. The operating rate of these devices is on the order of 10 to 15 characters per second (cps).

Other typewriterlike devices on the market operate at 30 to 40 cps. These are more correctly referred to as conversational terminals. Such typewriterlike terminals have also led to the development of much faster terminal devices that still have the same basic characteristics. Several manufacturers now offer printing devices (with or without keyboards) with speeds approaching 500 cps. These are usually called serial or character printers. The basic printing mechanism is like a typewriter in that each character is printed individually, and an entire line is constructed in a character-by-character manner from the left margin. This mode of operation is chosen to eliminate the need for some local storage at the printer site, since a character may be printed before the next character is received.

A Brief Classification Guide to Data Communications Hardware

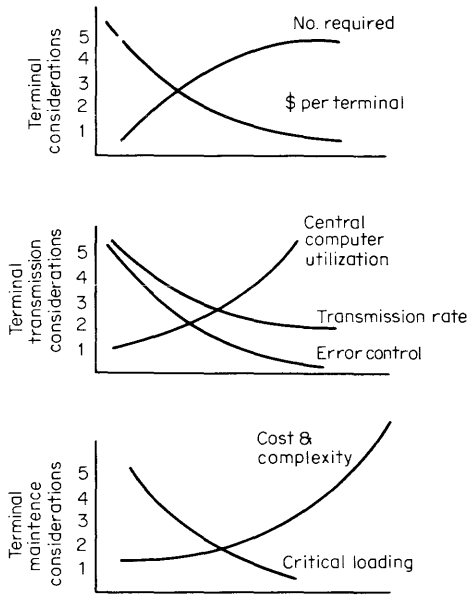


Figure 1. Relative-level relationships in terminal application classifications

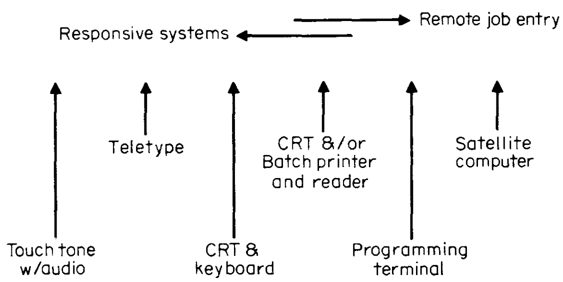


Figure 2. Terminal speed spectrum

Higher-speed printing mechanisms, identical to those used as computer peripherals, are also used as terminal devices. The entire line is printed in a single operation. Since characters are received through the communications medium in serial order, some local storage is necessary to assemble a print line. These printers are called line printers.

Datapro Comment:

The preceding class of printing mechanisms is called impact printers. The class also includes the relatively new daisywheel printer, in which the typing elements are placed on the ends of spokes connected to a central hub. The hub is twisted radially to position the elements for printing. Nonimpact methods are also becoming more popular because of their potentially higher speeds. Two frequently encountered methods are matrix printing and ink-jet printing. In matrix printing, characters are formed by patterns of closely spaced dots. The dot pattern is developed by selectively pushing tiny rods against sensitized paper through a hairbrushlike mechanism. In ink-jet printing, an ink spray is electrostatically deflected to form the shape of a character. The technique is practically the same as the method used to produce a picture on a television screen. The major disadvantage of nonimpact printing methods is that simultaneous copies cannot be made as they can with impact methods.

Wherever hard-copy output is not needed, CRT terminal devices can be used. The CRTs used in terminals have the same basic characteristics as those used in commercial television sets, and the associated electronics convert the digital bit sequences into signals acceptable to the display. These devices usually include a keyboard for data transmission.

Display devices may be of two kinds: limited and vector. Limited display systems include the ability to exhibit a limited set of symbols in specific locations on the face of the screen. This type is most often used in information-query types of commercial applications. A vector display permits arbitrarily drawn straight lines. By drawing short lines at various points, figures such as circles, letters, numeric digits, and other symbols can also be constructed.

The use of a high-speed line printer in a terminal provides for batch data output. If the terminal also includes a card reader, the terminal may be used to submit jobs to a central computer through the card reader and to receive output on the printer. In many applications, a keyboard is desirable. The keyboard is often associated with a display, so that the user can request certain limited information directly from the system. A terminal with four components (printer, reader, CRT, and keyboard) can be used for communicating with both batch-oriented and interactive systems.

Some terminals include a programmable minicomputer that can perform a limited amount of processing at the user's site. These satellite processor terminals are available to permit local data acquisition with local data editing performed by the minicomputer to edit and store the information. At the same time, the minicomputer is also used as a terminal to some larger system. The advantage of a programmable terminal is that its characteristics can be changed to emulate different terminal devices. If the terminal is used to communicate with more than one computer system, the programmable computer can be used to create the data transmission characteristics required by each individual computer system.

A Brief Classification Guide to Data Communications Hardware

Output Media

The choice of the terminal output media is an important factor in system design. The media selection must be made on the basis of the intended applications of the system, not just the availability of terminal devices. Some of the most common media are:

- Printed page.
- Punched card.
- Punched paper tape.
- Magnetic tape and disc.
- Visual display.

Printed pages may be typed by one of the various hard-copy mechanisms and are probably the most common output form of a terminal communications system. The volume and variety can dictate the use of either a low-speed typewriterlike character printer or a high-speed line printer.

Punched cards represent a very common data medium for high-volume applications. These can be prepared off-line, and provide a physical, tangible document from which to work. Punched cards are also used as an output medium in data communications, depending on the specific application. Their major use, however, is to input data.

Punched paper tape is a very common data communications medium. Many terminals are designed to handle only paper tape. Paper tape is often produced as a byproduct from accounting machines, and can be used as a direct data entry medium. Paper-tape speeds are also compatible with punched cards, with many of the same physical problems. Paper tape demands that some transcribing equipment be available if output is provided only on tape.

Punched cards and paper tape are now being replaced by magnetic tape as a local terminal storage medium. There are many terminal devices which use small audio cassettes, with relatively low recording densities. Magnetic tape and disc offer the potential for higher transmission speeds, but, the speeds of contemporary magnetic facilities at a terminal are normally designed to be compatible with the other media of the terminal.

Most visual display devices need to refresh the display several times a second. The refresh function is most often done by requiring the terminal to have some local, recyclable storage for refreshing the local display.

Hard-Wired vs. Stored-Program Terminals

When considering the application characteristics of the terminal, one may be faced with the question of selecting either a hard-wired or a stored-program device. Hard-wired terminals are generally satisfactory with low-speed, keyboard, or hard-copy applications. Magnetic-tape, punched-card, and punched-paper-tape terminals, however, are most often used where the output media will be direct input to a central processor. Because of normal program maintenance, it often becomes more desirable to make any required changes in the terminal program, rather than in the main processor program. Under these circumstances, a stored-program data terminal becomes more useful and more economical than a hard-wired terminal. It is also possible to utilize a stored-program terminal for off-line listing compilations, sorting, etc., by adding low-cost peripherals, if the traffic flow for communications permits this kind of scheduling.

Hard-wired terminals have traditionally had the advantage of lower cost capacity and in most cases, have required less maintenance. These factors are rapidly changing, however, with current technological advances, resulting in more reliable and lower-cost circuitry, particularly in the MOS/LSI semiconductor areas.

Datapro Comment:

The preceding description may sound a bit antiquated to readers who are familiar with the current buzzwords of super intelligent terminals, distributed processing, and the like, but the hard reality is that there is always a distinct and usually yawning gap between where the technology is going and where the user majority is. The most prevalent terminals are still the slow, venerable Teletypewriter Models 33/35 because they are still well suited to the processing tempo of older computer installations (there are a remarkably high number of IBM 1401's and 650's out there yet) and to the limited availability/high cost of highspeed communications links. True, microprocessors are becoming so inexpensive that many modern microprocessor-inclusive terminal offerings (intelligent and genius terminals) are practically indistinguishable from minicomputers, but few users have really exploited them extensively, and even fewer users have tied them together into the visionary distributed processing networks promised by cheap, wideband communications.

The important point is that the communications system planner is in the odd position of having one foot in yesterday and one foot in tomorrow. The old, slow teletypewriters can't be tossed on the scrap

A Brief Classification Guide to Data Communications Hardware

heap because they really work quite well for many mundane applications that really don't need super intelligence and highspeed communications. Yet the newer innovations can't be ignored because they presage a whole new phase of information processing that must be planned for now. The basic lesson is: Don't be afraid to hold onto a paper-tape device (for example) if it is still cost justified even though it may be a technological anachronism, but analyze its use carefully to make sure that it won't turn into a costly bottleneck in the system you plan to have running three or four years from now.

Error Control

The kind and type of error-control technique required will influence the selection of the data terminal. If the data traffic contains sufficient redundancy, relatively simple error detection with operator-initiated correction provisions may be acceptable. This function is carried out by either the processor or the human brain. If no redundancy or very little redundancy is present in the data traffic flow, or if the data is to be directly interpreted by a computer, it is important to keep the residual error rate very low and an elaborate error-control scheme will be required.

The choice of the terminal will be influenced by the complexity of the error control required. The more sophisticated codes generally require a stored-program terminal for execution. Character and block parity check schemes can be readily handled by hardwired machines.

Hidden Costs

Some hidden costs that may be overlooked but that must be examined when choosing data terminals for a system are personnel costs related to data preparation, data distribution, training and staffing, etc. The costs for parallel operations while switching over to a new data communications system must be calculated. Other charges,, such as the supplies for magnetic tapes, punched cards, documentation, etc., are sometimes significant.

MODULATING-DEMODULATING (MODEM) UNITS

The communication network used for voice and data transmission is an analog system. Voice generates an analog signal made up of a number of frequencies capable of having a number of levels. Data generated by or required by a digital computer are binary in nature, and the information signal within such a computer has only two states or levels: 1 or 0 value. In addition, the computer communicates between its internal components using direct current (dc) signals rather than frequencies. In order for a standard dc binary signal from a digital computer to function with a frequency multilevel signal communication network, a converter must be utilized. This converter is called the modem. The modem creates a signal compatible with the communication network from the dc binary signals. Conversely, the modem transforms a frequency signal from the communication network into a dc signal compatible with the computer.

The term modem is a contraction of the two functions to be performed, i.e., modulating and demodulating. Modulating means to convert the dc signal into a compatible frequency signal. Demodulating means to convert the frequency signal into a dc signal. The term data set is also used to identify a modem.

Modulation

There are three basic types of modem designs, classified with respect to their associated methods of modulating (and demodulating):

- Amplitude modulation (AM)
- Frequency modulation (FM)
- Phase modulation (PM)

An AM modem transmits a constant frequency. Amplitude, however, varies depending upon the value of the dc binary information. For example, a dc 1 level can be full amplitude, and a dc 0 level can be a lower amplitude.

Amplitude modulation is generally used for low-speed data transmission (up to 300 bps). While an AM modem is the least complex to design and manufacture, this technique tends to be susceptible to burst noise, which varies the amplitude or energy level of the transmission signal. Burst noise coincident with a 0 level can raise the amplitude so that the demodulator interprets the received signal as a 1 level.

An FM modem transmits a constant level or amplitude, and the frequency is changed according to the value of the dc binary information. For example, a dc 1 level can be frequency 1, and a dc 0 level can be frequency 2.

Frequency modulation is generally used for medium-speed data transmission (300 to 1,800 bps). While more complex to manufacture than an AM modem, the FM modem has a higher immunity to burst noise than the AM modem.

PM modems transmit a constant amplitude and frequency, but the phase according to the value of the

A Brief Classification Guide to Data Communications Hardware

dc binary information. For example, a dc 1 level can be phase 0 degree, and a dc 0 level can be phase 180 degrees.

Phase modulation is generally used for high-speed data transmission (2,000 bps and higher). The design and manufacture of PM modems is quite complex. However, since it uses constant amplitude and frequency, the PM modem has excellent noise-immunity characteristics. Phase modulation coupled with other data manipulation techniques can provide reliable data transmission at speeds of 9,600 bps over a typical voice circuit.

Some modems in operation today use AM for high-speed transmission (AM "vestigal sideband"); and some low-speed modems (under 300 bps) use FM rather than the AM techniques.

At the higher transmission speeds, circuit conditioning is mandatory. In addition to the circuit conditioning provided by the common carrier (C-1, C-2, C-4), some high-speed modems also contain equalizers (conditioners) as part of their design. The equalizers contained in these modems are intended to complement, not replace, conditioning provided by common carriers. These modem equalizers essentially accomplish the fine tuning required for the high transmission speeds.

Recent high-speed modems have automatic equalizers, an advancement which permits more practical application. With automatic equalization, the receiving modem can sample the delay variations present in a received signal and automatically introduce delay at the appropriate frequencies to ensure that all frequencies are received at the same time as the slowest frequency.

Timing

Modems can also be classified with respect to method of timing: asynchronous and synchronous.

An asynchronous modem reacts to any change occurring at its dc interface regardless of the rate of change. Realistically, if the rate of change exceeds the internal limitations of the modem, such as the bandpass of the output line filters, the modem will not operate properly.

An asynchronous modem designated as having a maximum 300-bps transmission rate will operate very reliably at only 100 bps or at any other speed within its stated limits. A synchronous modem, however, operates only at specified speeds. These modems also require a clock or timing signal, either generated internally by the modem or provided by the data processing equipment. This clock is used to sample the dc signal and activate the modulator. For received data, the clock identifies to the data processing equipment when a new bit has been demodulated. Asynchronous modems are generally used at up to 1,800 bps, while synchronous modems are generally associated with the higher transmission speeds.

In a low-speed data transmission application, the digital bit-per-second and transmitted baud rates may have the same value. For each unit of information on the digital side of the modem, there is a corresponding unit on the analog side. Such a system might utilize a modem device that is essentially passive and that generates a unit of analog information for each unit of digital information presented to the digital interface.

Multilevel Modulation

With higher-speed systems, modems perform functions other than simple modulation and demodulation. Because of the limitations of available communications channel bandwidths, or passbands, the digital data must be compressed so as to reduce the number of actual units of information transmitted. This is accomplished by transmitting to the communications channel a single unit of information representative of a number of digital bits of information generated at the interface between the data processing equipment and the modem. Such data transmission systems are referred to as multilevel modulation systems and are most common at the higher transmission rates.

If a transmitted unit of information can be used to recover more than one digital bit, the number of digital bits that can be transmitted in a given period of time may be greater than the transmission rate (baud) of the communications channel. For example, a system using amplitude modulation may employ two discrete signal levels to represent the different binary values. The maximum number of bits that can be transmitted in one second is equal to the baud rate of the communications channel. Assume that an amplitude modulation system actually had four discrete amplitude levels that could be transmitted and deleted. Two digital bits would be represented for each signal unit transmitted. In such a system the baud rate would be half the bit rate:

Transmitted signal level	Digital bit values
A	11
B	10
C	01
D	00

Other modulation schemes can also use this technique. By using four different frequencies in a frequency

A Brief Classification Guide to Data Communications Hardware

modulation system or four distinct phases (045, 135, 225, and 315 degrees) in a phase modulation system, the same type of bit rate compression can be achieved. This concept has been carried further with eight discrete steps where three consecutive bits are transmitted in a single unit of transmitted information. With these techniques, commercially available modems are able to transmit 4,800, 7,200, and 9,600 bps on a communications channel with less than 3 kHz effective bandwidth.

TRANSMISSION CHANNELS

Precise uniformity of transmission facilities is not a characteristic of today's communication network. Transmission facilities can be classified into three levels. The levels are based on the transmission capacity of the service as expressed in bandwidth. The wider, or larger, the bandwidth, the greater the potential capacity for data transmission. Table 1 presents these service levels and their associated characteristics. While all services can be used for data transmission, the voiceband is the most widely utilized at present. The transmission parameters of the voiceband are also applicable to the other two service categories. However, in view of the predominant use of voicebands, this discussion is limited to the voiceband characteristics.

As shown in Table 1 a typical voiceband has a nominal 4-kHz bandwidth. This means that when voicebands are frequency-multiplexed for long-distance carrier transmission, the centers of two adjacent voicebands are spaced 4,000 Hz apart. The actual bandwidth available for data transmission is significantly less. The actual bandwidth is identified as the circuit's effective bandwidth. Such transmission equipment as bandpass filters, inductive loads, amplifiers, etc. cause the difference between the nominal bandwidth and the effective bandwidth.

Designation	Bandwidth	Transmission rate	Remarks
Narrowband	Variable—generally up to 300 Hz	150—300 bps	Generally private line except for Telex and TWX services
Voiceband	Nominal 4 kHz	2,000—2,400 bps and higher with COAM modems	Private and dial lines
Wideband	48 kHz and up	40.8 kb per sec and higher	Generally private line except Data Phone 50 service from Bell System

Table 1. Transmission facility classifications

Voiceband	Conditioning	Delay distortion	Band
3002	None	1,750 μs	800—2,600 Hz
3002	C-1	1,000 μs	1,000—2,400 Hz
		1,750 μs	800—2,600 Hz
3002	C-2	500 μs	1,000—2,600 Hz
		1,500 μs	600—2,600 Hz
		3,000 μs	500—2,800 Hz
3002	C-4	300 μs	1,000—2,600 Hz
		500 μs	800—2,800 Hz
		1,500 μs	600—3,000 Hz

Table 2. Conditioning

When a modem transmits data, a series of frequencies is simultaneously generated. Generally, these frequencies can be considered the carrier frequency and the associated sidebands. In order for the receiving modem to detect the transmitted frequencies for either a 1 or 0 bit, the carrier frequency and a number of the sideband frequencies must be received within the duration of the bit. Electrical signals have different propagation times depending on their frequencies within a typical communication circuit. If the transmission rate is low (up to 1,800 bps) the bit duration is usually long enough to allow all necessary frequencies to appear at the receiving modem, permitting recovery of the bit. As the transmission rate increases (2,400 to 4,800 bps and higher) the transmission circuit must be treated to ensure that all frequencies are received at the required time. This is accomplished by slowing down the propagation of the faster frequencies to match the slower frequencies. This treatment is called circuit equalization or circuit conditioning. Table 2 presents typical conditioning available for a 3002-type voiceband data channel.

Transmission facilities are available on both a private line or dial line basis. A private line is a full-time circuit between two specified locations. The circuit charge is based on the airline mileage equivalent between the two locations. A dial line is established when needed and dissolved when not needed. The charge is based on the airline mileage equivalent and the time the connection was established.

Both of these services may be used for data. Since the dial line is established in a random manner, there is no way to predict the resulting connection quality. A dial connection may exhibit excellent transmission characteristics at one time and totally inoperative data characteristics the next time the same location is dialed. Because of this variability of characteristics, a dial line normally has a lower transmission rate capability than a private line. Conditioning, for example, is available only with a private line, since it is a permanent connection.

A Brief Classification Guide to Data Communications Hardware

Voiceband circuits are generally available on a half-duplex or full-duplex basis. Half-duplex circuits (2-wire) permit two-way, nonsimultaneous data transmission. Full-duplex circuits (4-wire) permit two-way, simultaneous data transmission. All the circuits between the telephone companies' switching offices are full-duplex (4-wire) lines. These interoffice circuits are called trunk facilities. The circuit between the telephone company end office and the customer is a pair of wires normally capable of transmitting one voiceband. These circuits are called local loops. In order to conserve on physical wire, and in view of the original voice transmission requirements, the local loops are half-duplex (2-wire).

When a private line is installed as full-duplex (4-wire), an additional local loop is assigned between the customer's location and the trunk facility. The switching centers, originally designed for voice, can accept only 2-wire local loops. A dial connection is, therefore, only half-duplex.

Since a dial connection is randomly established, a mismatch of trunk circuit characteristics normally occurs. This mismatch can cause the resulting connection to oscillate or echo, because of the reflection of the transmitted signal. In order to prevent this echo, which could destroy the quality of voice transmission, echo suppressors are installed in the telephone trunk circuits. These echo suppressors amplify in the direction aiding the higher-level signal and attenuate the lower-level reflected signals from the opposite direction.

When the other end of a connection begins to transmit, the higher-level signal will be coming from the opposite direction. The echo suppressor will then reverse its direction of amplification, the reversal being called the turnaround of the connection. This function takes a certain amount of time (turnaround time). Voice transmission can usually continue during this time without loss of intelligibility. For data transmission, however, this turnaround time is dead time. Information cannot be transmitted until the turnaround is complete. Turnaround time is normally stated as 150 ms, which can significantly reduce the effective throughput of a data transmission.

NETWORK MULTIPLEXERS

The nature and design of telephone carrier systems allows a number of independent simultaneous transmissions over the same physical trunk circuit. The same technique is utilized in digital multiplexer equipment.

The requirement that a separate physical circuit is needed to transmit an information signal can easily become an economical liability. Digital multiplexers allow a number of independently generated digital signals to be combined and transmitted on a single physical circuit. The system network economies that are possible through the use of multiplexing are the primary consideration for the use of this technique; that is, the basic justification for the use of digital multiplexing equipment is a reduction in system communications cost. Fundamentally, there are two types of digital multiplexers: frequency division multiplexers (FDM) and time division multiplexers (TDM). Each has applications in today's telecommunication systems networks.

A typical FDM configuration consists of a number of small-bandwidth circuits that are combined and transmitted simultaneously over a wider bandwidth circuit. The wider bandwidth circuit, also known as the trunk circuit, is frequency divided into a number of smaller bandwidth channels.

The primary disadvantage of frequency division multiplexing is the limited number of low-speed circuits that can be combined into a high-speed trunk. If the low-bandwidth circuits are to be used at a lower transmission rate, i.e., 75 bps, additional lowspeed channels could possibly be realized, since each low-speed channel could be filtered into channels narrower than 300 Hz. The electronic circuitry is extremely simple: to a large extent, the majority of the electronics consists of passive analog circuits. The only active element is associated with the amplification circuits.

The cost per low-speed channel for an FDM is generally lower than for a comparable TDM. This cost advantage is primarily evident when only a small number of low-speed channels are to be multiplexed. Once the number of low-speed channels to be multiplexed increases, time division multiplexing becomes more attractive economically. This particular economic crossover varies with each manufacturer's equipment, because of variances in pricing and system modularity. Whenever a number of low-speed circuits are multiplexed using frequency division, they must be demultiplexed before the low-speed data signals can be introduced to a digital computer. In evaluating computer communications system cost, operational duplication of FDMs must be considered.

Another advantage of frequency division multiplexing is the lack of required synchronization. Each low-speed channel before and after the multiplexing function operates as an independent transmission circuit. As such, the timing technique (asynchronous or synchronous) is indicated solely by the termination equipment or each low-speed channel. Between the FDMs there is no timing or synchronizing activity required. The multiplexers are passive and transparent transmission elements with respect to the information and data transmission activity.

A Brief Classification Guide to Data Communications Hardware

A time division multiplexer, unlike an FDM, utilizes the full spectrum of the trunk circuit. A typical configuration for a TDM system also consists of a number of low-speed circuits being combined for transmission over a high-speed trunk. The combining, however, is achieved by an allocation of time rather than an allocation of space or frequency spectrum.

If a bit of information requires X ms to be transmitted on a low-speed channel, and the high-speed channel can transmit this same bit of information in one-twentieth of the time (X/20 ms), the capacity of the TDM is theoretically 20 low-speed channels. For example, assume 75-bps low-speed circuits to be multiplexed using time division and a trunk circuit transmission rate of 2,400 bps. The theoretical capacity is 32 such low-speed circuits.

The determination of theoretical multiplexer capacity using frequency division was achieved by dividing the bandwidth of the low-speed channels into the bandwidth of the trunk circuit. With time division, the theoretical capacity of the multiplexer system can be calculated by dividing the band rate of a low-speed channel into the band rate to be used on the high-speed trunk circuit.

The most important consideration, however, in the design of a TDM is the synchronization necessary over the trunk circuit. The TDM operates on a cyclical basis, such that a cycle must be completed within the transmission time of a unit of information on the low-speed circuits. A complete cycle or scan of all low-speed circuits is called a frame. Each frame transmitted over the trunk circuit is divided into a number of bytes, where a byte is a unit of information that can consist of a data bit or a data character depending on the multiplexing equipment.

If the TDM is designed to multiplex data bits of information, one bit from each low-speed circuit would be scanned and transmitted in the sequence. More commonly, a complete data character is multiplexed and would, therefore, be transmitted over the trunk circuit in sequential bytes. In both cases, each frame must contain unique synchronizing information which is used to synchronize the demultiplexing equipment with the multiplexing equipment. A unit of information obtained from low-speed Circuit No. 1 by the multiplexer must be properly delivered to low-speed Circuit No. 1 by the demultiplexer. This order is assured by designating one or more bytes of the frame as synchronizing bytes, containing unique characters which cannot be duplicated by data. The byte following the synchronizing bytes is always assigned to the first low-speed circuit followed by the byte from the second low-speed circuit, etc.

The application of digital multiplexers is primarily of value when a system uses a number of low-speed circuits (less than 300 bps) and voice-grade trunk circuits. The other requirement for the use of multiplexing equipment is that each low-speed circuit, and hence each device, required continuous access to a central point. If the access to the central point can be scheduled, the use of the switched or dial communications generally offers a more economical system solution to the communication network requirement.

The major advantage of a TDM is the high multiplexing capacity. While a significant amount of common electronics is used by all the low-speed circuits, present TDMs generally represent a level of reliability comparable to FDMs. It is also interesting to note that if the data to be time-division-multiplexed is destined for a digital computer, it is not mandatory that the high-speed trunk be demultiplexed before interfacing with the digital computer. It is possible to interface the high-speed trunk directly to the digital computer and allow the computer to synchronize and demultiplex the data stream.

COMPUTER COMMUNICATIONS CONTROLLERS

The communications controller is a peripheral to the central computer, and generally interfaces with the modems or data sets through a separate data set adapter (also called a line termination unit) for each modem. For each modem there is a data set adapter (DSA) that performs the interface functions of matching the computer's internal conventions to those of the data set supplied by a communications vendor. Most data transmission is performed serial by bit, while digital computers deal with bit clusters called bytes or characters.

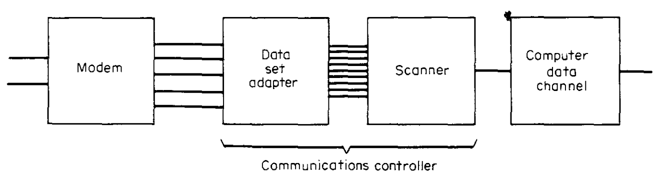


Figure 3. Communications controller configuration

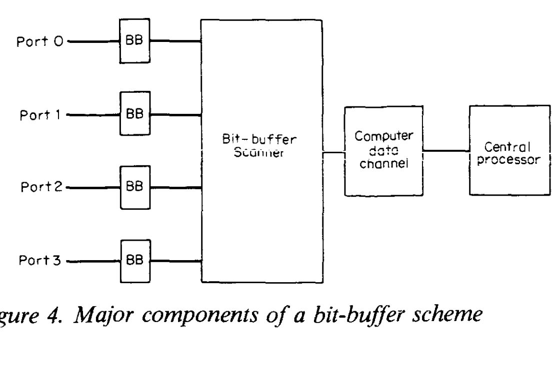


Figure 4. Major components of a bit-buffer scheme

A Brief Classification Guide to Data Communications Hardware

The data set adapter must contain the logic necessary to commutate the character's bits one by one for transmission and to assemble bits into characters on input for parallel presentation to the computer. It is important to note that not all DSAs do character-bit manipulation. Most data set adapters do, however, carry certain control lines from the data set into the computer (see Figure 3) to detect and handle various nondata control functions.

The simplest kind of I/O communication scanning transfers bits, as received, into the computer on demand. The bits are not normally deposited directly into main memory but are deposited into a special register in the central processor and are made accessible to a program in memory. This memory implies that the data set adapter must have a one-bit memory, and for that reason is called a bit buffer. Figure 4 shows the major components of a bit-buffer scheme. Each individual modem and associated transmission channel is referred to as a port when viewed from within the computer.

Since bits arrive through different ports of the scanner asynchronously, some control outside the scanner itself must be responsible for timing in this kind of system. Each port is equipped with a bit buffer that senses and holds the current logical state of the communications channel. During one bit interval, the multiplexer scans each active port and presents the current logical state of each port to the central processor.

The scanner that controls the bit buffers typically works as follows for data reception. Assume that only port number 2 is active, and the maximum bit rate is 150 per second. Each 6-2/3 ms, an incoming bit obliterates the last bit. Each 6-2/3 ms the computer is interrupted by the scanner with the address of port 2. The current state of the bit buffer for that port is presented to the central processor and is made available to the programs designed to do communications servicing.

Each time the scanner senses the state of port 2, a logical 1 is detected, since no data is being sent from the terminal, and the communications channel is in the quiescent state. When a key on the terminal is depressed, the next port sample yields a logial 0—the start bit. The next eight bits are then collected and transferred as data.

Each port of the system has two unique words in central memory. Once the start bit is detected, the character accumulation register (CAR) is cleared to zero, and the bit counter is set to eight as shown in Figure 5. After one bit interval has elapsed since the start bit (6-2/3 ms in our example), the received bit is shifted into character accumulation register,

Time since start bit (ms)	Bit received	Car	Bit counter
			8
$6\frac{2}{3}$	1	00000001	7
$13\frac{1}{3}$	0	00000010	6
20	0	00000100	5
$26\frac{2}{3}$	1	00001001	4
$33\frac{1}{3}$	0	00010010	3
40	1	00100101	2
$46\frac{2}{3}$	1	01001011	1
$53\frac{1}{3}$	0	10010110	0
60	Stop bit	10010110	0

Figure 5. Scanner operation

and the bit counter is decremented by one. Bits are placed into the character accumulation register by shifting the previous contents left by one bit position and by placing the newly received bit in the low-order (vacated) position of the register. This corresponds to the standard convention of transmitting the most significant bit first.

After all bits have been collected, the bit counter eventually reaches zero. Then, the software that accumulates the bits into characters recognizes that subsequent one-bits represent the quiescent state of the communications link, and a zero in the bit counter means a new start-bit is to be expected. This may happen as soon as one bit interval (6-2/3 ms) has elapsed after the CAR has been filled. Therefore, the communications software must unload the character accumulation register as soon as possible. This is accomplished, in the software, by passing the contents of the accumulation register to a buffer elsewhere in central memory reserved for this port. The buffer is a consecutive group of memory locations that hold the successively received characters.

Assuming a 30-port system, with all ports active, all terminals will be simultaneously transmitting or receiving. The scanner that controls bit buffers will generate heavy demands on the software system. The bit-sensing operation of the scanner must be performed 150 times each second for each of 30 ports, or 4,500 times each second, whether data is being received or not. To be able to continue to transfer logical bits of data, a computer interrupt must be performed 4,500 times each second. If each interrupt requires an average of 50 us, 450 ms per second are used up, or 45 per cent of the processor's capacity for computing is usurped by the communications bit servicing.

Under the worst conditions, each terminal could supply 15 characters each second, and 450 characters must be placed into buffers each second. Normally, output and input are not performed concurrently in half-duplex operation. If the software requires 100 us to install one received character or to select one character to be transmitted, another 45 ms are required.

Finally, the buffer-switching operation will occur about every 30 characters, which is the approximate length of the average record received from and transmitted to a terminal. Then, the buffer switching must take place 15 times each second. This operation might take an average of 250 μs each, requiring a total of about 4 ms out of each second.

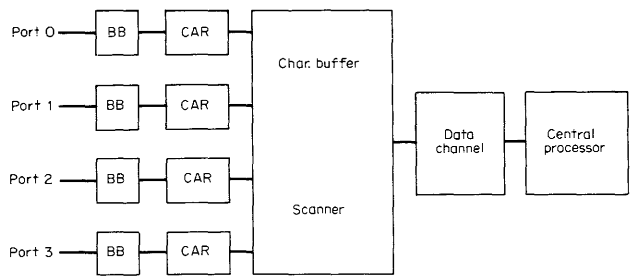


Figure 6. Character-buffer data set adapter

The bit-buffer controlling technique, in this example, has used about half the time available in the central processor at full system load. The number of bits per character, the operating bit rate of the channel, and the synchronous/asynchronous network controls are all under the control of the communications software. These parameters can be varied as required by modifying the software. As terminals on ports change dynamically (by dialing or by mixing varieties of terminal types on a multidrop polled line), the appropriate software changes can be made automatically.

A character-buffering scanning system configuration permits fixed parameters to be established for the three attributes of:

- Maximum number of bits per character.
- Bit rate.
- Synchronous or asynchronous timing.

Efficiency can be gained by providing a more sophisticated data set adapter. Such a data set adapter includes the necessary logic to collect bits into characters.

The character-buffer data set adapter (see Figure 6) includes the bit buffer, as well as a character accumulation register and counting logic. The hardware in the DSA is responsible for detecting the start bit, accumulating characters, and notifying the central processor when a character is done. Many data set adapters of this type include a pair of character accumulation registers; one contains the last complete character received (or next full character ready for transmission), and the other is empty (in the quiescent state) or partly used. If only a single character accumulation register is available, then when a character is received, the central processor must clear the CAR between characters. With two accumulation registers, the processor has a full-character interval time in which to clear the holding CAR. This time permits a load leveling that can be used to enhance system efficiency through scheduling in the software.

Once a character being received is detected as complete by the data set adapter, the active character accumulation register's contents are transferred to the holding register, and the processor is notified of the character's completion. Once the software answers the interrupt, the character may be read from the data set adapter through the scanner into the central processor memory.

Characters received must be placed into the memory buffers, and characters to be transmitted must be selected from these buffers, with the associated software actions. The efficiency of a character-buffering system is higher than that of the bit-buffering system, since software is not involved in collecting bits into characters. If the character-servicing software still requires 100 μs per character, then the worst conditions of system activity will still yield a consumption of 45 ms per second (or 4.5 per cent) of the processor's capacity.

	Bit interrupt	Character interrupt	Buffer interrupt
Bit program time	45.0%	—	—
Character program time	4.5%	4.5%	—
Buffer program time	0.4%	0.4%	0.4%
Total program time	49.9%	4.9%	0.4%

Table 3. Impact of multiplexing on CPU overhead

Since buffer servicing is also identical with the bit-buffering case, this kind of data set adapter will require the associated software to occupy a total of about 5 per cent of the processor's time. The next obvious step is an interface which not only collects characters, but deposits them in buffers in the central memory, interrupting only when buffers are full. Such scanners are called buffer controllers. The buffer controlling scanner services data set adapters, which include individual bit buffers and character accumulation logic for each port.

When a character is received, the data set adapter collects the bits into a character. Once the entire

A Brief Classification Guide to Data Communications Hardware

character has been recognized, the scanner deposits it into central memory. The data set adapter contains the logic required to determine the number of bits which constitute a character, and the time interval between bits. Generally, this fixes the characteristics of the terminal devices which may be serviced on a particular port. However, different ports may be characterized by different data sets and adapters to support the specific mix required by the installation.

The buffer control word pointer table is known to the multiplexer. The servicing of a particular buffer is determined by adding the port number to the base address of this table to arrive at the location containing a pointer. Assume that a character has been sensed and collected for Port 3. The base address is added to 3, and the pointer in this computed address is read from memory. This content points to a set of buffer control words. To distinguish between different buffer control words for a particular port, this will be called BCW3.

The buffer control word basically contains two addresses. As with the software-controlled buffering schemes of previous techniques, these pointers refer to the next and end character positions. When the character arrives, and the buffer control word is accessed via the pointer, the character may be deposited in the location addressed by the next character pointer. The pointer is then incremented by 1, and checked against the end-of-buffer pointer. If the end has not yet been passed, the multiplexer recognizes that servicing is finished for this character for this port. If, on the other hand, the buffer has been filled, an interrupt is generated.

The impact of the multiplexing method on the central processor is summarized in Table 3. For the purposes of an example, this case has been chosen to be consistent for all three communication controller transfer techniques. The purpose of all higher level communications control for even the most complex networks is exactly the same as the example just given for a single processor—to reduce the demands on processing intelligence to the smallest possible fraction of a process time by relegating the administrative chores of communications management to a separate, dedicated controller. The controller can be as little as a circuit in a mainframe or as much as a pair of IBM 360/158's within a large network of many independent communicating modes.□

The Integrated Uses of Hardware in Network Configurations

Problem:

The leap from separate hardware devices to the techniques of integrating the devices into a controlled structure called a network requires a substantial mental readjustment for the engineer/designer who is accustomed to thinking of a system design job as a set of neatly separable tasks. Communications networks are relatively new, but they have already evolved into extremely complex multilayered systems of hardware, software, and protocols that can no longer be considered independently. The "leap" is thus an adaptation to the necessities of interdependent hardware design. The adaptation is neither simple nor easy, but it must be made.

This report is offered as a basic definition of a network because it properly relates all the hardware pieces into the network context and explains the elemental patterns of common network configurations. Later reports in this service will grapple more directly with the discrete problems of network control architectures, message switching, and packet switching.

Solution:

A data communications network consists of a number of different elements, each with a distinct role to play in enabling the network as a whole to meet its objectives. In the first half of this report, we consider the basic elements and their functions, to provide the foundation for the subsequent discussion on network configurations in the second half.

At its simplest, a data communications facility consists of some form of input or output unit, a communications link, and a host processor. Most networks contain many more than these three basic ingredients, but it is worth remembering that they are all designed to facilitate, expedite, or make more efficient the basic functions of inputting, transmitting, processing, and outputting data.

The network elements considered in this report are:

- terminals
- lines
- modems
- concentrators
- multiplexers
- control units
- front-end processors

From *Planning for Data Communications* by John E. Bingham and Garth W.P. Davies. © John E. Bingham & Garth W.P. Davies. Reprinted by permission of MacMillan, London and Basingstroke.

The Integrated Uses of Hardware in Network Configurations

- host processors
- communications software.

TERMINALS

In a data communications system, a terminal provides an interface between the user of the system and the computer. Thus, a terminal may be either a means for the user to submit input into the computer or to receive output from it; alternatively, the terminal may fulfill both of these functions. Terminals are in the main—though not always—located remotely from the computer to which they are connected.

There are a number of ways in which terminals may be classified, but the most useful approaches are to classify them by function or use. According to this type of classification, the main types of terminal are:

1. teletypewriter terminals
2. printing terminals
3. visual display terminals
4. remote batch terminals
5. graph plotters
6. point-of-sale terminals
7. factory data capture devices
8. banking terminals.

Of these categories, types 1-5 may be described as general purpose in that they can be used for wide variety of applications, whereas the remaining three categories may be called application specific in that they are suitable for only a subset of the total number of data communications systems.

Teletypewriter terminals comprise, as their name implies, a simple keyboard analogous to that of a typewriter, together with a few simple control keys. Similarity to the typewriter is enhanced by the usual provision of a hard-copy printout of the information entered on the keyboard. Keyboards can be combined with one or more other input or output features, for example, a paper tape reader, to provide a more versatile terminal.

At the simplest level, the terminal is linked directly to the communications channel and hence to the computer, but since this limits the data transfer rate to the performance of the terminal operator—often someone with 'hunt and peck' typing ability—various measures have been adopted to improve data transfer rates. Two of these, the preparation of paper tape off-line and buffering, require further comment. Where considerable quantities of data have to be transmitted, many teletypewriter terminals have the facility to handle an intermediate medium (usually paper tape, although punched cards and cassette tapes are also used) in an off-line mode (that is, with the terminal not connected to the computer). This step can be taken at any convenient time, and later, either at the user's discretion or according to a predetermined schedule—the choice depending on the application—the data may be transmitted to the computer and perhaps the keyboard—printer combination can then be used interactively to assist the process.

Simple keyboard/printer devices (with or without intermediate data storage) are commonly used for data collection from large numbers of remote points, for example, all branches or depots of a wholesale or retail business (because they are cheap and easy to use), and as terminals to timesharing services.

Printing terminals range from the simple teletypewriter type of machine described above in use as an output device to variants of the standard high speed computer line printer (in this context, a device capable of producing 350 or more lines of printed output per minute).

Visual display terminals consist of a keyboard associated with a cathode ray tube (CRT) or other display screen closely resembling that of a television set. In many of these devices a buffering mechanism is included to substitute for the intermediate storage media used in keyboard terminals. The buffer is a store into which a certain amount of information may be entered. The operator then presses a 'release' key to transmit the data. The size of the buffer is usually linked to the size of the screen so that the operator may enter enough data (which will be displayed on the screen to enable a visual check to be made) to fill the screen or complete a particular record or request before transmitting data.

In a number of cases the display terminal incorporates a 'forms feature.' This feature enables the data (either entered via the keyboard or received from the computer) to be displayed in a predetermined format or formats corresponding to clerically completed forms, and may display the headings concerned. When used for data entry (that is, operator keyboard use), the position into which the operator will key the next character of information is marked by a cursor that usually takes the form of underlining the character concerned. In 'forms feature' operation, the cursor automatically skips to the next space to be completed.

Other forms of display terminals utilize light pens or other devices that enable the operator to enter information or data on to the screen directly without using a keyboard. Typically, this will involve the selection of an item from a displayed list by pointing the light pen

The Integrated Uses of Hardware in Network Configurations

at that item. Devices of this type make it easier for the operator to conduct an interactive dialogue with the computer via the terminal.

Remote batch terminals. Remote batch operations consist essentially of providing the facilities at a distance for a computer to be used in conventional batch mode, that is, collecting data into groups or 'batches' of transactions and processing at predetermined or otherwise convenient intervals. The terminals used to support this type of operation are therefore somewhat similar to conventional computer peripherals. The essence of batch operations generally requires relatively high volumes of input or output data to be handled.

A typical remote batch terminal provides a number of facilities to the user and normally features both input and output devices, enabling the user to effectively have 'hands on' use of the computer for his own application. A typical remote batch terminal configuration might include a card reader, a card punch, a line printer, and an operator's console.

These peripherals are connected to a control unit. Other common peripherals are magnetic tape readers/writers and, disk storage devices. Many such devices are called 'intelligent,' which essentially means that they can be programmed by the user in just the same way as a small computer (which is exactly what they then are). It is usage rather than the configuration of the device itself that determines whether it is a remote batch terminal. Thus, the same device may be usable as a stand-alone local computer doing batch processing, a data capture device of the keydisk type, and as a remote batch terminal, and indeed may operate in all three modes in the course of a day. There is an increasing tendency for intelligence to be built into terminals in this way both for economic and systems reasons. Economically, the arrangement may save costs by transmitting summarized or partly processed data. Many systems advantages may accrue because data validation and similar queries can be dealt with conveniently close to their source.

Graph plotters used as terminals are exactly the same as those used as peripherals for a central computer. Their application is suitable where the presentation of numerical data in hard copy visual form is required. Devices of this type may be associated with other forms of input/output as part of a remote batch terminal or can be used on a stand-alone basis for output.

Point-of-sale terminals are devices designed specifically for data capture in a retail sales environment. Typically they serve both the need to provide a 'cash register' and also as a means of capturing basic data on stock movement. Among the devices available are those that only possess specialized keyboards, and those that make use of other technologies such as a laser-based light pen. This device, which resembles a largish pen in shape and size, is passed over a specially encoded slip attached to the merchandise and captures such information as size, style, color, price category, etc., depending on the nature of the goods. The captured data will immediately (and usually within the terminal itself) produce a conventional cash register listing for the customer and virtually simultaneously produce a record on a storage medium, for example, a cassette tape, for subsequent transmission and processing on the company's central computer facility.

Factory data capture devices represent a wide and growing range of terminal devices designed specifically to capture data in factory locations. Three fundamental principles are, however, common to all types: simplicity of operation, robustness, and reliability. In the majority of cases, terminals of this group are designed for operation by employees who are employed for skills completely unrelated to data capture: moreover, the company generally has a great interest in minimizing the amount of time such 'productive' employees spend on 'administrative' tasks. Thus simplicity of operation becomes a design criterion of paramount importance. The location of many such terminals adjacent to work stations in factories means that they have to be built to the same standards of robustness as the other equipment in that location. Thus, dirt, heat, oil, chemical corrosion, etc., may all have to be resisted—hazards uncommon in conventional office environments. Reliability is also important since unless a high level of serviceability can be guaranteed, high costs are incurred by the necessity to provide adequate backup. Characteristics terminals in this category include a card or badge reader that collects data from a job ticket or batch card accompanying the work through the factory, for example, a piece of machinery being assembled. This card or badge is read at each terminal in the factory as it physically reaches that point, and a certain amount of additional information is added, for example, employee number, by means either of keyboard entry or a second card or badge. Some data, for example, terminal identity, time, etc., may be captured without operator intervention other than pressing a single key. Data captured will be transmitted for either batch or on-line processing.

Banking terminals probably represent the largest category of specialized data communications terminals in use today. Within the United Kingdom and the Federal Republic of Germany, for example, they exist in tens of thousands. The nature of the facilities provided naturally varies somewhat depending on the country concerned and the type of bank but will often include the ability to update a pass or savings book, produce a local hard copy of the transaction (for audit

The Integrated Uses of Hardware in Network Configurations

purposes), and capture the transaction on cassette tape for subsequent transmission.

In addition to the types of terminal discussed above, many other types of equipment may be regarded as terminals in certain applications. This is especially true of many types of specialized data capture equipment, for example magnetic ink character readers, bar code readers, optical character readers, etc.

COMMUNICATIONS LINES

The actual communications channels along which data are passed are frequently known as lines. The use of this term is, however, strictly speaking incorrect since the channel may actually take the form not only of physical lines but also of microwave radio. Moreover, the actual form of the channel is not known to the user and may consist of a combination of lines and radiowave links. The term 'communications channel' is more accurate but tends to cause confusion with the term channel as used in conjunction with central processors. Hence the words 'communications link' are the most widely acceptable term.

Much the most important form of communications link is provided by the standard public telephone and telegraph networks, owing to their almost universal availability. The facilities available ditter somewhat depending on the location concerned. Some of these facilities may also be reserved for the exclusive use of a particular company or concern, although these are in the main also provided by the same authority as the public services.

Links of the type now used for the vast majority of data communications traffic were not originally conceived for this type of use, and special measures have to be taken to create an interface between the originating and receiving points and the link handling the transmission of the data. (Data transmission over a telephone network occurs by modifying the basic waveform of the link to indicate either of the binary states 0 or 1). This function is met by a device at each end of the link called a modem, which will be considered in the following section. It should be noted in passing, however, that over very short distances and where it is possible to use communications links that are not part of the public network, this type of interface may be dispensed with. At this level, the communications link is effectively an extended version of the standard cable link between the central processor of a computer and its peripheral devices.

Data is transmitted over a link in one of a number of codes. These codes represent the data in the form of a string of binary digits and also enable certain essential instruction information to be included in the traffic passed over the link. The codes available break the stream of binary digits (bits) into groups of 5 to 8 bits. Each group customarily represents a character of data or a control character. A code of five-bit units enables 25 or 32 different combinations, a six-bit code 26 or 64 combinations, a seven-bit code 27 or 128, and an eight-bit code 28 or 256 combinations. For most purposes it is desirable to provide combinations of bits or codes for the numerals 0-9 and the complete alphabet (possibly in both upper and lower case), a number of special characters (punctuation marks, £ or other currency signs), the control characters (which include characters to control the devices at the end of the link, for example, carriage control movements) as well as control the link itself. It might therefore appear that the number of combinations or codes required would preclude the use of anything less than a six-bit code. In practice however, the member of combinations in even five bits can be rendered sufficient by the use of 'shift' codes. These special codes modify the value of succeeding characters until another shift character is encountered. Effectively this means that the same five-bit pattern can represent as many different characters as there are shift characters (although for obvious reasons control characters have the same meaning in all shifts). The most commonly used example of this ingenious principle is the CCITT Alphabet No. 2 popularly known as the Baudot code (figure 1). Although widely used for telegraphic purposes, this code is not ideal for data communications purposes, and a considerable number of alternatives have been developed.

One of these is the eight-bit Extended Binary Coded Decimal Interchange Code usually known as EBCDIC which is widely used for data communications with computers (figure 2).

It should be noted that the codes in use for communications purposes differ from those used by computers internally in structure as well as format and code translation is therefore a common necessity.

The codes described above indicate the number of bits that will be necessary to represent any given character; they do not indicate the number of bits that will be necessary to transmit that character. To actually transmit a character of information, additional bits are required. The number involved (which is of importance to the analyst when configuring the network) depends on a number of factors. Among the most important of these are error checking and message headers.

Error checking is commonly performed by adding a separate bit to the character code to achieve odd or even parity. Similarly a parity character may be added to each block of data to achieve block parity. These techniques are illustrated in figure 3.

The Integrated Uses of Hardware in Network Configurations

BIT 1	BIT 2	BIT 3	BIT 4	BIT 5	MEANING	
					Lower case	Upper case
0	0	0	0	0	no action	no action
1	0	0	0	0	T	5
0	1	0	0	0	carriage return	carriage return
1	1	0	0	0	O	9
0	0	1	0	0	space	space
1	0	1	0	0	H	
0	1	1	0	0	N	comma
1	1	1	0	0	M	full stop (period)
0	0	0	1	0	line feed	line feed
1	0	0	1	0	L	right bracket
0	1	0	1	0	R	4
1	1	0	1	0	G	
0	0	1	1	0	I	8
1	0	1	1	0	P	0(zero)
0	1	1	1	0	C	colon
1	1	1	1	0	V	equals sign
0	0	0	0	1	E	3
1	0	0	0	1	Z	plus sign
0	1	0	0	1	D	who are you?
1	1	0	0	1	B	question mark
0	0	1	0	1	S	apostrophe
1	0	1	0	1	Y	6
0	1	1	0	1	F	
1	1	1	0	1	X	oblique stroke
0	0	0	1	1	A	minus sign
1	0	0	1	1	W	2
0	1	0	1	1	J	bell
1	1	0	1	1	figure (upper case) shift	figure (upper case) shift
0	0	1	1	1	U	7
1	0	1	1	1	Q	1
0	1	1	1	1	K	left bracket
1	1	1	1	1	letter (lower case) shift	letter (lower case) shift

Figure 1. CCITT alphabet number 2 (the Baudot code). Note: Unallocated combinations are used for specific national symbols and should, therefore, never be used for international transmission

Bit code	Meaning	Bit code	Meaning
1100 0001	A	1111 0000	0
0010	B	0001	1
0011	C	0010	2
0100	D	0011	3
0101	E	0100	4
0110	F	0101	5
0111	G	0110	6
1000	H	0111	7
1001	I	1000	8
1101 0001	J	1001	9
0010	K	0000 0100	PF
0011	L	0101	HT
0100	M	0110	LC
0101	N	0111	DEL
0110	O	0001 0100	RES
0111	P	0101	NL
1000	Q	0110	BS
1001	R	0111	IDL
1110 0010	S	0010 0100	BYP
0011	T	0101	LF
0100	U	0110	EOB
0101	V	0111	PRE
0110	W	0011 0100	PN
0111	X	0101	RS
1000	Y	0110	UC
1001	Z	0111	EOT
1000 0001	a	0100 0000	blank
0010	b	1100 0000	>
0011	c	1101 0000	<
0100	d	0100 1010	?
0101	e	1011	.
0110	f	1100	←
0111	g	1101	(
1000	h	1110	+
1001	i	1111	‡
1001 0001	j	0101 1010	!
0010	k	1011	
0011	l	1100	*
0100	m	1101	)
0101	n	1110	;
0110	o	1111	¢
0111	p	0110 1001	/
1000	q	1011	,
1001	r	1100	%
1010 0010	s	1101	_
0011	t	1110	–
0100	u	1111	±
0101	v	0111 1001	"
0110	w	1010	:
0111	x	1011	#
1000	y	1100	@
1001	z	1101	'
		1110	=
		1111	√

(PF through EOT: Control characters)

Figure 2. Extended binary coded decimal interchange code (EBCDIC)

0	1	0	0	0	1	0	0	0	0	1	0
1	0	1	0	1	0	1	1	1	1	1	1
0	1	0	1	0	1	1	1	0	0	1	1
0	0	0	1	1	0	1	1	0	0	1	0
0	1	0	1	1	0	1	1	0	0	0	0
0	0	0	0	0	0	0	1	1	0	0	1
1	0	0	0	1	0	0	1	0	0	1	1
1	0	0	0	1	1	1	1	1	0	0	1

Character parity bit (odd parity)

Block parity bits (odd parity)

Figure 3. Character and block parity checking. Note: The circled bit effectively serves as a double check since it must be correct parity both longitudinally and vertically

Message headers are characters necessary to transmit to enable the receiving end of the link to identify and/or process the data. The volume of this information varies according to the nature of the communications network. Where asynchronous (or start—stop) transmission is used, each character of information transmitted is preceded by a bit pattern that alerts the receiving station to the fact that data will be transmitted and establishes the synchronization necessary for the receiving station to interpret it correctly. Similarly, the data charater is followed by a bit pattern that re-establishes synchronization for the second data character, and so on. This arrangement obviously imposes a high overhead of network administration data to actual data and may be of the order of 30 percent. This can largely be avoided by using a

The Integrated Uses of Hardware in Network Configurations

stream of bits at periodic intervals to synchronize the transmitting and receiving devices. This transmission mode, known as synchronous transmission, is only satisfactory if the data is to be transmitted in continuous blocks, for example, the contents of a terminal buffer. if the data flow stops, it is normally necessary to resynchronize by the appropriate bit stream.

Another feature of links is the mode in which they operate. The terms simplex, half duplex, and duplex are in common usage. The usual meaning of simplex is that data can be transmitted along the link in one direction or the other but not both. half duplex generally means that data can be transmitted in either direction or the other but not both. Half duplex interpretation of duplex (or full duplex) operation is that such a link enables simultaneous transmission in both directions.

Our next consideration is that of the speed at which data can be transmitted over a data communications link. Basically, the capacity of a link to transmit data is limited only by the bandwidth associated with it. Within this limitation the modems in use actually determine the amount of data that can be carried in any given time. The rate at which data can be transmitted may be expressed in bauds or bits per second. Much confusion exists over the correct use of the term baud, which may or may not equal the transmission rate in bits per second. In general usage, the term 'bauds' is taken to be equivalent to bits per second, but the formal definition of a baud is 'the number of code elements per second'. Therefore, it is only true to say bauds equal bits per second for codes with equal bit length. The non-data-communications technician is well advised to use bits per second (bps) exclusively.

Finally, in our discussion of links we must consider the topic of contention and polling. For every data communications link there must be a set of rules to govern the transmission of data across the link. It will be readily observed that in addition to the use of a common code at both ends of the link, rules must exist to determine when data is transmitted and when any individual unit attached to the link changes from operating in 'receive' mode and starts to operate in 'send' mode.

Two main approaches exist to this latter problem; contention and polling. Contention basically gives the right to either end of the link to seize control over it if it is vacant. This is achieved by the sending end of the link originating an enquiry; if the receiving end of the link is able to accept data, it acknowledges the signal, and transmission begins. This precaution prevents loss of data by transmission to an invalid (or switched off) destination. At appropriate intervals the receiving unit will acknowledge that the data has been received and that the transmission has met the parity and other checks in force (if it fails to meet this standard, the receiving unit makes a different signal and the data sent since the last acknowledgement is retransmitted). The potential problems arising from simultaneous enquiries from both ends or from a busy receiving device is also handled by the provision of timing devices in both the sending and receiving units to enable the unit to either repeat the action or move on to something else (for example, create an operator message) if no response is received within a predetermined time. This prevents a unit from going into a permanent 'wait' state.

The second approach to controlling the business conducted over the data communications link is that of polling. In this approach the host computer (or a nominated intelligent device) effectively 'asks' each sending station if it has any data to be transmitted. If the 'answer' is yes, the link is established, and the message is sent. When transmission is complete, another device is asked if it has data to transmit, and so on. All devices may be asked sequentially, or those that handle the most data may be asked or polled more frequently than the others.

MODEMS (MODULATOR/DEMODULATOR)

Modems are devices that provide the translation between the digital signals used by computers and terminals and the analogue signals used on the communications link. Effectively this means loading the digital signal (in binary representation) on to the normal carrier wave (the latter usually being in sine wave form). This process is reversed at the receiving end of the link where the digital signal is 'stripped' from the carrier wave. The digital signal can be loaded onto the carrier wave in one of three ways (see Figure 4): by amplitude modulation, frequency modulation, or phase modulation. Each of these techniques refers to the manner in which the basic carrier wave is modified (modulated) so that the receiving unit is aware that it is carrying information. Amplitude modulation tends to be more susceptible to interference than either frequency or phase modulation, and for that reason the latter techniques are to be preferred. From the user point of view, however, a modem is a 'black box,' and the technique, actually used for transmission need not be known; what is of paramount interest is the specification of the link/modems provided. The two vital criteria are speed of transmission and error rates. Although it is usually the speed of the link that is referred to, it is really more accurate to refer to the speed at which the modem can load signals on to the link.

In many cases (indeed in most cases in Europe) the attachment of a modem to any given communications link is not entirely at the discretion of the user but rather at the command of the relevant telecommunica-

The Integrated Uses of Hardware in Network Configurations

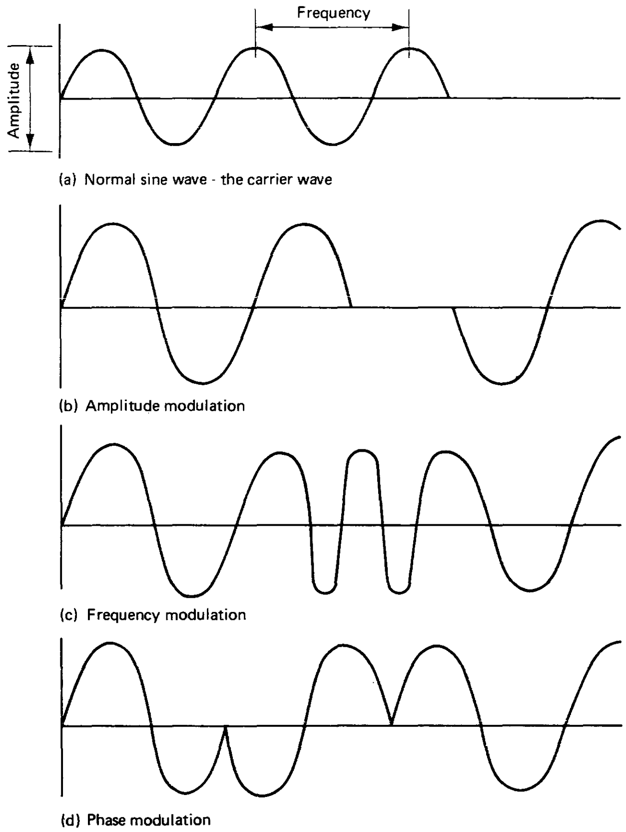


Figure 4. Modulating techniques: (a) normal sine wave—the carrier wave; (b) amplitude modulation; (c) frequency modulation; (d) phase modulation

tions authority. This control is enforced by maintaining a register of approved modems that can be attached to any given type of communications link.

Most modems are attached directly to the communications link in use and remain permanently in one physical position. Some modems are, however, constructed instead as accoustical couplers, which makes them more portable. The acoustic coupler modulates the digital input in such a way that it can be transmitted over a normal dialed telephone link. This device, which is physically a small 'black box' with a rest or cradle on to which a standard telephone handset can be placed, is often combined with a simple keyboard/hard-copy printer unit to provide a terminal of average briefcase size and portability. Such acoustically coupled terminals are being increasingly used because of their convenience. They can operate in offices, meeting rooms, etc.—anywhere a normal telephone is available.

CONCENTRATORS

Conceptually the simplest way of linking a number of remote locations to a single host processor is to provide a separate communications link to each remote location. The cost of the communications links in this type of network can quickly become prohibitive, and various approaches have been adopted to overcome the problem—two of the approaches being concentrators and multiplexers. The concentrator is effectively a computer in its own right. As its names implies, it concentrates the data travelling to or from a number of remote locations into 'bulk loads.' Thus, a concentrator may 'collect' data from a number of relatively slow-speed terminals using appropriate low speed and probably asynchronous links and interleave them for transmission over a higher performance link to the host processor.

Being intelligent, concentrators may be programmed to perform a variety of tasks, which fequently include data validation, and for this purpose storage devices containing a variety of files may be attached. Similarly for data flowing outwards from the host processor, that is, towards the terminal, output formatting may be performed, for example, to provide appropriate screen layouts for visual display units. Finally, the possibility of using concentrators to provide at least some of the essential back-up and/or fail-safe options in a network should be mentioned. Thus, a concentrator may perform sufficient processing to enable its part of the network to continue essential processing if one or some of the communications link(s) to it become inoperable. Furthermore, it may store essential data if a link or device on either its inboard (host processor) or outboard (terminal) side fails.

MULTIPLEXERS

Compared with a concentrator, a multiplexer is basically an unintelligent unit that performs the basic role of reducing total communications link costs but has no other function in the network. In this sense, the multiplexer plays a solely economic role with regard to data communications links whereas, as we have seen, a concentrator may have a functional (back-up, data validation, etc.) as well as an economic role. This distinction will probably become less clear as an increasing number of multiplexers are based on programmable minicomputers.

There are two main technqiues used in multiplexing: frequency division and time division. The trend is towards the latter at the expense of the former. Schematic representations of both techniques are shown in Figure 5.

Frequency division multiplexing (FDM) achieves economies in communications links cost by subdividing a medium or high capacity link into a number of parallel subchannels. Such a multiplexer effectively works as a telephone exchange connecting terminals and the host processor (or other unit as a telephone

The Integrated Uses of Hardware in Network Configurations

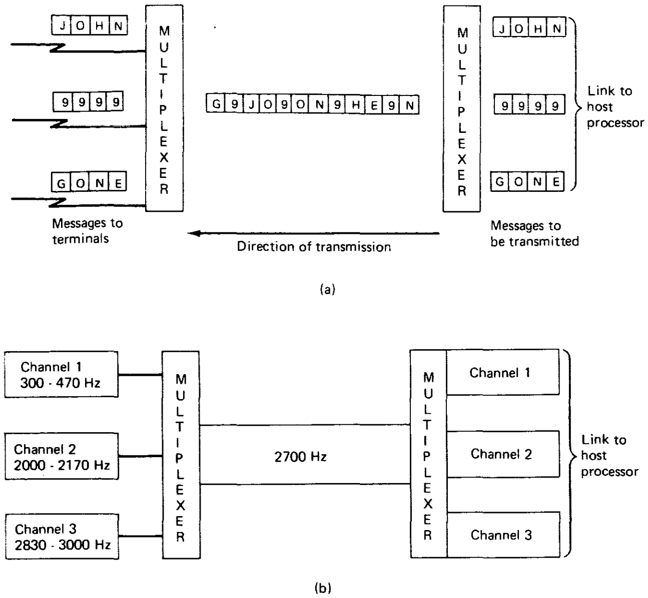


Figure 5. Multiplexing techniques: (a) time division multiplexing—this simplified diagram, with traffic in one direction only, shows how three messages emanating from the host processor are transmitted to the appropriate terminal (note: control characters omitted for clarity; (b) frequency division multiplexing (schematic)

exchange connecting terminals and the host processor (or other unit in the network) using an available subchannel.

Time division multiplexers (TDM) on the other hand may be regarded as analogous to a container train running regularly from one end of the multiplexed link to the other. Each terminal has its own 'container' in each train. This approach is achieved in practice by transmitting sequentially signals from each terminal over the single common path at different instants of time.

CONTROL UNITS (COMMUNICATIONS PROCESSORS)

Even though many modern data communications links are capable of being operated at high speeds, there remains a significant difference between the rates of transmission and the speeds at which processors are able to process the data. Thus there is almost invariably a need for some type of buffering device between the link and the processor. Moreover, most data is actually transmitted in bit serial form (that is, in a sequential pattern of binary characters) whereas processing is typically carried out on parallel characters.The two roles of buffering and code conversion can most accurately be ascribed to a control device which, depending on manufacturer, may also be referred to under the names: communications controller, transmission control unit, communications processor, or by a variety of trade names, acronyms, and initials. These devices, although normally located centrally, may also be dispersed throughout a network and may also include some of the functions described elsewhere in this report for concentrators and front-end processors.

FRONT-END PROCESSORS

A front-end processor is in essence a separate computer interposed between a data communications network and the host processor to perform a substantial part of the communications-oriented (as opposed to applications-oriented) processing. Thus, the range of functions extends from the control unit approach through to a full-scale computer that pre-processes all data communications traffic, including such activities as data validation, and may even provide a fairly extensive set of standby processing capabilities in the event of host processor failure. Thus, a front-end processor may in fact range from a simple dedicated minicomputer tailored to a specific communications environment (and virtually unusable for any other purpose) up to a standard programmable medium or even medium-large general purpose computer with a full range of input, output, and storage devices and could readily be used independently as an 'ordinary' computer. The decision to use a front-end is a practical and economic one and conceptually makes little difference to the user of the network.

HOST PROCESSORS

A practical definition of a host processor is an element (or elements) of a network that performs the applications (as opposed to message handling) processing. Throughout this report the term host processor has been used in a way that indicates that there are definite focal points of a data communications network responsible for the applications work performed by means of the network. Traditionally this has mostly been based on a single host computer with a 'star' or 'tree-type' of structure. Increasingly, however, the term 'host processor' in relation to a network is being used in the plural as more and more networks include more than one applications processing computer. Prototypes for networks with multiple host processors exist already (for example, the ARPA network and the European Informatics Network), and this concept is a major feature of the EURONET project. As has also been noted above, there are occasions when another element of the network may temporarily assume some or (less frequently) many of the functions of the host processor. Typically this occurs when a concentrator or front-end carries out essential processing in the event of a failure in the host processor.

The Integrated Uses of Hardware in Network Configurations

NETWORK CONFIGURATIONS

Now that the various elements of a network have been identified and described, and an approach towards the determination of user needs considered, it is appropriate to discuss the problems of configuring the network.

The process of determining user needs has identified the necessity for data originating in one place to be processed and/or made available at another. The response-time requirements will moreover have established the need to use data communications technology.

The information will also have indicated the type of system that will be required, that is, off-line transmission, remote batch, on-line data collection, enquiry/response, or real-time. The approach to configuring the network is basically the same in each of these cases. The three factors that have to be considered are: the equipment to be installed at the periphery, the equipment to be installed at the center(s), and the communications links that connect them. Clearly the decisions that have to be made are as much economic as technical, and many times the analyst will find himself evaluating tradeoffs between one alternative and another.

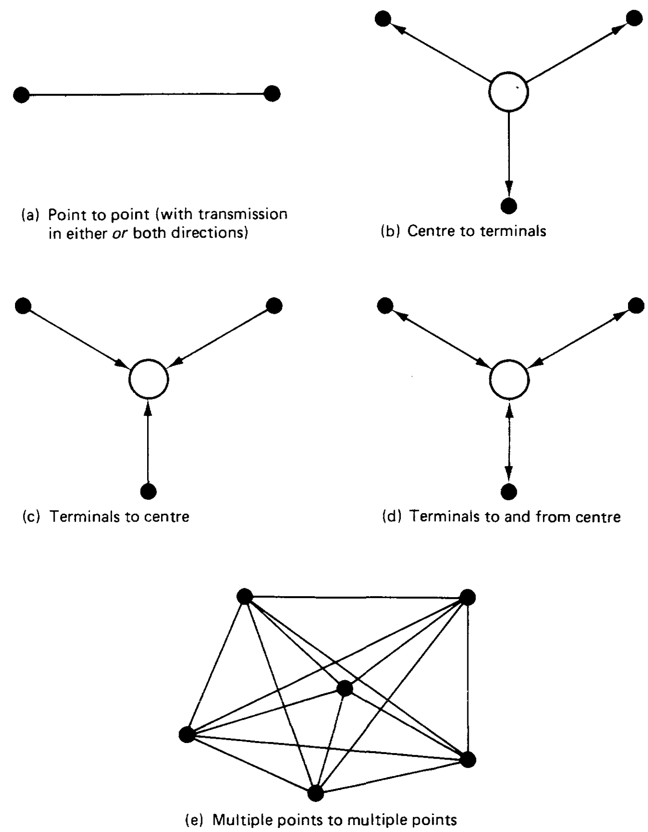


Figure 6. Logical data communications networks: (a) point to point (with transmission in either or both directions); (b) center to terminals; (c) terminals to center; (d) terminals to and from center; (e) multiple points to multiple points. Note: Structures (b), (c), and (d) are often called 'star' networks

The Logical Network

Once the basic parameters concerning the proposed network have been established, the analyst should proceed to produce a visual representation of the logical network. This merely shows the location(s) at which data will be entered into the network, the location(s) at which they will be processed, the location(s) at which they will be output, and the distances between each location.

The main types of communications network patterns are shown in figure 6. Note that at this stage the lines drawn between the input, output, and processing locations do not necessarily represent the paths along which the data will ultimately be transmitted but merely indicate the logical paths or flows.

The next stage in the process of configuring the network is to add the traffic volumes and response times to the logical network diagram.

Where any doubt exists about the frequency of transmissions, for example, in a remote batch environment, it might not initially be clear whether transmission should occur daily or twice daily. The diagram should show the data for the shorter periods since these will determine the communication link requirements. Response times are, of course, shown at the locations at which the data is required. A simple cross-reference scheme is usually necessary to ensure readability.

A logical network diagram annotated with response times, traffic volumes and a brief statement of the applications is shown in figure 7.

From the logical network diagram a preliminary view of the actual network is beginning to emerge, and once this has been established the analyst can commence the design of the actual network to be used. At this point, the inexperienced user is well advised to seek specialized advice from the equipment vendors, and consultants since for any but the most simple point-to-point transmission system a level of technical expertise will be needed that cannot be easily obtained without specialized guidance.

Clearly, the complexity of the task of designing the network will vary with the type of system, for example, off-line transmission, real-time, the extent of the network, that is, the number of locations serviced, and the extent to which existing facilities can (or have to be) utilized, for example, an existing computer center

The Integrated Uses of Hardware in Network Configurations

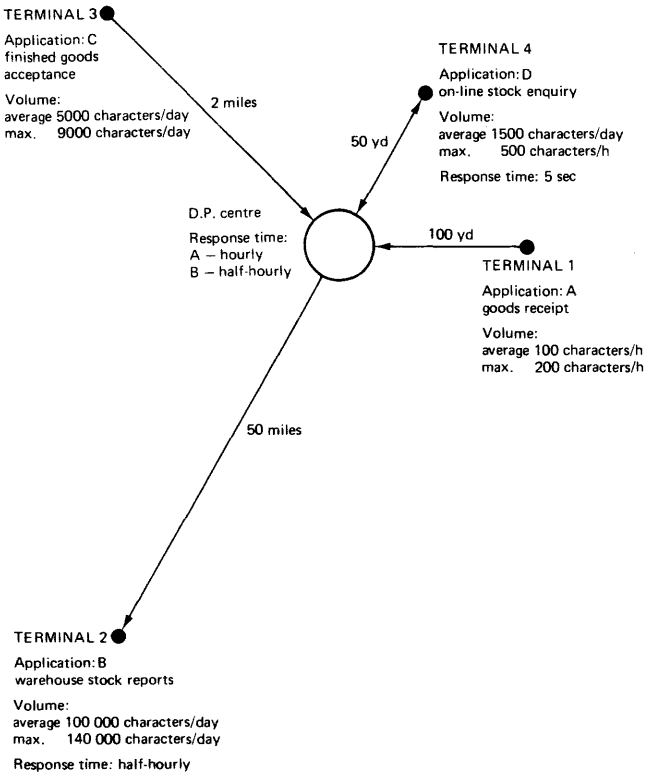


Figure 7. A logical data flow chart with response times

and equipment. As noted above, however, the basic principles are in all cases the same.

NETWORK CONFIGURATION ALTERNATIVES

When the analyst actually starts to determine which piece of equipment goes where and how pieces are to be linked together, he is faced with a bewildering range of possibilities. In the following paragraphs the major decisions to be made are isolated and the major factors determining the decision identified. The major decisions to be made are:

- use of public or private communications links
- speed of communications links
- the physical arrangment of communications links
- use of concentrators and/or multiplexers
- number, capacity and location of processing centers
- use of front-end processors
- terminals and distribution of intelligence
- physical arrangement of terminals

In any given situation the decision to be taken may, of course, have been pre-empted by previous decisions, for example, an installed private communication link for voice traffic, which is also availble and suitable for data traffic. The above decisions, while they may be considered separately are, moreover, interrelated, so each decision as taken will impinge on all the other decisions, making the complete decision process a re-iterative one.

User of Public or Private Communications Links

One of the fundamental decisions to be made in the early stages of configuring a network is whether private (also called leased) or public communications links should be used. The first consideration is, of course, whether the required link is available as both a public and private service. In many cases the higher quality communications links are only available as private or leased facilities. A second prime consideration is the economics of each alternative. The most effective way of answering the economic question is to draw a graph of the costs of the required line speed for various periods of utilization above and below the estimated workload. Care should be taken to base the charges for public lines on total connect times at the rates applicable at the time the transmission will take place, since it is common practice to charge different rates depending on the time of day and/or day of the week. A sample graph is shown in figure 8. Among the other considerations to be evaluated are link quality (and the associated error rates) and the security and reliability factors. Quality is, as a generalization, higher on private than on public links, and this alone may be sufficient to persuade the analyst to select the

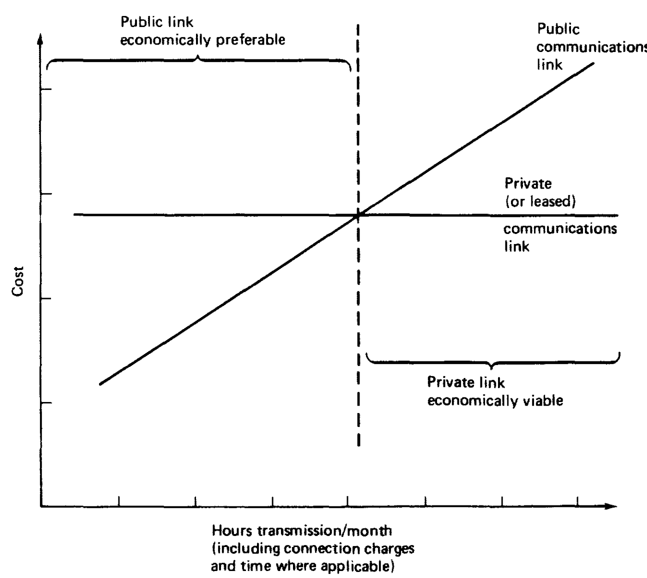


Figure 8. Economic viability of public v. private communications links for any specific link speed and distance

The Integrated Uses of Hardware in Network Configurations

private link even if the public link offers somewhat more favorable economics in the case under consideration. Security and reliability factors may, paradoxically, suggest opposite conclusions. More extensive security measures may be possible with private than public links, although for most commercial organizations the anonymity of the public links offers a high degree of security. Reliability is generally better with a public service because of the multiplicity of paths between source and destination offered by the common carrier. The vulnerability of the private link in this case can, however, be easily overstated because the private link user may well be able to utilize public links for back-up purposes.

In any given network it is quite possible and practical to specify both public and private communications links to meet specific needs.

Physical Arrangement of Communications Links

The analyst working on any network making use of private communications links has the task of defining the basic topology of the network to support the applications for which he is responsible (the analyst electing to use public facilities is spared this task, since when the data traffic is 'handed over' to the common carrier, the actual route it follows is both transparent and irrelevant to the user organization).

The basic objectives in determining the physical arrangements of communications links are:

1. to minimize the total communications link distance (and therefore cost)

2. to provide alternative routings for as much data traffic as possible

3. to equalize as far as possible the workload over the available links and equipment.

Clearly, these objectives are to a certain extent incompatible, and thus some careful judgements are required. Among the generalized conclusions to which these points lead are that the above objectives are more difficult to meet for the 'spider's web' type network (figure 9) than for circular networks (figure 10) or a network in which trunk and local routes are separated (figure 11), except where it is feasible to have multiple switching centers. The typical organization tends to lend itself to the latter approach rather than the former in as much as it is common for an organization to have a concentration of activities in a few locations, each of which is remote from the major processing site. It becomes economically attractive in this environment to utilize high-speed trunk communications links, each of which serves multiple terminals, con-

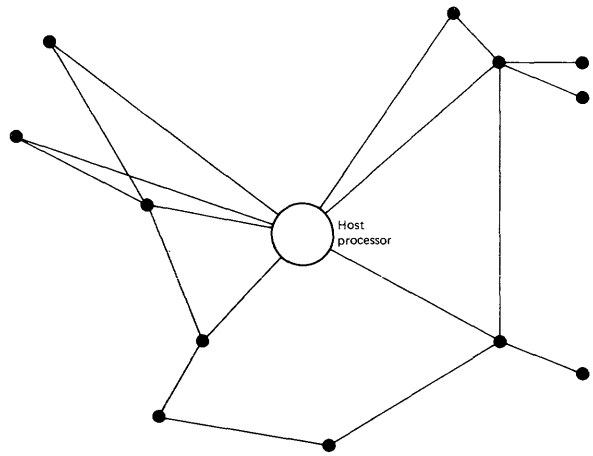


Figure 9. A 'spider's web' network

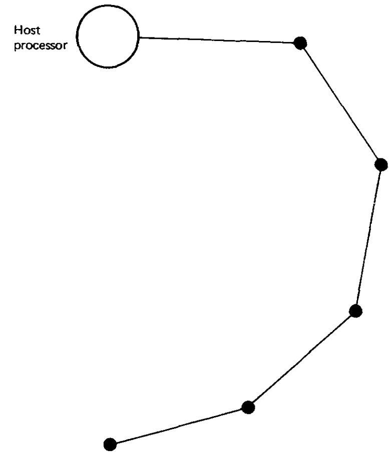


Figure 10. A circular network (utilizing multidrop lines)

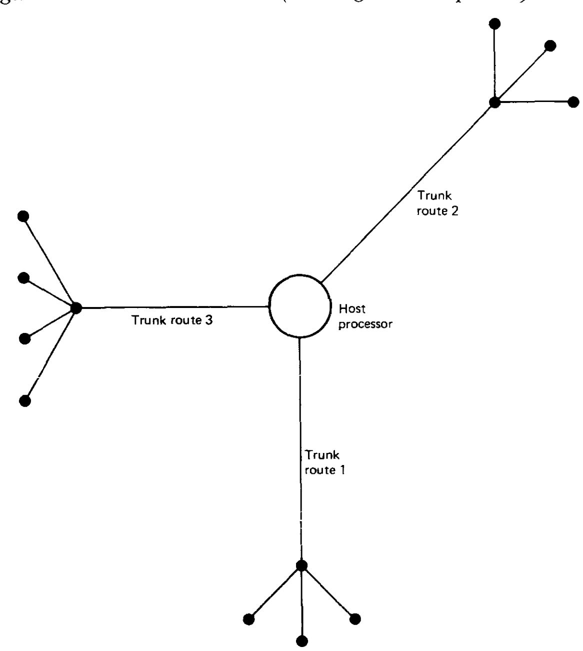


Figure 11. Separation of trunk and local routes in a network using concentrators or multiplexers (tree structure). Note: Local routes may be as shown or use the multidrop approach

The Integrated Uses of Hardware in Network Configurations

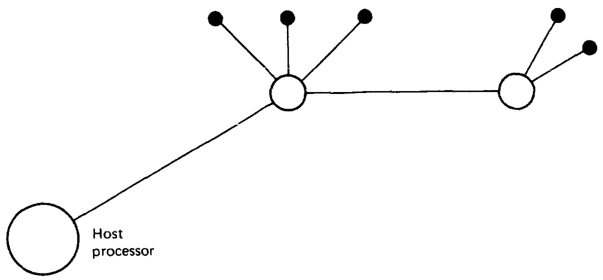


Figure 12. Multidropped multiplexers or concentrators

centrators, or multiplexers being used to assemble and distribute the 'main line' traffic.

Use of Concentrators and/or Multiplexers

The topology of a network cannot be finalized until decisions about the use of concentrators and/or multiplexers have been taken. Although these two types of equipment are basically different, they nevertheless have a common purpose in facilitating the type of network shown in figure 11. There are a number of ways in which these devices can be utilized to develop a practical and economic network. Concentrators can, moreover, also have the effect of distributing (decentralizing) intelligence throughout the network, thus reducing the total volume of data transmitted. Two ways in which multiplexers and/or concentrators can be featured in a network are shown in figures 12 and 13. The basic decisions facing the analyst are: firstly, which is the most economical solution for the situation he is facing, and, secondly, which provides the highest level of performance and reliability? Further, if these criteria produce different solutions, is the improvement worth the cost? There are two basic approaches that can be used to help in determining where and how to use concentrators and/or multiplexers. The first is a straight hand calculation method—identifying the various alternatives and costing them out individually. The second approach is the use of software packages.

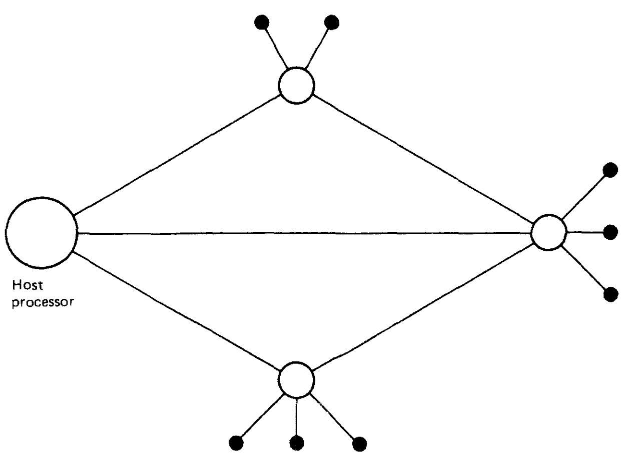


Figure 13. Interlinked concentrators

Number, Capacity and Location of Processing Centers

So far we have implicitly considered networks in which there is a single host processor (whether or not this physically consists of two or more central processing units). In practice, it may be necessary or desirable to provide multiple host processors. The capacity of each host processor must, of course, be sufficient not only to enable it to fulfill its role in the network but also to handle any batch work for which it may also be used. It is notoriously difficult to forecast workload, performance, and consequently the capacity required in a data communications environment. The use of simulation can materially assist in this calculation, and the analyst must also consider the question of additional equipment for reliability. This situation may arise when a concern decides to link a number of existing computer centers into one network, when the applications needs are best served by different host processors or when, for workload or security considerations, the use of two or more host processors is desirable.

In most cases the siting of the host processor(s) will be predetermined by the location of existing computer centers or the particular application(s) they are intended to support. Where this is not the case, the optimum site for each host processor is that which minimizes total network costs. A major factor in determining this site will be line costs (where all lines are of equal cost per unit distance this site will be the one that minimizes total network mileage), but in addition such factors as property prices, suitability (and planning approval) for building and the local common carrier facilities may also play a major role in the choice.

The use of multiple host processors implies the use of a switching center in the network. The function of a switching center is to receive all messages inward and outward bound and to route them to the appropriate destination. The most appropriate approach to siting switching centers is to use manual matrix techniques or suitable software packages.

Although theoretically a switching center (there may, of course, be multiple switching centers) may be located anywhere, in practice it will almost always be located at a node of the network. Once the objective of minimizing the total network link distance is agreed on, the mathematics of locating the switching network becomes fairly simple, although it is, of course, necessary to weight the volumes of traffic originating at each point (or, more accurately, passing through each node) in order to reach the optimum solution.

The Integrated Uses of Hardware in Network Configurations

Use of Front-end Processors

A front-end processor may be regarded as either a centralized concentrator or as a decentralized piece of the central processing unit or host processor. Clearly, from a communications link viewpoint, a centralized concentrator offers no economic advantages, so the justification of a front-end processor must depend on its ability to perform some of the work of the host processor more cost effectively than the host processor itself. The growth in recent years of the front-end processor clearly reflects that many manufacturers have been able to demonstrate that this objective can be achieved.

Terminals and Distribution of Intelligence

The key issue in configuring the network is how much processing will take place at the terminal and/or other distributed locations in the network and therefore how much workload will need to be transmitted over the network as a whole. Clearly, the installation of intelligent terminals will, in general, reduce the transmission requirements. Similarly, the installation of minicomputers as concentrators will tend to reduce the processing load on the host processor(s), the extent, in all cases, depending on the exact solution chosen.

Physical Arrangements of Terminals

As noted above, it is a typical situation to find the requirement for a number of terminals close together at a remote distance from the host processor, for example, several terminals may need to be located close together (forming a so-called 'cluster') at a factory some distance from the central data-processing facility. In these cases the alternatives of multi-dropping, a shared control unit, or the installation of a multiplexer or a concentrator should always be investigated. Which of these solutions is adopted will depend on the particular circumstances, but the more traffic there is to and from the terminals the more likely it is that the use of a concentrator will be found desirable.

SOFTWARE PACKAGES FOR NETWORK DESIGN

A number of software packages have been developed to assist in the planning of a network. These packages utilize simulation and modelling techniques and can be used to calculate reliability and response-time performance as well as an optimum network layout. The term 'optimum' should be used with care in this connection since the package will need to be supplied with projected traffic data, and the projection made will obviously only be accurate if the source data is correct. Not all packages, moreover, handle the full range of equipment options available, but within these limitations such packages can be a valuable aid. One final word of caution also needs to be stated: where these packages are used by vendors to 'help' the client design his network, the analyst must take particular care to ensure that the information supplied is both accurate and used as intended, since if it is only partially used or if unrealistic assumptions are made, the results may well be meaningless or, worse, misleading.□

The Categories and Functions of Data Terminals

Problem:

Two likely upcoming candidates for the over-used word morgue are "ubiquity" and "pervasive," along with a lot of other once useful words/phrases like "system" and "office of the future." But they are good, fulfilling descriptors for important trends in current information processing developments, so we will latch on to them quickly while they are still fresh.

Ubiquity (presence everywhere or in many places simultaneously) aptly qualifies the evolving data terminal scene, and pervasive (diffused throughout every part of) certainly characterizes the evolving data communications scene. Terminals are becoming ubiquitous because of drastically falling prices, rapidly rising capabilities, and cheaper/easier communications paths among other terminals, computers, and a myriad of other devices. So-called intelligent and genius terminals, which are two multifunction devices that can do practically anything a computer can do and cost almost as much, are at the top of the terminal hierarchy. At the other end of the scale are "dumb" terminals that cost little more than a color TV set. So ubiquity comes in many forms. Your problem as a user or potential user, is how to find exactly the right match between your applications and needs and the hundreds of different terminal products offered by the vendors . . . at the right price. This report won't give you all those answers, but it will help you to sort out the general product classifications and applications potentials and will give you a point of reference for the reports that follow in this segment.

Solution:

The most obvious trend in the human interface to a computer system has been the movement towards direct interactive communication between the end user and the system. This is evidenced by the substantial rise in the number of terminals of all kinds. A study by the Stanford Research Institute projects that the number of terminals in use will exceed five million in 1984 as shown in Figure 1. The hardware to accomplish this started with the teletype-like device and has proceeded to the widely used remote batch terminals and video terminals with keyboards. A wide variety of other interface media is also in what must be considered limited use, including some of the following:

- Voice response
- Speech recognition

From *Computer Technology Impact on Management* by George A. Champine. © 1978 North-Holland Publishing Company. Reprinted by permission of the publisher.

The Categories and Functions of Data Terminals

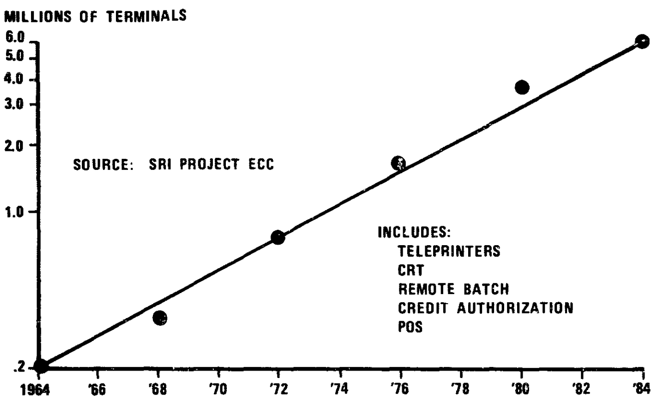


Figure 1. Projected growth of number of terminals in the U.S.

- Badge readers
- Bar code readers
- Graphics
- Hand-held terminals
- Keyboardless video terminals
- Computer output microfilm (COM)
- Magnetic ink character recognition
- Optical character recognition (OCR)
- Portable, acoustic coupled terminals
- Point-of-sale terminals

These interface devices will of necessity be the interface devices of the future. Said another way, the human interfaces of the future already exist, the only thing that will change will be the economics.

While hardware progressed from teletypes to remote batch and video terminals, software progressed from special interfaces, to batch systems, to time sharing, to transaction, to query, and now to tutorial systems.

Batch processing has served data processing for the 30 years of its history and continues to be a viable interface. However, there are many applications in which batch processing is a barrier to use of a computer, as evidenced by the over two million terminals in use today. The latest entry in the human interface to information systems is the word processor, and current trends suggest strongly that the word processing systems, the computer system, and the communications system may combine to form the integrated information system of the near future, also called the electronic office.

This report reviews the human interface technologies expected to be important in the 1980's, including data entry, hardcopy output, and portable terminals. The technologies of transaction systems, graphics, voice response, and speech recognition are discussed, and examples are given. The electronic office is described in terms of the interface of the future.

DATA ENTRY

The principal data entry device in use today is the same device used 20 years ago, namely the keyboard. Of course, the media has changed from paper tape, to cards, to magnetic tape, to diskette, to direct entry; and also the prompting has been improved greatly with new selection automatic cursor control, data validation, editing, and automatic provision of repetitive data. Although the productivity of data entry has been substantially improved, it remains largely constrained by the keyboard. Because of this, data entry remains very labor intensive—operator costs account for about 80 per cent of the total data entry expenses, and data entry accounts for more than 30 per cent of all EDP expense in a typical installation. These high costs are the result of a low level of automation (still being labor intensive rather than capital intensive) and a very stereotyped and unforgiving system interface which is far from the conversational mode desired.

The movement from key punches to CRT technology, including key-to-tape, key-to-disk, and direct entry, improved productivity substantially by providing immediate visual feedback and easy correction/editing. The ability to have immediate verification and error correction can significantly reduce cost and improve timeliness of data, sometimes improving turnaround by as much as 50 days.

The addition of even limited intelligence to the CRT terminal provided a significant upgrade in productivity by providing for:

- Error control—range checks and data type
- Format control—right/left justification, zero fill
- Local processing—COBOL or RPG
- Security—passwords, suppressed fields, suppressed update
- Programmed entry sequence
- Automatic cursor control
- Menu selection
- Fill-in-the-blank

The Categories and Functions of Data Terminals

- Table look-up
- Audit logs
- Communications
- File access
- Production statistics

The addition of intelligence to the CRT terminal has made it indistinguishable from the small business system used in an interactive environment; in fact, the hardware can be identical. The microprocessor has played the key role in making these functions available at an attractive price, and will also play a major role in future enhancements in functionality and price/performance. Other data entry devices have been available for some time, including magnetic ink character recognition (MICR), optical character recognition (OCR), and magnetic strip recording. Up to the present, these devices have not found widespread use, partly because of the admittedly high cost and partly because of user inertia.

A great deal of information exists in electronic form at the point of the transaction, for example, cash registers, gasoline pumps, credit card imprinters, and transaction terminals. It would seem that the evolution in the future must be to increase the electronic data capture at the source instead of continuing with manual transcription. Simple economics will be the motivating force to accomplish this change. To further identify the total magnitude of the data capture cost, it is estimated that gathering of data costs at least three times more than the actual data entry costs.

Of course, technology exists to solve these problems. Supermarket checkout systems using optically encoded product codes are available and are slowly gaining acceptance. Point-of-sale terminals also are gaining in acceptance in larger organizations. Portable bar code wand readers are now being used with some point-of-sale terminals to enter product codes and prices. Portable data collection devices are another way to obtain source data entry, and are increasing in use. By whatever means, a trend is clearly in process for the people originating the data to also enter it in the system, usually as a natural function of doing their job. The benefit of source data capture is to eliminate entering of the same data two or even three times as is sometimes done now. A result of this is the reduction in size and in some cases the elimination of a separate data preparation center.

Perhaps the most widespread near-term solution to the data entry problem will be the continued growth of on-line transaction systems. Here the data is captured at the source as a natural part of the transaction and can be used throughout the system. Point-of-sale systems are good examples of this type of system.

HARDCOPY OUTPUT

Relatively little has happened recently in the traditional high speed printer area of 2000 lines per minute. Conversely, there has been substantial development in the low speed area of 10 to 100 characters per second and in the very high speed area of 20,000 lines per minute.

The cost of low speed printers has not changed appreciably in the last several years, although the performance and reliability of low speed printers has improved substantially through the use of new mechanical approaches such as the matrix printer, daisy print wheel, and electrostatic technology. The benefitting areas have been hardcopy terminals and small business systems. It is now possible to obtain highly flexible correspondence document quality printing at more than 50 characters per second, or at much higher speeds when interfaced through photocomposition equipment.

The key to the substantial flexibility to these devices is the current practice of putting a considerable degree of functional capability into the reproduction device (variable line spacing, variable character spacing, and variable print font) and putting the intelligence to control this flexibility into the processor.

The new technology in the high performance area is the xerographic process applied to printers. The process provides a wide variety of functions at about 20,000 lines per minute, including the following:

- Self-generating forms from a mask
- Variable size characters
- Variable line spacing
- Variable type fonts
- Graphics

The xerographic approach can be expected to have continuing benefits as the functionality can be brought into the lower performance areas at an appropriate cost. It also can lead towards the 'intelligent copier', which accepts information in digital form from a variety of sources.

It is generally agreed that organizations today use too much hardcopy output, in the form of reports not wanted but generated because of inertia or "someone might want it". The following story illustrates that a little discipline can go a long way toward reducing unnecessary hard copy output:

A large software development organization was using hard copy output in the usual wasteful way, taking extensive storage dumps and generating voluminous listings. Cubic yards of waste paper were hauled out every day. When paper became hard to get in 1974, management insisted on reducing paper usage by using interactive techniques for program development, and taking dumps to mass storage for later study interactively. After much moaning and groaning and saying it couldn't be done, people reluctantly changed to interactive methods and the amount of paper used dropped substantially. Later when paper became easy to obtain and management eased off, the workers liked the interactive methods better than batch, and paper usage did not rise.

Although some applications clearly require hardcopy output such as billing and payroll documents, most organizations would be better off using less.

VIDEO, OR DISPLAY, TERMINALS

For years it has been forecasted that "the world is going on-line." Now it is actually happening, and the principal interface device is the video terminal. Video terminals are a very good match to the human interface because they can supply about as much information as a human can absorb at one time at an appropriate rate.

The CRT (Cathode Ray Tube, also known as a television screen) has long provided the technology vehicle for the video terminal, both for alphanumerics and graphics. Costs have been gradually reduced as production has increased and as LSI storage and logic have become available. Costs will continue to improve.

The CRT is normally associated with a keyboard; but, for nonexpert users, a touch panel on the screen is much more useful. It may also turn out to be more useful for expert users in some situations. Another new feature is the use of color, which is becoming quite popular in Japan, in ordinary commercial (e.g., hotel reservation systems) applications. Color, which has always been popular in command and control situations, is coming down in price to the point that it can be quite attractive.

The only technological competitor for the CRT in the video terminal area is the plasma display, which is a matrix of small illuminated gaseous discharge dots, usually 50 to 100 to the inch. Although present costs make it about three times more expensive than CRTs, it has several advantages—small physical size, useful life, self refresh, and easy graphics implementation. A manufacturing cost breakthrough would make it very attractive.

In the mid 1980's, most human interaction with computers will take place through video terminals (or telephones as described later.) This is because the video terminal is a very efficient way to use people, even though it is a very inefficient way to use computers; about 30 per cent more computational power is required to support an interactive interface than a batch interface. Yet, considering the cost of people versus the cost of computer hardware, this is a very favorable tradeoff.

PORTABLE TERMINALS

Portable computer terminals have been available for some time, and are slowly increasing in use. Their importance lies in allowing a worker to do his job away from the office at a very low cost.

Portable terminals come in two varieties:

- Hardcopy terminals
- Video terminals

Each kind is briefcase size and has an acoustic coupler that allows access to a computer over dial-up telephone lines. A number of companies now use these portable terminals to allow people who use a computer to work away from the office, usually at home. In one using company, workers can perform assignments at home but rarely stay away from the office more than a few days at a time because of extensive interaction required with other people. In another case, the workers are mostly working mothers who formerly were in-office programmers, but who now do all of their work at home and rarely go to the office.

Another variety of portable terminals on the market is the hand held terminal, which is similar to a hand calculator.

Given that slightly over half of the U.S. work force is engaged in "information handling" in the broadest sense, the consequences of the portable terminal in the long run could be very large if substantial numbers of people could work in a local or regional office, or at home. It could be possible to reduce rush hour traffic, reduce energy consumption, and reduce air pollution.

The most common current application of portable terminals is the portable data capture terminal, which is becoming widespread in the retail and distribution industry.

The Categories and Functions of Data Terminals

A portable data capture terminal consists of a hand held keyboard about the size of a calculator. It records information into a solid state storage or portable cassette recorder attached to the handset. The terminal usually incorporates local validity checks on the recorded data, such as field length, check digit verification, double key entry, and invalid code numbers. Information is either hand-keyed on the keyboard or recorded by passing a hand-held wand scanner over the bar code information. Bar codes generally accepted include Monarch Codabar, the American Universal Product Code (UPC) symbol, and the European Article Number (EAN) code. Information can be sent to the computer by acoustic coupler or modem.

In addition to the common use for taking inventory, portable data terminals can be used for capturing other data such as payroll, goods received, and branch orders. They can even be used as alternatives to point-of-sale terminals.

Savings are brought about through reduced inventory and faster stock turnover. Also, there is less handling of merchandise because it can be taken directly from receiving to the display shelves rather than warehousing. The payback period for this equipment is often about one year.

TRANSACTION SYSTEMS

Perhaps a credible view of the future of information systems can be obtained by looking at the existing data processing complex for Los Angeles County. The county is one of the largest governmental bodies in a single locality in the U.S., with a population of some seven million people and an annual tax revenue of $3 billion. Because of the relatively high concentration of people, the many social services provided, and continuing pressure to control costs, the handling of data has been automated to a very high degree on several transaction systems. This automation has been carried to the point that almost every involvement of a citizen with a government-supplied service begins and ends with a CRT terminal operator. This is true for people who are:

- Going to or leaving a hospital
- Receiving welfare
- Going to or leaving jail
- Applying for a drivers license
- Applying for unemployment compensation
- Being born
- Being married

Even in death, one must be processed by the CRT operator before being allowed to rest in peace.

As is typical of information systems, the total cost of this system is about one per cent of county revenue. The strongest advocates of an adequate EDP budget for the county are the other departments serviced by the EDP system because they see obvious cost saving by having the on-line information system available.

The following example of the welfare information system illustrates the magnitude of the available cost savings.

Welfare administration is a major activity for state and local governments across the nation. The County of Los Angeles has 1.2 million currently active welfare cases. Keeping information current on these cases is a large job because there are about 10,000 changes in status, address, and eligibility each day. Access to the data is required by 82 county social services office, seven hospitals, and the Department of Collections. Yet, timely access is needed to prevent people from getting welfare who are not eligible and to coordinate with the state medical insurance system.

To date, phase one of the three-phase implementation plan has been completed, saving the welfare department an estimated 500 staff positions. Chances for welfare overpayment have also been substantially reduced, and collections from the state medical insurance have been improved. When phase three is completed, it is estimated that staff requirements will be reduced by 900, and the system will save $10 million per year over its cost.

In another, even larger transaction system, the United States Social Security Administration is planning a very large, distributed data base system to provide interactive access to social security information on a individual basis. It is estimated that implementation of this system will so improve efficiency relative to present batch techniques that 20,000 job positions can be saved.

The Social Security Administration employs 85,000 persons at 1300 locations to make annual payments of $100 billion to 40 million beneficiaries. The present EDP system is batch oriented and centralized, using 23 large scale systems and about 50 smaller ones. Each system now maintains its own data, with interchange accomplished by tape. Over 400,000 reels of tape are required to support the present process including a one trillion (10^{12}) character master file. System failures now require 30-40 hours recovery time.

The Categories and Functions of Data Terminals

The present system is greatly in need of upgrading, but conversion is estimated at 50,000 man years. The approach that is currently being planned is to re-systematize the operation to distribute the data base and processing in order to regain control. The new distributed data system is planned to handle 100 transactions/second against a 100 billion character direct access data base, partitioned according to the location of residence of the beneficiary. The data base will be integrated with a whole person record for each client.

GRAPHICS

Graphics remain the best way to communicate large amounts of information between a computer and a human. A number of sophisticated graphics terminals exist, and the use of a minicomputer as part of the graphics system allows it to communicate with the host processor over a relatively low-speed data link. However, the number of graphics terminals in use is small compared to alphanumeric terminals because of their higher cost and specialization.

Perhaps a good indication of the future direction of graphics, if they are to achieve widespread application, is illustrated by the system developed by the Central Statistical Office of the British treasury. Here, a tutorial graphics system has been designed for use by nonprogrammer personnel in conjunction with an extensive and current data base on the British economy. The user directs the system by means of a hierarchy of menu selections to retrieve and present the desired data and then manipulates the data by means of time series analysis including:

- Smoothing
- Interpolation/extrapolation
- Correlations
- Normalization
- Scale expansion/contraction
- Multicycle function plots
- Parameter estimation

Data is normally presented in graphical form, although tabular data also is available. Input is provided through a light pen or keyboard.

A novice can use the system effectively with less than one hour of training, and yet the system is sophisticated enough that it is used on a real time manner during meetings to answer questions about what has happened in the British economy during the week.

VOICE RESPONSE

Computer controlled voice response systems have been available for some time, but the cost and effectiveness are now attractive enough that we may expect widespread application in the next few years. This is largely because it is now possible to purchase a voice response unit for approximately $10,000. The characteristic that makes the voice response technology so attractive is that it converts every telephone into a computer terminal, where the push buttons are used for input, and the computer controlled voice response is the output. With more than 120 million telephones in the U.S. and a comparable number in Europe, this is a powerful tool indeed.

A number of systems have been implemented using voice response technology. One such system is a catalog mail order system implemented by Simpson-Sears of Toronto. In this system, the customer dials the Simpson-Sears computer directly and then enters merchandise catalog numbers and other data requested by the voice response through the push-buttons to complete the transaction. For each piece of data entered, the voice response confirms the last input and requests the next input.

The system is able to reply with a variety of responses based on data in the system such as:

- "The price on this item has been reduced to ___"
- "This color is out of stock"
- "A special sale is on: you can buy a second one for ½ price"
- "This item is out of stock but a similar one of higher quality is available for the same price. The catalog number is __________ "

Another one of the largest catalog mail order companies is seriously considering a similar voice response system for catalog ordering and is now implementing a pilot project.

In a financial application, approximately 20 cities in the United States have banks with computer controlled voice response systems to provide account status and to allow transactions to be performed such as transferring funds from one account to another and paying bills at (typically) 1,000 different establishments.

Voice response systems are now widely used in telephone systems to provide information on wrong numbers and out-of-service numbers. In command and control applications, voice response systems

The Categories and Functions of Data Terminals

have been used to provide aircraft advisory information on local air traffic and weather, based entirely on sensor information.

Possibly the largest voice response system implemented to date is that of the Japanese National Railroad. Here, it is possible to book a reservation on any JNR train, directly through the JNR computer, entirely by telephone dial input and voice response.

For those telephone systems where push-button input is not possible, inexpensive tone generators are available that fit over the mouth piece and are keyed by a 12 button touch pad. These tones then are decoded by an appropriate modem.

A system block diagram for a voice response unit is shown in Figure 2. The voice response unit appears to the system as just another video terminal, and in fact can be attached to a terminal multiplexer along with other video terminals.

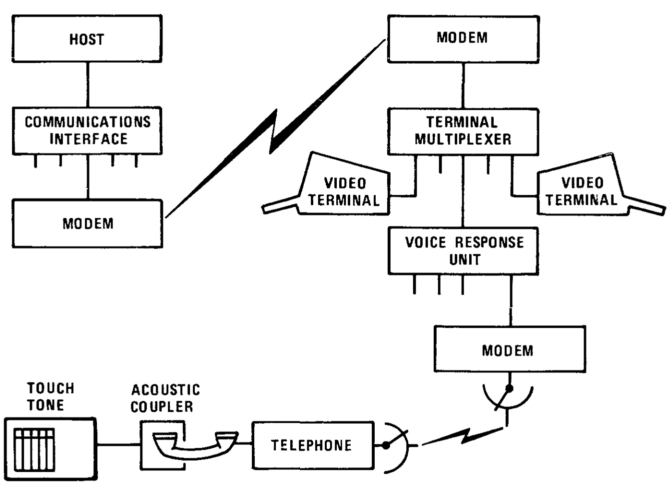


Figure 2. Typical voice response components

Because of the large number and low cost of the telephones (they are usually already available) and the effectiveness in making small amounts of information available, it would appear that voice response systems will become much more widely used in the future. The list of potential applications is very long and includes almost every kind of data stored in a computer. Perhaps one example will suffice: a company selling home computers now allows a purchaser to place his order for a home computer or expansion thereof via a voice response system.

VOICE INPUT

Another technology that has been in the laboratory for many years is speech recognition, or computer recognition of spoken words (not to be confused with voice recognition, which is computer identification of people). Although long claimed to be just on the verge of widespread usage, speech recognition seems finally to be coming into its own, and off-the-shelf products are now available from several vendors at prices that start at less than $200.

At the lower end of the range, the systems can accommodate a vocabulary of 30-50 words for a single speaker on an isolated word basis over a high fidelity direct wire or wireless connection. Higher priced units can accommodate a larger vocabulary and multiple speakers.

Many applications to date are justified on a functional, rather than a cost, basis where conventional data entry techniques cannot be used or are very clumsy. These applications include:

- Automotive inspection where both hands are required
- Inspection of metal thickness where both hands are required
- Data entry by workers wearing gloves
- Data entry in total darkness (darkrooms)
- Data entry on commodity exchange (where seconds are important)

However, the larger market for voice response appears to be as another way to implement source data capture. In all cases, a voice input function can always be performed by an operator at a keyboard. The ultimate impetus for widespread movement to voice input will be cost reduction, as is usually the motivation. Voice input is competitive with a single shift keyboard operator on a direct cost basis. On a two or three shift basis it is more than competitive. Other savings are available; data gathering cost is reduced by source data capture and training cost can also be lower.

Future improvements in semiconductor technology, described earlier, will reduce the cost of voice input system by 10-15 per cent per year and will provide larger vocabularies, multiple speaker operation, use of low fidelity or high noise environments, and phrase recognition.

Of course, there are many situations where voice input is precluded by the complexity of the application or where the operator must speak to a customer while using a terminal. Also, the cost of application development, which can easily exceed the cost of the voice input hardware, must be included in cost estimates along with the cost of and input verification device to insure correct entry. However, there are very likely many applications that could benefit from voice input now, and there will be more in the future due to cost reductions.

HUMAN INTERFACE OF THE FUTURE—THE ELECTRONIC OFFICE

Perhaps the natural next step of source data entry in the office environment will be the "electronic" office, which not only replaces current paper based technology with EDP methods, but also integrates the corporate information system into the office information system to form a natural flow of operational and strategic information to the end users.

The current cost of paper communications in the United States is immense. Some 65 billion pieces of first class mail are currently delivered annually. Eighty per cent of all mail is sent by business, and a substantial fraction of this is business letters with estimated total cost of $9.00 each. Also, some 15 billion long distance calls are made annually, again a substantial fraction of which are business calls. A near-term solution to the data communications problems is the use of facsimile, which having passed its 100th birthday can be considered a proven technology. However, rapid advances in cost and capability are being made, with a unit now on the market that can send a page in two minutes at a rental of $60 per month. This cost/performance ratio can be expected to improve even further as data communications improves.

The current use of paper to transmit information has worked well for 2,500 years, when it replaced clay tables, but now electronic methods offer advantages in several areas. Current, paper-based methods for transmitting information in the form of letters and memos are slow and expensive because these methods are labor intensive.

A block diagram of the components of an electronic office is shown in Figure 3. The users interface with the rest of the system via the CRT work stations. All communications are integrated in the local private branch exchange, and in fact all communications may be digital. The branch exchange may communicate with the common carrier telephone system, or through a satellite common carrier. The corporate data processing facility stores, processes, and controls transmission of all data.

An electronic office consisting of a CRT-based word processing/MIS approach with a digital communications system would solve the following problems:

- Memo/Letter Generation—The CRT allows fast generation, error correction, and recall of previously stored information.
- Mailing—The conventional mailing system would be placed by sending text information over a digital communications system to an "electronic mail box" (local mass storage) for each addressee. This would be much cheaper and would eliminate the current several-day time lag in mailing systems. Already, facsimile transmission is only half as expensive as sending a page of information via the U.S. Postal Service.
- Filing—A memo/letter can be filed in mass storage under several identifiers (e.g., date, topic, or customer), and then accessed by conventional data retrieval methods, thus eliminating the time, expense, and bulk of conventional filing-cabinet storage.
- Forms—Rather than having preprinted paper forms, the form will be held in digital form in mass storage to be called up on the CRT screen as required, and the blanks can be filled in at the screen.
- MIS Interface—Word processing interfaces are a simple, natural way to access conventional data base systems to obtain, for example, marketing reports, inventory data, shipping data, or financial summaries of various kinds.

A number of early products, especially in the word processing area, are now available as a first step toward the electronic office. A more advanced generation of products is now in design by several companies to provide pooling of work stations, more mass storage, correspondence-quality printers, and integration with the corporate MIS data base.□

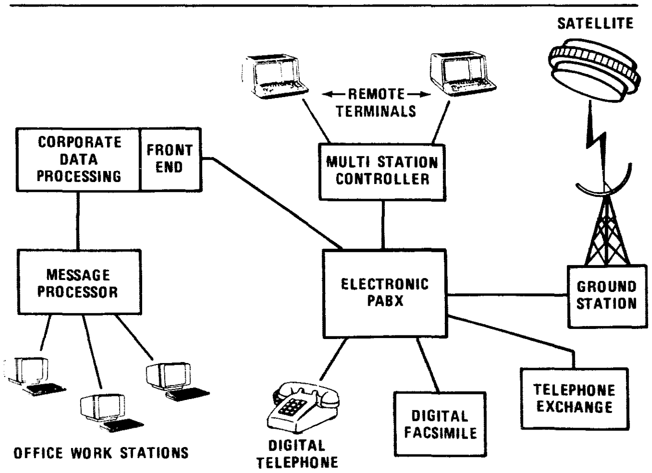


Figure 3. Electronic office system diagram

Typing and Printing at a Distance

Problem:

Occasionally we consider how difficult our job would become if the very convenient prefix "tele-" were to be excised from the English language. Tele- comes from an ancient Greek word, telos, which means afar or at a distance. We could probably get by with afarscope, or afarphone, or afarvision, but with a saddening loss of euphonious elegance.

Typing/printing at a distance is nothing new. It goes back to the earliest days of telegraphy. One of the first attempts to mechanize coded data transfers (as opposed to, for example, remote analog control of a stylus) grew out of Baudot's paper tape attachments to Morse code links in the 1870's. Baudot's next step was to develop a 5-level code to replace Morse code. In Baudot's scheme, each depression of a typewriter key produced a unique code that was transmitted electrically to another keyboard where the process was reversed to print the original character. This was the beginning of the ubiquitous Teletypewriter, many versions of which still operate with the century-old Baudot 5-level code.

Modern technology has affected the appearance and the circuits of the classic keyboard/printer terminal, but the basic constituents of a terminal-to-terminal connection are exactly the same now as they were in Baudot's time. Each terminal consists of a keyboard/printer with a local storage facility—paper tape in older equipment, some form of magnetic memory in modern equipment—and with a converter that translates each keystroke (character) into some kind of coded pattern in electronic form. The electronically encoded characters are transferred through an electronically compatible medium (wires, radio, microwave, etc.) to an identical receiver terminal. Each received character code is translated back into the mechanical world of solenoids, cogs, levers, and cams to select and activate a print hammer at the receiving terminal. The keyboard is bypassed at the receiving terminals, so we can safely and accurately telescope the whole keyboard/printer-to-keyboard/printer sequence into the single term "teleprinting."

This comprehensive report on the teleprinter market, currently served by more than 50 vendors, provides an incisive, succinct perspective of the teleprinter industry today. The report addresses the pros and cons of teleprinters versus alphanumeric display terminals, the tradeoffs between impact and non-impact printers, printer types and trends, and the makeup of the teleprinter industry and its anticipated growth. User

Typing and Printing at a Distance

experience, a vital aid to an intelligent decision, is also an integral part of this report: the experience of our subscribers with more than 14,700 terminals is reported in detail.

Solution:

Today's teleprinters feature a host of significant advances over their early predecessors. Modern teleprinters are available with a variety of printing techniques and a wide range of print speeds. What's more, they offer a variety of useful features such as programmable format control, adjustable forms control, upper and lower case printing, interchangeable character styles (fonts), bidirectional printing and paper feeding, selectable character and line spacing, additional keys such as a numeric keypad, status indicators, and portability. Of course, not all these features are to be found in any one teleprinter, but some vendors include most of them in their top-of-the-line models.

The microprocessor has found its way into teleprinters just as it has with most other types of terminals. Vendors have found that the magic device has substantially cut design, development, and production costs, and it easily lends itself to a variety of applications that can be implemented by either the vendor or user. What's more, the microprocessor precludes rapid obsolescence, since future applications can be implemented via reprogramming.

From the terminal user's point of view, the advent of microprocessor technology offers one major advantage: price. In the highly competitive terminal marketplace, cost savings resulting from implementation of microprocessor technology are often passed on to the customer.

Microprocessor-based programs (firmware) reside in ROM or PROM memory. ROM-resident programs, which are inexpensive when reproduced in large quantities, control those features which are permanent and unchangeable; while PROM-resident programs are typically produced in smaller quantities and implement customized or modifiable features. Either type can be replaced by simply removing the old chip and putting in a new one. This flexibility is highly beneficial to the manufacturer, since older equipment can be updated and nonstandard customer specifications fulfilled without costly hardware changes. Theoretically, program interchangeability might also benefit the user, but in practice it is doubtful that the requirements of a particular user will change often enough to make it a great advantage. The fact that PROM replacement generally must be done at the factory or by a field service technician precludes frequent PROM replacement.

In addition to controlling basic terminal functions, the microprocessor firmware can provide protocol emulation, define the character/code sets to be generated by the keyboard, implement special features, set control parameters, etc. Firmware specifications are generally determined at the time of order, and once the firmware is in place, execution is transparent to the user. Some vendors have predetermined programs from which to choose; a few permit the user to submit his own firmware specifications.

TELEPRINTER OR TUBE?

Teleprinters have traditionally been used as interactive terminals for two reasons: 1) they were the *only* type available before the CRT era, and 2) their costs (particularly for the Teletype models) were substantially below those of the early display terminals. However, cost is no longer the determining factor for selecting a teleprinter over a display terminal. With the introduction of the microprocessor, CRT terminal costs have plunged, and many of the so-called "dumb" CRT terminals are now available at substantially lower costs than teleprinters. For example, Teletype-compatible display terminals are currently available for as little as \$800 in single quantities and less than \$600 in quantities of 100 or more. Keyboard/printer terminals range upward from \$1,200, and are typically priced between \$2,000 and \$4,000. Printer mechanisms are more costly to produce than electronic components, and unless a new technique eliminates the printing and paper-movement mechanisms or new production techniques are implemented, teleprinter costs will typically continue to be substantially higher than those of basic display terminals.

Then why do teleprinters continue to constitute a large and viable segment of the interactive terminal market? Simply because there continues to be a strong demand for printed copy; some applications cannot survive without it. Some typical examples are messages or records that must be retained for reference, reports that must be distributed, program

Typing and Printing at a Distance

development, and unattended reporting (such as transmission after office hours, when rates are lowest). Although it is possible to attach a peripheral printer to a display terminal, it is generally less expensive to purchase a comparably-featured printer terminal.

Another important factor is portability. This aspect is important to a traveling business-person whose needs are satisfied by a small, light-weight, hand-carried terminal. A fairly wide selection of portable printing terminals in the 13- to 18-pound weight class is currently available from such vendors as Texas Instruments, Computer Devices, and Computer Transceiver Systems. Currently only one display terminal vendor, Digilog, makes a portable keyboard/display that is comparable in weight and price to the portable teleprinters.

Users who do not really need hard-copy output or portability should consider the numerous advantages of display terminals; Report CS15-150-101 presents a detailed discussion and survey of the current alphanumeric display terminals.

INDUSTRY PROFILE

The computer terminal market has been and will continue to be the fastest-growing segment of the computer industry. Estimates of its growth rate vary between 15 and 30 percent per year. However, industry forecasts predict that the teleprinter segment will be outpaced by other terminal sectors. This is not surprising, considering the meteoric growth of the display terminal market. Keyboard/printer terminals currently account for about 25 percent of total terminal installations; by 1983, it is predicted that the percentage will drop to about 16 percent.

This is not to say that the teleprinter market is dying. On the contrary, the industry has been extremely active in realigning its DP role, redefining traditional terrorities, and developing specialty markets that have not been penetrated by CRT's. The smaller, more specialized teleprinter market that is beginning to emerge is as active, competitive, and fast-moving as that of displays. And, as with most periods of adjustment, the process is causing some upheaval: acquisitions, dropouts, and new entrants are not uncommon among the participants. For instance, although the number of vendors listed in this year's report is the same as last year (54), 6 vendors whose equipment was presented last year have been dropped, and 6 new vendors have been added. This represents a turnover of only slightly fewer teleprinter vendors than display terminal vendors from 1978 to 1979. Even more to the point are the changes in the terminal models presented in the comparison charts: among the 47 manufacturers (excluding leasing companies presenting duplicate products) who were included both last year and this year, 104 models were offered last year. Of these only 83 (or 80 percent) are still being actively marketed, and 25 new models (or 23 percent of the current models) were added by these vendors during the past year. This compares to a retention of 78 percent of older models and a 27 percent rate of new model introductions for the display terminal market.

As you can see, the dynamics of the teleprinter segment are nearly as fluid as those of the display terminal sector of the terminal. A new generation of teleprinter terminals is slowly but steadily emerging from this upheaval. One really bright spot in this new generation is the portable teleprinter, which is becoming increasingly popular and is less susceptible to replacement by CRT's. Those teleprinter vendors who continue to compete with the glamour and sheer size of the display terminal market may find that their business is down. Those who cannot adjust to the fact that what was once the exclusive domain of the printing terminals is now dominated by the CRT's will slowly be weeded out. But those who strive for a new niche in the DP picture have a good change of succeeding.

The Industry Giants

The proven success of four teleprinter manufacturers—in terms of both endurance and volume—deserves special recognition. Teletype Corporation, Digital Equipment Corporation, General Electric Company, and Texas Instruments Inc. are responsible for the delivery of a combined total of more than 1.7 million teleprinter terminals.

Teletype Corporation, a subsidiary of AT&T, is the traditional patriarch of the teleprinter terminal industry. Its family of teleprinters has dominated the terminal market for more than a decade and has long represented the primary de facto standard which most other manufacturers emulate.

Teletype not only outranks the other three companies mentioned in longevity, but also (not unexpectedly) surpasses the others by a wide margin in productivity, with about one million units shipped. Many of these units are included in private teletypewriter networks of AT&T and Western Union, and are not usually included in installed base statistics. In December 1974, Teletype announced the production of its 500,000th Model 32/33 terminal, which it gold-plated to commemorate the event.

Teletype holds a unique position in the market that sets it apart from all the other terminal manu-

Typing and Printing at a Distance

Printer Characteristic	IMPACT		NON-IMPACT	
	Full Character	Dot Matrix	Electrothermal	Ink-Jet
Noise Level	Generally noisy except for daisy wheel machines, although some models can be equipped with noise-reducing cabinets or hoods		Quiet operation	
Maximum Print Speed	Ranges from 10 cps (Teletype Models 33 & 35) to 120 cps (GE TermiNet 1200/1232)	Ranges from 30 cps (several models) to 360 cps (Control Data 9318)	Ranges from 24 cps (Telpar Model PS-48C) to 50 cps (Computer Devices Models 1204/5/6)	60 cps (Siemens Model PT80 Ink Jet)
Character Formation	Provides typewriter-like printing preferred for many applications. Mechanical limitations on size of character set; large character sets (i.e. many symbols) cause reduced printing speeds	Dot matrix print quality depends on resolution: the most frequently used configuration is a 7-by-7 dot matrix; several vendors offer higher resolution up to 9-by-7 dots	Dot matrix print quality depends on resolution: the most frequently used configuration is a 5-by-7 dot matrix, a lower resolution than that available on many impact teleprinters	A 12-by-9 dot matrix resolution comes very close to the quality of full-character printing
Legibility	Generally good to excellent, but can vary widely depending on ribbon condition, number of copies, and mechanical adjustment		Lack of contrast between specially-treated paper and thermal-image characters generally decreases legibility	Generally good to excellent
Printed Copies	Permits simultaneous printing of multiple copies, generally up to 6-part forms		Prints original document only; multiple copies must be produced sequentially	
Paper Type	Uses ordinary computer paper; forms can be preprinted		Uses specially-treated blank paper, which cannot normally be preprinted	Uses ordinary computer paper
Paper Feed	Available with friction, pin, or adjustable tractor feed, vertical forms control, and other support for forms registration and specialty printing requirements		Generally available only with friction feed	Available with friction or pin feed
Physical Size	Generally medium to large desktop or pedestal-mounted units		Medium to small desktop units and compact portables	Desktop or pedestal-mounted units
Reliability	Varies widely depending on durability of the printhead and number of moving parts		Machine components are subject to lower mechanical forces and therefore less wear-and-tear	
Price Range	Prices generally start at about $3,000 for a basic KSR unit; fully-featured programmable ASR versions can cost up to $8,500. Exception: Teletype's Model 33 KSR is priced at just over $800; the Model 43 KSR, about $1,200	Prices generally start at about $1,500 for most basic KSR units; a fully-featured programmable ASR version can range upward to around $9,000	Prices generally start at $1,200 for a basic KSR; a fully-featured programmable ASR version can range upward to around $6,000	$3,000 to $4,100, depending on options

General characteristics of serial teleprinters

facturers. As a subsidiary of AT&T, it enjoys the advantages of a huge built-in market. Teletype equipment produced for AT&T's Bell System is available from Bell only as part of specific communications services. Teletype equipment is also available directly from Teletype Corporation, but on a purchase-only basis.

Because of its unusual market position. Teletype is forced to operate under several constraints. The two major agreements that dictate Teletype's market approach are a 1956 antitrust consent decree signed by AT&T and a 1971 agreement with Western Union upon the sale of the TWX network from AT&T to Western Union. The antitrust decree prevents AT&T from marketing anything that it does not use in its own communications network; therefore, Teletype cannot market a product until AT&T offers the product through one of its own services. The agreement with Western Union prevented AT&T from offering low-speed teleprinters unders its Dataphone services until April 1976. The May 1973 introduction of the Teletype Model 40 system, an impressive medium-speed CRT display terminal, turned out to be the ace up Teletype's sleeve. In November 1976, Teletype also updated its teleprinter line by introducing the Model 43, a thoroughly modern electronic unit.

The other three manufacturers are running neck-in-neck for second place:

Digital Equipment's popular DECwriter line consists of pedestal-mounted and desk-top impact printers. Since the first DECwriter was introduced in 1975, more than a quarter million of these durable teleprinters have been produced.

Typing and Printing at a Distance

General Electric's TermiNet family, with over 250,000 units installed, has grown steadily since 1969 when the TermiNet 300 was announced. The family includes a wide variety of printing terminals, including serial matrix and full-character teleprinters and low-to-medium-speed line printers that can be equipped with remote communications interfaces.

In December 1978, Texas Instruments celebrated the production of its 200,000th terminal. Its Silent 700 thermal teleprinters, including a portable unit and two models equipped with bubble memory, and its Omni 800 buffered impact printer terminals typify the "new generation" of teleprinters aimed at a broad range of specialty markets.

LEASING COMPANIES

Teleprinter terminals, particularly those produced by Teletype Corporation, Digital Equipment, General Electric, Texas Instruments, Diablo, and other large manufacturers, are available from sources other than the manufacturers. These additional suppliers are third-party leasing companies that purchase OEM quantities of the terminals from the manufacturer and lease the terminals to users.

Service and installation are usually provided by the leasing firm. Prime-shift service is generally included in the lease price of the terminals. Additional maintenance coverage may be available at extra cost. Cancellation of the lease is generally permitted on 30 days' notice. Teletype Corporation provides classroom instruction on the servicing of its equipment for the benefit of leasing firms that market its terminals.

Service and installation are usually provided by the leasing firm. Prime-shift service is generally included in the lease price of the terminals. Additional maintenance coverage may be available at extra cost. Cancellation of the lease is generally permitted on 30 days' notice. Teletype Corporation provides classroom instruction on the servicing of its equipment for the benefit of leasing firms that market its terminals.

In addition, a large number of these distributors also make minor, and in some cases major, equipment modifications or enhancements to provide a unique products or configurations not available from the original manufacturer. Those units with major modifications are shown in the accompanying comparison charts under the leasing company's name.

Nationally prominent leasing firms include RCA Service Company, a division of RCA; Western Union Data Services, a division of Western Union; Alanthus; Carterfone; ComData; and Data Access. A list of the full names and addresses of these leasing companies is provided for your convenience along with the list of vendors in Report CS90-120-101 in Volume Two.

A SUMMARY OF TELEPRINTER TECHNOLOGIES

The majority of today's teleprinter terminals employ serial printers, so named because they print one character at a time. Serial printers are grouped into two broad categories: those that mechanically strike or "impact" the paper to produce a printed image, and those that produce a printed image by some other means. Based on this key distinction, printers are generally classed as either *impact* or *non-impact* printers. Teleprinters using an impact printing technique can be further divided into two sub-categories: those that produce a "full-character" (typewriter-like) image, and those that produce a character image formed by a matrix of dots. Non-impact teleprinters currently form characters by the dot matrix configuration only, using either an electro-thermal or ink-jet printing method. The salient characteristics of these printing techniques are compared in the accompanying table of general characteristics on the facing page.

Impact Printing

Numerous teleprinter terminals are currently available that feature full-character impact printing. These terminals generally operate in the range of up to 55 characters per second, with the exception of the GE TermiNet 1200 and 1232, which can reach a speed of 120 characters per second. Among the more popular terminals in this class are the IBM 2740 and 2741, which contain a version of the ubiquitous IBM Selectric typewriter, the GE TermiNet 1200, the Teletype family of teletypewriters, and the Univac DCT 500, to name a few. Each of these terminals employs a different printing technique. IBM uses a replaceable "golf ball" print element that permits the operator to change type styles rapidly by snapping out the existing element and snapping a new one into its place. General Electric employs a moving type belt and a row of actuators, one per print position. Teletype, in its Models 33 and 38, uses a rotating cylinder that contains the type face and, in principle, operates much the same as the IBM Selectric typewriter. In its Models 35 and 37, Teletype uses a type block with type pallets embedded in the block; a single actuator is used. Univac uses a helical print wheel and throw-away cartridge ink roller.

The Diablo HyType, Qume Sprint, and Perkin-Elmer Carousel impact printers, because of their

Typing and Printing at a Distance

novel approach, represent a significant contribution to the serial printer industry and a challenge to the IBM Selectric printer. With fewer than 12 moving parts, these printers (equipped with stepping motors) are rated at 2 to 3½ times the print speed of an IBM Selectric. Printing can be performed in either direction and paper fed either up or down. Character and line spacings are variable, with up to 120 increments per inch horizontally and up to 48 vertically to permit proportional letter spacing or incremental plotting. The print element used by the Diablo and Qume printers is a flat disk with petal-like projections called a "daisy," while that of the Perkin-Elmer printer is shaped like a cup with finger-like projections. At the end of each projection is an embossed character.

The Diablo, Qume, and Perkin-Elmer printers offer good-quality printing at a low noise level, easily changeable type fonts, and higher speeds than most other full-character printers. Many terminal vendors have included these printer mechanisms in their products, saving them the trouble of developing a proprietary mechanism.

General Electric is another company that has developed a high-speed, full-character impact printer for use in typewriter-style terminals. GE's TermiNet 1200, a high-speed version of the successful TermiNet 300 terminal, employs a line printing approach to produce printed copy at speeds up to 120 characters per second. The TermiNet's printing arrangement consists of a type belt containing two symbol sets that move horizontally in front of a row of print actuators. This "chain printer" technique has also been adopted by Teletype Corporation in its Model 40 printer, rated at 296 to 416 characters per second.

The speed limitation on full-character impact printers served as the impetus for printer manufacturers to seek a different approach that would extend the upper limit of printing speed for serial impact printers. Their effort led to the development of the matrix printer, a compromise (though it has been a successful one) between decreased character quality and substantially higher print speeds that permits serial print rates up to 180 characters per second on a number of teleprinter models (a few are even faster).

The matrix type of impact printer produces a printed image formed by a rectangular matrix of dots, typically 7 dots high by 7 dots wide. Printing is performed by moving a print head containing a column of 7 pins across the paper and selectively actuating the pins at 7 successive intervals to form each character. Control Data has attained a speed of 360 characters per second with its Model 9318. The 9318, a receive-only teleprinter, uses two printheads that move bi-directionally along the same axis and in unison, so that each printhead travels just half of the paper width. The Facit receive-only Model 4540 achieves a rate of 250 characters per second, using a single printhead equipped with electromagnetically-controlled hammers instead of wire pins. Though they contain comparatively few moving parts, matrix printers are subject to an increased amount of wear within the print head as a result of the succession of pin movements required to create each character.

Matrix teleprinters are typically less expensive than similarly-featured full-character teleprinters. Especially considering the improved print quality now available with higher-resolution dot matrix printing, careful thought should be given to whether full-character printing is worth the trade-offs in price and speed.

One development that has tended to improve throughput in newer teleprinters is the "logic-seeking" (also called "smart" or "optimized") technique for printing received data. This technique utilizes a print buffer plus a bidirectional printhead. The "logic-seeking" feature seeks out the shortest distance (left or right) from one line to the next and eliminates the time that might be taken for a full carriage return. By utilizing this technique, the Centronics Model 761, for example, can sustain an average data throughput rate of up to 500 bits per second with an actual print rate of 60 cps. Although this technqiue is currently used primarily on impact teleprinters, it is likely that non-impact teleprinters will also feature it soon.

Non-Impact Printing

Members of the other basic class of teleprinters—the non-impact units—employ various electronic and chemical techniques to produce printed images. All the non-impact teleprinters currently on the market utilize dot matrix character formation. Some of the non-impact printing techniques have evolved from the development of facsimile communications; others were specifically developed for use in high-speed printing applications, where print speeds of better than 2000 lines per minute are not uncommon, or as low-cost alternatives to impact printing.

The electrothermal (or thermal) printing technique is the most commonly used of the non-impact techniques and is employed in terminals produced by Anderson Jacobson, Computer Devices, Computer Transceiver Systems, NCR, Texas Instruments, and Telpar.

The ink-jet technique, used in the Siemens PT80 Ink Jet teleprinter, was simultaneously and independently developed by A.B. Dick and by Teletype Corporation for high-speed printing applications. A stream of electrically charged ink droplets is sprayed onto ordinary paper to produce printed characters. Character formation is performed by electrostatic deflection plates that control the direction of the charged ink droplets, in much the same manner as the electron beam movement is controlled within a cathode ray tube (CRT). The ink-jet technique is relatively expensive and has a limited market potential, as indicated by the smaller number of units delivered. Production of ink-jet printers has been terminated by both A.B. Dick and Teletype, but IBM uses the ink-jet technique to produce high-quality printed output in some of its word processing systems.

Reliability of most non-impact printers is comparatively high because they have few mechanical parts; 3000 hours or better between failures is not uncommon.

There are some quiet environments where the noise of certain impact printers simply cannot be tolerated. The virtually silent non-impact printers are especially desirable in these locations.

The non-impact printers' ability to produce only one copy at a time might be a crippling disadvantage if you normally require several copies. But if you don't mind the additional time required to run off the needed extra copies on a nearby copying machine, the limitation of one copy may not be detrimental.

HOW TO ASSESS TELEPRINTER TERMINAL CHARACTERISTICS AND FEATURES

One of the toughest equipment preacquisition hurdles to overcome is getting to know what all the characteristics, features, and options mean and how they all relate to your needs. To help you on your way, we have collected the unfamiliar terms from many manufacturers' spec sheets and have grouped and explained them in the following.

Compatibility

Most of the communications terminals currently on the market are designed as direct replacements for other popular terminals. In the teleprinter terminal market, replacement terminals generally fall into four categories: those designed to replace a Teletype Model 33 or 35 teletypewriter, those designed to replace an IBM 2740 Model 1 or Model 2 Communications Terminal, those designed to replace an IBM 2741 Communications Terminal, and those designed to replace an IBM 3767 using SDLC protocol. Datapro included these four entries to define the category of *compatibility.*

Model Configurations

Teleprinter terminals are typically available in any or all of three basic *model configurations:* Receive only (RO), which includes a *printer only;* Keyboard Send Receive (KSR), which includes a *printer and keyboard,* and Automatic Send-Receive (ASR), which includes a *printer, keyboard, and a storage device* such as a punched tape reader and punch, a cassette or cartridge tape drive, a diskette drive, random-access memory (RAM), or the more recently introduced bubble memory. For many years, the conventional teleprinter ASR configuration always included a combined punched tape reader and punch because it was the only available low-priced storage device. But in more recent years, magnetic tape cassette and cartridge recorders have been replacing punched tape equipment on computer terminals as a result of quality components, decreasing prices, ease of use, and operating flexibility. The diskette or "floppy disk" also belongs in this category. RAM memory is becoming increasingly popular with the rising availability of large-capacity RAM modules at diminishing prices. Bubble memory, as introduced by Texas Instruments on its 763 and 765 Electronic Data Terminals, is a promising replacement for other forms of terminal storage in the future.

Some terminals provide an *auxiliary or second serial (RS-232C) interface* for attaching a user-supplied I/O device, such as a cassette or diskette unit.

Terminals that are designed to be hand-carried (usually in a suitcase-like enclosure) are noted in the entry *portable case.*

Significant General Features

Teleprinter terminals are available with a variety of potentially useful features and capabilities. No one terminal has them all, however, and some stripped-down economy models offer very few of them.

User programmability can be defined in different ways. Datapro defines the term as 1) operating under the direction of a user-created application program stored within the terminal, 2) operating under the direction of user-defined parameters that may be changed according to user needs, or 3) operating under the direction of a user-created data entry format or limited text-editing program. The *program loading technique* identifies the way in which user programs are entered into the terminal; e.g., via the keyboard, via an external device such as a cassette tape or diskette unit, or via downline loading from the host computer.

Typing and Printing at a Distance

The use of a buffer between the terminal and communications facility promotes communications economy through increased transmission speeds and enhances terminal flexibility through additional capabilities such as message editing prior to transmission. Buffering can be performed by input/output media such as punched or magnetic tape, and often is (e.g., in the Teletype ASR terminals). However, some manufacturers provide an internal buffer (usually composed of a semiconductor shift register), which is used to gather keyed or received data prior to transmitting or printing, respectively. The *line buffer capacity* in characters is presented where applicable.

Editing, by line and/or character, featured only on terminals that provide some form of buffering, allows the operator to correct data that has been erroneously keyed prior to transmission. Some terminals, such as those that include a punched tape capability, provide editing by character only. Those that contain an internal buffer, however, usually permit the entire buffer to be erased so that a line containing an error at the beginning can be quickly retyped instead of having to backspace character-by-character to reach the erroneous entry. On some of the more flexible terminals, such as those that contain dual cassette recorders, the editing facilities include the ability to update an existing tape. Keyed data can be merged with data read from the existing tape to produce a new, updated tape.

Parity checking and/or generation are important terminal features that safeguard the integrity of transmitted data. Some terminals only perform parity checking on received data, while others only generate character parity for each transmitted character. Still others provide both checking and generation. Many terminals allow the operator to select odd or even parity or to inhibit the parity functions.

Terminals that are designed to operate in a multistation environment (i.e., multidropped from a leased line) must include a *polling and addressing capability* so that computer messages can be directed to a specific terminal and terminal messages can be selectively transmitted to the computer; otherwise, the multidropped terminals would be required to contend with one another for the computer by "bidding" for use of the line.

The *automatic answer* feature permits the terminal to respond automatically to a call via the dial network from the remote computer. The terminal responds by readying itself to receive and print the incoming message.

Printer Characteristics

Printer *type* categorizes the printer as an impact or non-impact printer; *technique* specifies the printed character image as full character or dot matrix and describes the printing technique in a concise, simplified manner.

The total number of print positions in which the printer can print on each line is specified by the entry, *character positions per line.*

Print rate specifies the maximum rated printing speed of the printer in characters per second. Some terminals offer more than one rated printing speed to facilitate matching the communications characteristics of the remote devic. In most cases, manual selection is provided to switch among the available speeds.

Character set specifies the total number of print symbols provided by the printer. Typically, the character set is composed of upper case alphabetics, numerics, and special symbols including punctuation. *Lower case alphabetics* are usually available as standard or optional, however they are not required in many cases and tend to reduce printing speed. Where more than one character set is available, the entries distinguish between standard and optional sets.

Horizontal pitch defines the spacing between the centers of successive characters printed in the same line, and is presented in characters per inch. *Vertical spacing* defines the spacing between print lines, and is presented in lines per inch.

Forms feed specifies the type of paper-feed mechanism employed by the printer, usually as friction feed, pin feed, or tractor feed. Some terminals are available with more than one type, but typically offer pin feed or tractor feed as an option. Most non-impact printers feed paper without tractor or pin feed mechanisms.

Horizontal tabulation and *vertical formatting* facilitate control of the format of the printed output. In most cases, this level of sophistication is not required but it can be very helpful for registration of preprinted forms and other specialty printing jobs.

Features other than those listed in the standard comparison chart entries, such as split platen, bidirectional printhead, last character visibility, or low-paper indicator, are presented as *other features.*

Keyboard Characteristics

The style of *keyboard arrangement* defines the key/symbol relationships. There are two basic keyboard

Typing and Printing at a Distance

arrangements, typewriter and keypunch style. Teletypewriter keyboards, such as those provided with the Teletype terminals, can generally be categorized as typewriter arrangements. The keypunch arrangement is often referred to as a data entry keyboard. Some terminals are available with more than one keyboard style to permit the user to satisfy his particular need.

Character set refers to the total number of character codes and the code set that the keyboard is designed to generate. Each keytop symbol, represented by a corresponding bit pattern, is independent of its corresponding character code and can be interchanged with other symbols without affecting keyboard operation.

Keyboard *features* include such entries as numeric pad or character repeat. Some terminals offer these features as standard capabilities; others make them available as options only.

Transmission

Each teleprinter terminal contains a communications interface that enables communications between the terminal and the central computer site. *Mode* and *technique* define the operating mode and the method in which data is transmitted. There are three operating modes: simplex (transmission in one direction only), half-duplex (transmission in both directions, but not simultaneously), and full-duplex (simultaneous transmission in both directions).

Data is transmitted synchronously or asynchronously. Asynchronous transmission is characterized by the transmission of data in irregular spurts, where the duration of time can vary between successive transmitted characters; the transmission from an unbuffered teletypewriter is a good example. Synchronous transmission implies the transmission of data in a steady stream. Each transmitted character is clocked, and the time interval between successive characters is always precisely the same. The communications interface either provides clocking or accepts external clocking signals from the data set.

The transmission *speed* of the terminal is specified in bits per second and is usually limited by the speed of the printer or other I/O device unless the terminal contains an internal buffer. Buffered operation permits the printing to be performed at the rated speed of the printer, although the transmission speed may be much greater. Most teleprinter terminals are unbuffered due to cost considerations and therefore operate at low transmission speeds.

The transmission *code* refers to the bit pattern of the transmitted characters. The ASCII code is prominent and has been accepted as an industry and government standard; it is now by far the most commonly used code. Other transmission codes popularly employed by teleprinter terminals include Correspondence (a Selectric terminal code introduced by IBM) and two paper tape transmission codes, PTTC/BCD and PTTC/EBCD. Correspondence, PTTC/BCD, and PTTC/EBCD are all 7-level codes (including character parity); ASCII is an 8-level code, which also includes character parity. A few vendors also offer transmission using EBCDIC or Baudot code patterns.

The *unit code structure* specifies the total number of bits transmitted for each character. Asynchronous operating conventions require a single start bit and one or two stop bits to be combined with the character code for each transmitted character; therefore, an 8-level code such as ASCII is transmitted as a 10- or 11-unit code. Following Teletype's lead, the 11-unit code structure has been generally adopted for transmission at 10 characters per second; 10-unit codes are typically used at higher operating speeds.

Terminals that are capable of operating at more than one transmission speed typically feature *operator selectable speeds* via switch selection.

Transmission *block size* refers to the length in characters of a transmitted message. Unbuffered terminals transmit each character as it is keyed; therefore, the entry reads "character-by-character." Buffered terminals transmit data in multi-character blocks whose length is usually limited by the buffer capacity.

The terminal's *communications interface* generally meets the standard EIA RS-232B/C and CCITT specifications and connects to a modem or acoustic telephone coupler. Teletype terminals and their independent replacements are also available with a 20 or 60 milliampere dc current loop interface designed for use on telegraph-grade or private-wire facilities.

Some terminals contain an *integral modem* that can be connected directly to a communications line via a Bell System Data Access Arrangement. In some cases the manufacturer also provides an acoustic and/or inductive *telephone coupler* so that the terminal can be connected to a conventional telephone handset.

User Experience

To assess the current level of user satisfaction with the installed teleprinter terminals and to determine some usage patterns, a Reader Survey Form on Teleprinter Terminals was included in the February 1979 supplements to *DATAPRO 70* and *DATAPRO REPORTS ON DATA COMMUNICATIONS.*

Typing and Printing at a Distance

Terminal Supplier and Model	No. of User Responses	No. of Terminals in Use	Overall Performance					Ease of Operation					Keyboard Feel and Usability					Print Quality					Hardware Reliability					Maintenance Service				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
Anderson Jacobson—																																
830 Series	12	26	3.8	9	3	0	0	3.8	9	3	0	0	3.8	10	2	0	0	3.4	6	5	1	0	3.6	7	5	0	0	3.5	7	4	1	0
Other models & unspecified	7	31	2.9	2	3	1	1	3.3	2	5	0	0	3.0	1	5	1	0	2.9	0	6	1	0	2.7	0	5	2	0	3.1	4	1	1	1
Subtotals	19	57	3.4	11	6	1	1	3.6	11	8	0	0	3.5	11	7	1	0	3.2	6	11	2	0	3.3	7	10	2	0	3.4	11	5	2	1
Computer Devices—																																
Miniterm 1203	5	74	3.6	3	2	0	0	3.6	3	2	0	0	3.4	2	3	0	0	2.4	0	3	1	1	3.0	1	3	1	0	3.2	2	2	1	0
Other models & unspecified	3	12	4.0	3	0	0	0	3.7	2	1	0	0	3.7	2	1	0	0	3.0	1	1	1	0	3.0	1	1	1	0	3.0	0	3	0	0
Subtotals	8	86	3.8	6	2	0	0	3.6	5	3	0	0	3.5	4	4	0	0	2.6	1	4	2	1	3.0	2	4	2	0	3.1	2	5	1	0
Computer Transceiver Systems, all models	6	47	3.3	2	4	0	0	3.5	3	3	0	0	3.5	3	3	0	0	2.5	0	4	1	1	3.2	2	2	1	0	2.8	1	3	0	1
Diablo 1620	3	6	3.0	1	1	1	0	3.0	1	1	1	0	3.0	1	1	1	0	3.7	2	1	0	0	2.7	1	1	0	1	2.7	0	2	1	0
Digital Equipment Corp.—																																
LA 36 DECwriter II	50	472	3.4	22	25	3	0	3.2	15	29	4	1	3.2	18	25	3	3	2.9	7	30	11	1	3.5	27	17	3	1	3.1	18	18	7	3
LS 120 DECwriter III	16	67	3.8	12	4	0	0	3.6	9	7	0	0	3.5	8	8	0	0	2.7	4	9	3	0	3.4	6	10	0	0	3.2	5	9	2	0
Other models and unspecified	5	99	3.4	2	3	0	0	3.4	2	3	0	0	3.8	3	1	0	0	2.6	1	3	0	1	3.2	2	2	1	0	3.2	1	4	0	0
Subtotals	71	638	3.5	36	32	3	0	3.3	26	39	4	1	3.3	29	34	3	3	2.9	12	42	14	2	3.4	35	29	4	1	3.1	24	31	9	3
General Electric—																																
TermiNet 30	7	1,126	3.0	1	5	1	0	2.6	1	3	2	1	3.0	1	4	1	0	3.6	4	3	0	0	3.1	1	6	0	0	2.9	0	6	1	0
TermiNet 200 Series	4	10	3.0	0	4	0	0	2.3	0	1	3	0	3.0	0	3	0	0	3.3	1	3	0	0	2.5	0	3	0	1	2.7	0	2	1	0
TermiNet 300 Series	6	140	3.3	2	4	0	0	3.3	2	4	0	0	3.4	2	3	0	0	3.5	3	3	0	0	3.3	2	4	0	0	3.2	2	2	1	0
TermiNet 1200 Series	7	21	3.0	1	5	1	0	2.7	0	5	0	1	3.0	0	5	0	0	3.0	1	5	1	0	3.4	3	4	0	0	3.0	3	1	3	0
TermiNet, unspecified	5	32	2.8	0	3	1	0	2.5	0	3	0	1	2.5	0	3	0	1	2.5	0	2	2	0	2.5	0	3	0	1	2.0	0	1	1	1
Subtotals	29	1,329	3.0	4	21	3	0	2.7	3	16	5	3	3.0	3	18	1	1	3.2	9	16	3	0	3.1	6	20	0	2	2.8	5	12	7	1
Harris, all models	4	69	3.3	3	0	0	1	3.3	2	1	1	0	3.7	2	1	0	0	3.5	3	0	1	0	3.3	3	0	0	1	3.3	3	0	0	1
IBM—																																
2740/2741	10	101	3.1	3	6	0	1	2.9	1	7	0	1	3.2	1	8	0	0	3.2	3	6	1	0	3.1	2	7	1	0	3.3	3	7	0	0
3767	6	85	2.8	0	5	1	0	3.3	1	2	2	1	3.0	1	4	1	0	3.0	0	6	0	0	2.5	0	4	1	1	2.5	1	5	0	0
Other models	3	6	3.0	1	1	1	0	3.3	1	2	0	0	3.3	1	2	0	0	3.7	2	1	0	0	3.7	2	1	0	0	4.0	3	0	0	0
Subtotals	19	192	3.0	4	12	2	1	2.8	3	11	2	2	3.1	3	14	1	0	3.2	5	13	1	0	3.0	4	12	2	1	3.4	7	12	0	0
Teletype—																																
ASR 33	10	24	2.9	2	6	1	1	2.7	3	1	6	0	2.4	1	3	5	1	2.7	1	7	0	2	3.0	2	7	0	1	3.2	3	6	1	0
KSR 33	5	51	2.8	1	3	0	1	2.6	0	3	2	0	2.4	0	3	1	1	2.4	0	3	1	1	3.2	3	1	0	1	3.2	2	2	1	0
Model 43	18	578	3.7	12	6	0	0	3.4	9	7	2	0	3.3	7	10	1	0	3.2	5	11	2	0	3.7	13	5	0	0	3.5	8	3	2	0
Other models	7	3,512	3.0	1	4	1	0	3.2	2	3	1	0	3.0	1	2	1	0	3.3	2	4	0	0	3.5	3	3	0	0	2.8	0	5	1	0
Subtotals	40	4,165	3.3	16	19	2	2	3.1	14	14	11	0	2.9	9	18	8	2	3.0	8	25	3	3	3.4	21	16	0	2	3.2	13	16	5	0
Telex, all models	4	765	3.3	1	3	0	0	3.3	1	3	0	0	3.3	1	3	0	0	3.3	1	3	0	0	3.3	1	3	0	0	2.3	0	1	3	0
Texas Instruments—																																
733	6	296	3.5	3	3	0	0	3.2	1	5	0	0	3.5	3	3	0	0	2.7	0	4	2	0	3.3	3	2	1	0	2.8	0	3	1	0
735	4	33	3.3	1	3	0	0	3.5	2	2	0	0	3.3	1	3	0	0	2.5	0	2	2	0	3.3	2	1	1	0	3.0	0	4	0	0
743	4	505	3.0	0	4	0	0	3.3	1	3	0	0	2.8	0	3	1	0	2.8	0	3	1	0	2.8	1	2	0	1	2.5	0	2	2	0
745	18	122	3.4	10	6	2	0	3.6	11	6	1	0	3.2	7	8	3	0	2.8	5	4	9	0	3.3	10	4	4	0	3.3	8	6	3	0
Other 700 Series models	8	2,565	3.1	2	5	1	0	2.6	2	2	3	1	3.1	3	3	2	0	2.8	1	4	3	0	3.1	3	3	2	0	2.8	1	3	2	0
810	5	339	3.2	2	2	1	0	3.4	2	3	0	0	—	0	0	0	0	3.6	4	0	1	0	3.4	2	3	0	0	2.8	0	3	1	0
All other models	3	3	3.3	1	2	0	0	3.7	2	1	0	0	3.7	2	1	0	0	3.0	1	1	1	0	3.0	1	1	1	0	3.0	1	1	1	0
Subtotals	48	3,863	3.3	19	25	4	0	3.3	21	22	4	1	3.2	16	21	6	0	2.8	11	18	19	0	3.2	22	16	9	1	3.0	10	22	10	0
Trendata—																																
4000 Series	6	68	3.7	4	2	0	0	3.5	4	1	1	0	3.3	4	1	0	1	3.8	5	1	0	0	3.7	4	2	0	0	3.3	2	4	0	0
Trendwriter	3	3	3.0	1	1	1	0	3.7	2	1	0	0	3.7	2	1	0	0	3.3	1	2	0	0	3.3	1	2	0	0	3.3	2	0	1	0
Subtotals	9	71	3.4	5	3	1	0	3.6	6	2	1	0	3.4	6	2	0	1	3.7	6	3	0	0	3.6	5	4	0	0	3.8	4	4	1	0
Univac DCT 500	4	13	2.5	1	1	1	1	2.8	1	2	0	1	2.5	0	3	0	1	2.0	0	1	2	1	2.5	1	1	1	1	2.8	1	2	0	1
Western Union—																																
33 (Teletype)	3	34	3.0	1	1	1	0	3.0	1	1	1	0	2.7	0	2	1	0	3.0	1	1	1	0	2.7	1	1	0	1	2.7	1	1	0	1
35 (Teletype)	3	7	2.7	0	2	1	0	2.7	0	2	1	0	2.7	0	2	1	0	2.3	0	1	2	0	2.3	0	2	0	1	2.3	0	2	0	1
1200/1232 (GE)	3	1,765	2.7	1	0	2	0	3.0	0	3	0	0	3.0	1	1	1	0	3.3	2	0	1	0	2.3	0	1	2	0	1.3	0	0	1	2
Other models	7	1,266	2.8	0	5	1	0	2.8	0	5	1	0	2.7	0	4	2	0	2.5	1	1	4	0	2.7	0	4	2	0	2.3	0	4	0	2
Subtotals	16	3,072	2.8	2	8	5	0	2.9	1	11	3	0	2.7	1	9	5	0	2.7	4	3	8	0	2.5	1	8	4	2	2.2	1	7	1	6
Xerox 1700 Series	5	25	3.4	2	3	0	0	3.2	2	2	1	0	3.6	3	2	0	0	3.6	4	0	1	0	3.8	4	1	0	0	4.0	3	0	0	0
All Others	22	377	2.9	3	14	3	1	3.1	5	13	3	0	2.7	2	9	5	1	2.9	4	12	4	1	2.8	6	7	6	2	2.7	4	7	7	1
Grand Totals	307	14,775	3.3	116	154	26	7	3.2	105	151	36	8	3.2	94	149	31	9	3.0	76	156	61	9	3.2	121	134	31	14	3.0	89	129	47	15

Weighted Averages and Response Counts

LEGEND: Weighted Average (WA) is based on assigning weights of 4 to each Excellent (E) response, 3 to each Good (G) response, 2 to each Fair (F) response, and 1 to each Poor (P) response.

Users' ratings of teleprinter terminals

Typing and Printing at a Distance

By the editorial cut-off date of March 30, 1979, 307 usable responses had been received from 149 users, representing user experience with a total of 14,775 terminals. (Many users reported on multiple models and/or vendors.)

The ratings which the users assigned to the various models are shown in the accompanying tables. Subtotals by vendor are presented to make group comparisons easier. Weighted averages of the user ratings are also shown to simplify comparisons between models with dissimilar numbers of responses. Some of the models were rated by only a few users, and the results in these cases are presented solely for information purposes; it would be unwise to draw firm conclusions about these models from the small samples represented. For many models, however, the number of responses appears to be large enough to represent a valid cross-section of their users' experience.

Several questions were asked to determine usage patterns. The percentage results reported below are based on the total number of responses (307).

The users were asked to describe the characteristics of their teleprinter terminals. Their responses can be summarized as follows:

	Number of Responses	Percent of Responses
TERMINAL CONFIGURATION		
ASR	83	27
KSR	169	55
RO	24	8
TRANSMISSION MODE		
Half duplex	202	66
Full duplex	126	41
TRANSMISSION FORM		
Character	265	86
Block	19	6
PARITY		
None	118	38
Character	112	36
Block	13	4
PRINTING SPEED		
10 char./sec.	52	17
15 char./sec.	9	3
20 char./sec.	2	1
30 char./sec.	198	64
45 char./sec.	3	1
60 char./sec.	3	1
80 char./sec.	3	1
120 char./sec.	38	12
Other speeds	18	6

	Number of Responses	Percent of Responses
TRANSMISSION SPEED		
Printing speed	245	80
Higher	37	12
TRANSMISSION CODE		
ASCII	168	55
EBCDIC	38	12
IBM Correspondence	11	4
EBCD	4	1
Other	3	1
TRANSMISSION FACILITY		
DDD (dial-up)	215	70
Private line	74	24
Locally connected	66	21
X.25 (packet)	5	2
Other	9	3
TYPE OF MODEM		
Acoustic	129	42
Integral	66	21
Bell System	75	24
Locally connected	53	17
Other	26	9

Those users who utilize a storage option for their terminals were asked to provide the following information:

	Number of Responses	Percent of Responses
TYPE OF DEVICE		
Punched tape	25	8
Cassette tape	25	8
Diskette	3	1
Other or unspecified	12	4
SOURCE OF DEVICE		
Terminal vendor	50	16
Other vendor	8	3
APPLICATIONS		
Off-line message preparation	32	10
Editing prior to transmission	33	11
Data (file) storage	28	9

We also asked the users whether they planned to replace their existing teleprinter equipment in the near future, with these results:

	Number of Responses	Percent of Responses
Yes, with a CRT display	23	7
Yes, with another teleprinter	44	14
No	180	59

Typing and Printing at a Distance

When looking at the usage figures, keep in mind that they are based on the number of responding users, not on the number of terminals. Also, some skew is introduced because not all of the users responded fully. In addition, some users with multiple installations of the same teleprinter model provided multiple answers to some questions.

This year's survey results were remarkably consistent with last year's results, except for a decrease in the number of users responding. The profile we drew in 1978 of the predominant usage of teleprinters as operating over the public telephone network at 30 characters per second and using ASCII is still valid.□

Display Terminals—The Soft Copy Alternative to Teleprinters

Problem:

The key word in the evolving relationship between user and terminal is interaction, or perhaps more accurately, ease of interaction. You can interact with a teleprinter and even walk away with a hard copy of your conversation, but the hard medium of paper is not as friendly to the user as is the soft medium of a television screen. The modern nonprinting display terminal offers such an enormous range of on-the-spot editing, easy correcting, simple reformatting, and other features that you may want to seriously reconsider your hard copy requirements to take advantage of the features offered by display terminals. In fact, for a little more money you can have the best of both worlds by getting a hard copy option that you can turn on and off at the display terminal. Read this report carefully before you go out to buy your terminals, and remember that the price ranges we quote here may be a full 20%-30% lower by the time you're ready to make your acquisition.

Solution:

The introduction of the alphanumeric display terminal was a major milestone in the development of computer communications. Its inception in 1965 revolutionized the data communications environment, which had previously been the sole domain of the teleprinter. The display terminal market soon exploded with scores of vendors and a plethora of models in the wake of accelerated user demands.

Initially, display terminal prices were prohibitively high for many applications as a result of low-volume production and the material and assembly cost of discrete components. But significant technical achievements in the semiconductor industry drove display terminal prices down as terminals composed of discrete components gradually gave way to those containing ever-increasing amounts of integrated circuitry. Large-scale integration, involving the full use of integrated circuitry, was considered a major technical achievement that would cause a trend toward price stability in the industry.

But then a bombshell exploded in the semiconductor industry, driving terminal prices to incredibly low levels. The bombshell was the microprocessor—another milestone that is revolutionizing the whole computer industry.

From the terminal user's point of view, the advent of microprocessor technology offers only one advantage: price. In the highly competitive terminal marketplace, prices continue to plunge. In fact, the full capabilities of a unit can often be grossly underutilized and its cost still justified.

General Categories

Display terminals fall into one of three general categories: dumb, smart, and user-programmable. Naturally, there is some overlap between these categories.

Dumb terminals offer a limited number of functions; most feature Teletype compatibility.

Display Terminals—The Soft Copy Alternative to Teleprinters

Smart terminals offer extended functions, such as editing and formatted data entry. In some cases, the user can tailor the terminal to fit his own application via a limited degree of programming, such as format creation and parameter definition.

User-programmable terminals feature software support. The vendor typically provides an operating system, an assembler- or compiler-driven programming language, subroutines, I/O utilities, one or more protocol emulators, and one or two application programs, such as data entry and text editing. By definition, all user-programmable terminals have a user-accessible random-access memory (RAM) or equivalent—as little as 4K bytes or as much as 256K bytes or more. User-programmable terminals provide the user with the highest degree of flexibility and capability, permitting complete tailoring to the user environment. Moreover, the same programmable terminal can be tailored to different application environments within the same company, precluding the necessity to use a different terminal for each application. Some programmable terminals are available as turnkey systems, which means the program is furnished by the vendor and the terminal is ready to use as soon as it is installed; there is no need for the user to create his own programs before usage.

We have not identified a separate category of "intelligent" terminals because the industry does not exhibit a consistent correlation between the name and the device functions. Some "intelligent" terminals are programmed via factory-installed firmware and give the user no more capability to create programs than the "smart" terminals defined above. Other terminals marketed as "intelligent" are fully user-programmable.

But what about price? As usual, price is in proportion to capability. Dumb display terminals are the least expensive and typically range between $800 and $1,500 in purchase price for single quantities. Programmable terminals typically range upward from $4,000 and can become quite expensive when loaded with memory, additional display stations, and peripherals, such as diskette or disk storage and printers. Smart terminals are generally priced between dumb terminals and programmable terminals, with some overlap in both directions. (Naturally, added capabilities, such as program function keys and additional display stations, raise the price.) Quantity discounts available from some vendors can reduce per-unit costs, typically by 10 to 30 percent.

Some of the more prominent dumb terminals are those offered by Applied Digital Data systems (ADDS), Beehive, Hazeltine, Infoton, and Lear Siegler. Some of the more prominent smart terminals include the Hewlett-Packard 2640 series, the IBM 3270 Information Display System, the Racal-Milgo 40+ and System 400, the Teletype Model 40, and the Univac Uniscope terminals. Prominent examples of programmable display terminals include the ADDS System 70, the Beehive B 550, and the display terminal systems produced by Datapoint, Four-Phase Systems, Incoterm, and Sycor.

Microprocessor Mystique

Most of the display terminals introduced in the past two or three years are microprocessor-controlled; industry sources estimate that at least 50 percent of the installed display terminals will feature microprocessor control in 1980, as compared with only 10 percent in early 1975. Vendors have found that the magic device has substantially cut design, development, and production costs, and easily lends itself to a variety of applications. Moreover the microprocessor precludes obsolescence, since future functions can be implemented via reprogramming.

Microprocessor-based programs (firmware) reside in ROM or PROM memory. ROM-resident programs, which are inexpensive when reproduced in large quantities, control those features which are permanent and unchangeable; while PROM-resident programs are typically produced in smaller quantities and implement customized or modifiable features. Either type can be replaced by simply removing the old chip and putting in a new one. This flexibility is highly beneficial to the manufacturer, since older equipment can be updated and non-standard customer specifications fulfilled without costly hardware changes. Theoretically, program interchangeability might also benefit the user, but in practice it is doubtful that the requirements of a particular user will change often enough to make it a great advantage. The fact that PROM replacement generally must be done at the factory or by a field service technician precludes frequent PROM replacement.

In addition to controlling basic terminal functions, the microprocessor firmware can provide protocol emulation, define the character/code sets to be generated by the keyboard and displayed on the screen, implement special features, set control parameters, etc. Firmware specifications are generally determined at the time of order, and once the firmware is in place, execution is transparent to the user. Some vendors have predetermined programs from which to choose; a few permit the user to submit his own firmware specifications.

Display Terminals—The Soft Copy Alternative to Teleprinters

Display Media

Alphanumeric display terminals, the subject of this report, are designed principally to display messages composed of alphanumeric characters, although a limited graphic capability may be an added feature. Alphanumeric terminals are attracting most of the attention and generating most of the revenue in the current display market. Graphic display terminals account for only a small portion of the overall market.

Until recently the CRT (cathode ray tube) was virtually the sole means for displaying dynamic visual information, for business as well as entertainment use (in commercial television). Other devices for displaying information are now gaining popularity. These non-CRT devices include LED's (light emitting diodes, such as those used in calculators), plasma (gaseous) displays (such as the Burroughs Self-Scan panel), liquid-crystal displays, etc. But CRT displays still dominate the display industry, because they are still by far the cheapest method for displaying large amounts of data. Solid-state displays, such as those mentioned above, are currently limited to displays ranging from a few characters to a line of some two or three dozen characters (although Burroughs uses its Self-Scan panel to display up to 480 characters in a 12-line by 40-character format in its TD 730 unit.) Because of their prohibitive costs for displaying large quantities of data such as the 2000-character displays in current use, solid-state displays are still a long way from replacing the ubiquitous CRT.

Industry Profile

The computer terminal market has been and will continue to be one of the fastest-growing segments of the computer industry. Estimates of its growth rate vary between 15 and 30 percent per year. And within that market, display terminals—especially the user-programmable units at the high end and the dumb terminals at the low end—are moving most rapidly. According to optimistic industry sources, the terminal marketplace has barely been penetrated, the demand is "insatiable," and the potential is "seemingly endless." Predictions that, as office equipment, display terminals will soon become as familiar as telephones or typewriters do not seem unreasonable.

Excluding specialized terminals for dedicated markets such as brokerage houses, banks, and point of sale (POS), the alphanumeric display terminal industry has focused its attention on three principal markets: Teletype replacement, IBM 3270 replacement, and user-programmable terminals.

The most active of these three markets has been, and will probably continue to be, that of Teletype replacement, because it represents the greatest profit potential for the small terminal manufacturers.

Replacements for the IBM 2260 have long passed their peak of market penetration and are now offered by a few independents on an "as available" basis. The IBM replacement industry is now directed mainly toward 3270 replacements. Terminals based on microprocessors or minicomputers promise to capture both the IBM 2260 and IBM 3270 replacement markets by virtue of their software emulation capability.

As the emphasis on computing/communications networks and distributed processing become stronger, the need for user-programmable terminals continues to grow, both as a low-end expansion of the minicomputer market and as a high-end progression of the editing terminal market.

Display or Teleprinter?

If you are using teleprinter terminals for communications and plan to add to or replace your current units, be sure to consider display terminals as an alternative. Based on current prices, many display terminals can now compete with the lowest-priced teleprinters (even the ubiquitous Teletype units); the average unit price of display units continues to shrink as the result of microprocessor technology and high-volume production techniques.

In choosing between teleprinters and display terminals, you'll want to consider these factors:

- *Output medium*—Although in some applications the need for all output to be recorded in print is justifiable, in many cases it is simply custom or habit. Display terminals provide faster, more convenient access to required information, and can usually be equipped with auxiliary printers to produce hard-copy records of the displayed data when required.

- *Operating speeds*—Teleprinters are generally far slower in operation than their counterpart display terminals. Typical print speeds range from 10 to 30 characters per second for most teleprinters (though a few are capable of 120 cps or even high speeds). Typical display speeds range from 300 to 1200 characters per second.

- *Editing and formatting*—Teleprinters are designed primarily for message or data communications, and they generally do not provide sophisticated capabilities for data editing or formatting. When implemented, editing is usually limited to a single line (the one just keyed) unless the teleprinter includes a cassette tape unit, which can significantly enhance the edit capability but naturally increases the cost. Formatted data entry/output is featured on some printers, but again the cost is driven upward. Unless there is a definite requirement for printed copy, a display terminal will usually be the better buy.

- *Reliability*—Most of the current teleprinters are reasonably reliable devices, but like all mechanical devices they are subject to wear and misalignment. Display terminals offer generally higher reliability as a result of their totally electronic operation.

HOW TO ASSESS DISPLAY TERMINAL CHARACTERISTICS AND FEATURES

One of the toughest equipment preacquisition hurdles to overcome is getting to know what all the characteristics, features and options mean and how they all relate to your needs. To help you on your way, we have collected the unfamiliar terms from many manufacturers' spec sheets and have grouped and explained them in the following.

General Terminal Features

Display terminals are available in one of two basic terminal configurations: *stand-alone* and *cluster.* Stand-alone units are typically those that contain all components that support the operation of the terminal including display, keyboard, interface, and power supply within a single cabinet. Auxiliary units such as printers, cassette tape drives, etc., are usually external devices. Sometimes a stand-alone unit includes separate cabinets for terminal control and keyboard/display sections, and it may even include one or two separate displays. A cluster configuration typically includes a terminal control unit and a number of individual cable-connected keyboard/display units, which can often be located several thousand feet from the controller. In some cases, the vendor provides a multiplexer that accommodates a cluster of stand-alone terminals. A *local cluster* arrangement refers to a terminal that can be attached directly to a computer I/O channel and can operate as an on-line peripheral subsystem. A *remote cluster* arrangement refers to a terminal that is connected to the host computer via a communications facility. The size of a cluster arrangement is defined by the *maximum number of displays per controller.*

Some terminals are designed as direct replacements for other terminals. In the alphanumeric display terminal market, replacement terminals fall into four principal categories: those designed to replace an IBM 3270 and/or 3275, those designed to replace an IBM 2260 and/or 2265, those designed to replace a Teletype Model 33 and 35 teleprinter, and those designed to replace a Teletype Model 40 display terminal. Some vendors provide compatibility with *other* terminals such as those produced by Burroughs, Digital Equipment, Honeywell, and Univac. For example, no fewer than six vendors—Ann Arbor, Infoton, Datamedia, Dataview, Teleray and Ontel—are currently marketing units compatible with Digital Equipment's popular VT-52 terminal, and several more have plans to offer VT-52 compatibility in the near future.

Either of two types of compatibility may be offered: transmission compatibility or "plug-to-plug" compatibility. Transmission compatibility requirements include identical protocol, code and unit code structure, timing, asynchronous or synchronous operation, and transmission speed. Some vendors even provide identical cables, which is a cost-effective consideration in a local cluster environment. Most vendors with transmission-compatible units offer additional features and functions that the original vendor's equipment does not have, implemented via minor changes in host software. Units with true plug-to-plug compatibility not only have identical transmission parameters, but also identical features and functions; no alteration to host software is necessary, but no enhancements beyond the original vendor's equipment are available. For example, although numerous vendors offer IBM 3270 compatibility, only a few, including Courier, Memorex, Telex, Trivex, and Wordstream (Genesis One), make a true plug-for-plug replacement for the 3277 display station.

Programmability for processor-controlled terminals can be implemented via a combination of different techniques. *User-programmable* defines the capability for the terminal to operate under the direction of a user-created application program stored within the terminal. This requires the provision of an assembly-like language at the very least. Programmability via user-defined parameters or user-defined firmware refers to the use of fixed programs, such as a data entry program where the user defines field length and type, duplication, skipping, etc.

The entry *self diagnostics* denotes the terminal's capability to identify failures via self-generated test procedures. Failures are typically indicated by displayed text patterns, by indicator lamps, or by messages appearing on the 25th line of the display screen. Self-diagnostics are typically performed while the terminal is in the off-line mode.

Display Parameters

Printed information is generally arranged according to an orderly format consisting of a maximum

Display Terminals—The Soft Copy Alternative to Teleprinters

number of printed lines per page and characters per line. This orderly arrangement is also used to characterize the arrangement of data display on the face of a CRT screen or other display device. The electronic circuitry that produces the display image is designed to a specified set of parameters that define the capacity (i.e., the maximum number of display positions) and the display format (i.e., the maximum number of displayable lines and displayable characters per line). The most common display capacity is 1920 characters arranged in 24 lines of 80 characters. A few vendors, including Alanthus, Datagraphics, DEC, and ECD, offer 132-character display lines, which can eliminate the need to revise or patch software designed for standard 132-column printers or to maintain dual sets of programs for 80-column and 132-column output. Information is displayed in a rectangular area smaller than the total surface area of the display device. The factors that determine the required size of the display area are the display arrangement and the size of the displayable characters, which is normally a fixed parameter.

Symbol formation and the set of displayable symbols are functions of the character generator, which accepts coded characters (typically ASCII) from the computer and keyboard and converts them to a number of dots or strokes so that the form of the symbol or image can be displayed. In CRT's, characters are formed by a variety of techniques, including dots, strokes, starburst, or monoscope. The dot technique is by far the most popular. Each character is formed within a matrix of dots, and only those dots required to form the specific character are intensified. Typically, a dot matrix contains 35 dots arranged 7 dots high by 5 dots wide. Characters can be made clearer by increasing the number of dots within the matrix. The stroke technique forms characters by drawing short straight lines between specified points.

Solid-state display devices, such as plasma (gas) and LED (Light Emitting Diodes) are gaining popularity, but at present are generally limited to small display capacities consisting of a few characters. These typically form a character image in much the same way as a CRT display (i.e., via a dot matrix), though some form symbols through line segments.

Display arrangement, display medium, and symbol formation all have a great impact on display clarity. Test several units to decide which is easiest on the operator's eyes.

Attention can be drawn to vital information and different types of significant data can be visually separated by the use of the following display features:

- Color—characters or fields can be separated by color, which can also be used to identify conditions or types of data. A few vendors, including Applied Digital, Intelligent Systems, Megadata, and Terminal Data, offer up to eight colors as a standard feature; several other vendors offer a color option.

- Reverse video—displays a *negative* image of data, i.e., data normally displayed in white on a dark background is displayed in black on a white background. Characters or fields can be displayed in reverse video.

- Programmable brightness levels—visually separates different kinds of displayed information by displaying each type of a different intensity level, such as a fixed format and the entered data.

- Character and/or field blinking—vital information consisting of a single character or an entire field is blinked to attract attention.

Some terminals offer several of these display features, which can be combined to produce even more effective results.

Some applications require viewing more data that can be displayed at one time. The following features satisfy this need:

- Roll (or scroll)—this feature moves all displayed lines of data up or down by one line as a new line is added and an existing one removed. In some cases, the first line is linked with the last so that the data is rolled but not lost. Typically, data is lost as it rolls off the screen. This feature permits the user to scan through a volume of data to locate key information.

- Paging—this feature stores two or more frames or *pages* of data and displays any selected page.

Although roll and paging features can be software implemented in the host computer, the comparison chart entry applies to *only* those terminals that implement the features via hardware or firmware.

Many terminals provide the roll feature, but relatively few provide paging. Some provide both features.

The cursor marks the position on the screen where the next character will be read or written from memory. Cursor controls enable the operator to maneuver the cursor on the screen and facilitate the input and output of data. Typical cursor controls include:

Display Terminals—The Soft Copy Alternative to Teleprinters

- Move left (L)—moves the cursor one space to the left, which can be from the initial character position of a line to the last character position of the previous line if the terminal features wraparound.
- Move right (R)—moves the cursor one space to the right, which can be from the last character position of a line to the first character position of the next line if the terminal features wraparound.
- Move up (U)—moves the cursor to the same position on the previous line, which can be from the first line to the last line if the terminal features wraparound.
- Move down (D)—moves the cursor to the same position on the following line, which can be from the last line to the first line if the terminal features wraparound.
- Home top (H)—moves the cursor to the initial character position of the last line.
- Tab—moves the cursor forward to the next tab stop or backward to the previous tab stop (backtab).
- Return (RT)—moves the cursor to the initial character position of the next line; this is identical to the carriage return function of a typewriter.
- Backspace—moves the cursor one space to the left.
- Line Feed—moves the cursor to the same position on the following line.

Some cursors blink, others keep moving as long as the control key remains depressed. All cursors should be of the nondestructive type. Different manufacturers use a variety of symbols to indicate the cursor position on the screen. Some terminals also have *addressable/readable cursors,* which enable the position of the cursor to be written or read by the host computer under program control.

Most businesses use printed forms for daily activities such as billing, ordering, payroll, etc. Some CRT terminals can duplicate the printed form on the face of the screen, and data can be keyed into the blank spaces just as the typist enters data into a printed form. This "fill-in-the-blanks" approach to data entry requires a *protected format* capability. Display terminals that incorporate this feature treat the mixed format differently from keyed data. Field identifiers such as "name" or "salesman number" are protected from inadvertent key entry, and the data entry is confined to the variable fields (blank spaces) following the field identifiers. Some terminals automatically *tab* to the beginning of the next variable field immediately following the entry of the character that completes each field. The tab key is used where a field is partially filled.

Having completed entry into the fixed format, the operator transmits the data to the central computer. A feature called *partial screen transmit* promotes line economies by transmitting only the keyed data; the fixed format remains displayed and the "blanks" are erased for the next entry. This feature is also useful for transmitting only a portion of the displayed data such as a field, line, or block.

Editing features in a display terminal can consist of any combination of the functions listed below, although the best terminal for editing purposes would include all of them. Each function is performed with respect to the current position of the cursor. The desirable editing functions are:

- Character insert—the capability to insert a character into an existing line of displayed text; the remaining characters shift to the right or "spread" to accommodate the added character. The spreading capability may terminate at the last character position of the line or at the last displayable position on the screen. Data is lost when it is spread beyond the termination point.
- Character delete—the capability to delete a character from an existing line of displayed text; the remaining text closes up when the character is deleted.
- Line insert—the capability to insert a line of text into existing text; the text spreads to accommodate the added line.
- Line delete—the capability to delete a line of text from existing text; the remaining text closes up when the line is deleted.
- Erase—the capability to erase a character, line of text, message, field, or the complete screen. Most terminals include character erase and some form of display erase, which may erase the entire contents of the display, just that portion following the cursor location, or a combination of both functions. Line erase is optional in many terminals.
- Character repeat—enters a continuous sequence of symbols as long as the appropriate key remains depressed.

Display Terminals—The Soft Copy Alternative to Teleprinters

Keyboard Parameters

Keyboard *style* defines the general arrangement of keys; e.g., typewriter or data entry (keypunch) style. Data entry keyboards have a numeric keypad embedded in the alphabetic part of the keyboard which is accessed via a numeric shift. The *character/code set* refers to the set of symbols that appear on the keytops and, in many cases, to the actual character codes generated for each key depression, such as ASCII, EBCDIC, APL, etc. Some terminals are available with more than one keyboard style to satisfy particular user needs.

Keyboards that can either fit flush against the display or be located some distance away via cable connection are referred to as *detachable* keyboards. This feature provides increased configuration flexibility and operator convenience.

Some terminals are available with *program function keys*. These are special keys whose character codes are interpreted by the user's program. A function key is used to reduce the number of required input keystrokes to save time and reduce the number of input errors. Depressing one key could instruct the system to "sell one seat" or "call Chart A," for example.

A *numeric keypad* is a special keyboard feature that includes a set or block of 10 numeric keys, usually located to the right of the main keygroup. These numeric keys are arranged in an adding-machine format and are particularly useful for applications that require a high volume of numeric entries or arithmetic calculations.

Ancillary Devices

External I/O devices can add considerable flexibility to the applications possibilities for display terminals. A *cassette tape drive* or *diskette drive* can be used to store display formats, data to be transmitted, or user programs in the case of intelligent terminals. A *serial printer* provides hard copy when required.

These devices can usually be added to a terminal by the user via the terminal's RS-232 serial interface. The device is attached between the terminal and the external modem.

Although the above I/O devices are the most common, *other devices* can be and are used, such as industry-compatible 7- or 9-track magnetic tape drives, disk drives (cartridge or pack type), line printers, card readers, etc. Many units have an audible alarm which sounds whenever the operator's attention should be drawn to the prompting message area of the screen. Composite video permits multiple monitors to be attached to the terminal so that data may be viewed on more than one screen at the same time.

Transmission Parameters

Nearly every display terminal contains a communications interface that enables communications between the terminal and the central computer site. Mode and technique define the operating mode and the method in which data is transmitted. There are three operating modes: simplex (transmission in one direction only), half duplex (transmission in both directions, but not simultaneously), and full duplex (simultaneous transmission in both directions).

Data is transmitted synchronously or asynchronously. Asynchronous transmission is characterized by the transmission of data in irregular spurts, where the duration of time can vary between successive transmitted characters; the transmission from an unbuffered teletypewriter is a good example. Synchronous transmission implies the transmission of data in a steady stream. The time interval between successive characters is always precisely the same. The communications interface either provides clocking or accepts external clocking signals from the data set.

Communications protocol refers to the type of line discipline (control code sequence and control characters) that the terminal employs. The two most commonly used protocols are ASCII and IBM's binary Synchronous Communications (BSC) technique. IBM's latest protocol, Synchronous Data Line Control (SDLC), will be widely used in the future. Other large mainframe vendors such as Burroughs, Honeywell, and Digital Equipment Corporation (DEC) have produced their own communications protocols.

The tranmission *code* refers to the bit pattern of the transmitted characters. Two codes are prominent: EBCDIC and ASCII. The latter has been accepted as an industry and government standard, and is now the most commonly used code by display terminals.

The CRT terminal is a high-speed device that is usually capable of transmitting and receiving several thousand characters per second; however, it must run at a speed that is compatible with the communications system in which it is used. Most terminals are used on voice-grade facilities, which limit the transmission *speed* to a practical maximum of 4800 bits per second over the dial network and 9600 bits per second over leased or private lines.

Message *format* refers to the way data is transmitted, e.g., by block, by line, or by character. Terminals that are designed to be transmission-compatible with a Teletype unit transmit a character for each key depression. Buffered terminals transmit data in multi-character blocks. The line or block mode permits data to be composed and edited prior to each transmission and generally permits more efficient utilization of the communications facility. Some terminals offer manual selection between the modes.

Multipoint operation characterizes terminals that are capable of operating in a multiple-terminals-per-line environment such as that employed by the IBM 3270 and 2260/2265 display terminals. Basic to implementing this capability is the ability of a terminal to distinguish a control message intended for it alone. Polling invites the terminals to send data. Addressing informs the terminal that a message from the central computer is coming, so that it will be conditioned to receive. Central control of the message traffic is maintained by the central computer.

Auto answer refers to the facility for unattended operation on the dial network whereby incoming calls are automatically answered and messages are received without human intervention.

Auto call refers to the facility for unattended operation on the dial network whereby outgoing calls are automatically "dialed" and messages are transmitted without human intervention.

Display terminals usually have a *terminal interface* that meets the standards of the EIA RS-232B/C specification or some other standard interface and connects to an external modem or acoustic telephone coupler.

Some terminals contain an *integral modem* that can be connected directly to a communications line. In some cases the vendor provides an integral *acoustic telephone coupler,* so that the terminal can be connected to a conventional telephone handset.

A SUMMARY OF SELECTION GUIDELINES

In selecting a display terminal, as in acquiring most other types of computer equipment, your chances of picking the unit that's best for your installation will be far greater if you're willing to take the time to go about it in a systematic, logical way. The following selection procedure should help you get the maximum gain in computer throughput per dollar spent.

1. *Define the essential parameters* for a display terminal that will satisfy your needs. The entries in the comparison charts in this report, together with the text that explains them, can serve as a comprehensive checklist of the parameters and features that may be essential and/or desirable in your application.

2. *Find out who supplies the terminals* with the parameters and features you have selected. Use the accompanying comparison charts to determine which manufacturers produce terminals that appear to satisfy your needs. Then check the User Experience section of this report to see how users rate each manufacturer's products. You'll probably be able to narrow down the list of potential suppliers to a few firms that have demonstrated their ability to supply and service, at competitive prices, the specific type of terminal you need.

3. *Check the maintenance provisions.* Since maintenance is one of the key differentiating factors among the independent suppliers, you'll want to pay especially careful attention to this important area. Find out what organization supplies the maintenance service and learn all you can about it. Check the total size of the organization, the location and staffing of the closest service point, the promised response time for emergency service, the hours during which service is available, the nature and frequency of preventive maintenance, the size and location of the spare parts inventory, the procedure for handling engineering change orders, and the scope of the supplier's training program for his service technicians.

4. *Talk to users.* The terminals that appear most promising at this point should now be further investigated by conferring with present users. The users' ratings given in this report should serve as a first cut. Then ask each supplier for a list of customers. Be selective. Ask for installations similar to the one you're planning, at least with respect to communications discipline and number of terminals. And don't take no for an answer. Then, find out all you can from each user. Ask why he chose that unit, when it was installed, what problems were encountered in installing it, how many failures have occurred, how quickly they were corrected, and whether any incompatibilities have been detected. Finally, ask how he thinks the terminal or the associated support could be improved. The answers to these questions are likely to be enlightening, not only about the display terminals but about mainframe support as well.

5. *Choose the vendor and model.* By now, you should have all the information you'll need to choose the terminal that will satisfy your requirements at the lowest overall cost. If so, it's just about time to place your order.

6. *Negotiate a sound contract.* Now that you know which terminal you want, don't just sign the supplier's standard contract or order form. If you do, you're

likely to end up with a lot less security and support than the user who's willing to take the time and trouble to indulge in some old-fashioned haggling. What's more, you may even be able to shave some more dollars off the price tag.

USER EXPERIENCE AND RATINGS

To assess the current level of user satisfaction with display terminals, and to determine the patterns of usage of these terminals, Datapro conducted an extensive user survey. A Reader Survey form was included in the December 1978 supplements to DATAPRO 70 and DATAPRO REPORTS ON DATA COMMUNICATIONS and mailed to all subscribers. By March 1, usable responses had been received from 373 users with a total of 19,716 installed display stations.

Because many of the users reported on more than one model of display, the user replies generated a total of 619 responses or individual equipment ratings and profiles. The orientation of the users participating in the survey can be shown by the following table:

	Responses		Displays	
Responses on:	Number	Percent	Number	Percent
IBM displays	218	35	9,330	47
Other displays	401	65	10,386	53
Total	619		19,716	

Overall, the average number of displays per response was 32, while the average number of displays per responding user was 52.

The users were asked to rate the overall performance, ease of operation, hardware reliability, maintenance service, and software and technical support for each display by assigning a rating of excellent, good, fair, or poor. The resulting ratings for over 60 popular display models or families are summarized in the accompanying table. Any model or category that received more than two user responses is identified by manufacturer; models, categories, or manufacturers receiving only one or two responses were categorized as "other." Prospective buyers should note that the small sample sizes for some of these models make it unwise to draw firm conclusions from the indicated ratings.

To put the raw counts into a form more readily grasped, Datapro calculated a weighted average for each rating category. Each user response was assigned a weight of one, and the ratings were weighted on the conventional scale of 4, 3, 2, and 1 for excellent, good, fair, and poor, respectively. The data is presented as an additional information source, not as the final word on the worth of the displays represented. Individual vendor's ratings are tabulated on pages -110 and -111.

The ratings assigned by the responding users can also be combined to form this overall picture of current user satisfaction with the IBM displays, other manufacturers' displays, and all displays:

	Weighted Averages		
	IBM displays	Other displays	All displays
Overall performance	3.4	3.3	3.3
Ease of operation	3.2	3.3	3.3
Display clarity	3.2	3.2	3.2
Keyboard feel & usability	3.2	3.1	3.1
Hardware reliability	3.4	3.1	3.2
Maintenance service	3.2	2.8	3.0
Software & technical support	3.1	2.5	2.8
Number of responses	218	401	619

With the exception of IBM display clarity, which increased from 3.1 to 3.2, the ratings assigned by users of both IBM displays and other (non-IBM) displays were slightly lower or the same as in last year's survey. IBM clearly continues to maintain its traditional strengths in hardware reliability, maintenance, and technical support, but is nearly equal to other manufacturers in the remaining categories. IBM's rating decreased from 3.4 to 3.2 in ease of operation, the only category in which IBM now rates lower than other manufacturers. The increase in the IBM display clarity rating, combined with a 0.1-point decline in the display clarity ratings for other manufacturers, puts IBM on equal ground with the other manufacturers in this category. Both IBM and the other manufacturers dropped 0.2 point in the maintenance category, a significant comment from the users concerning the trend in field maintenance service.

The users were asked whether they were using their terminals as plug-compatible replacements for another vendor's terminals. Of the total 619 users responding, 279 were using IBM 3270, IBM 2260, Burroughs TD Series, Honeywell VIP Series, DEC VT-50/52, or Univac Uniscope terminals; 237 were using terminals made by another vendor to evaluate one of these or to emulate Teletype's Model 33/35 teleprinters; and 103 were using terminals made by another vendor but were not emulating any other terminal. The users not using one of the abovementioned terminals can be tabulated as follows:

Display Terminals—The Soft Copy Alternative to Teleprinters

Display Supplier and Model	No. of User Responses	No. of Displays in Use	Overall Performance					Ease of Operation					Display Clarity					Keyboard Feel and Usability					Hardware Reliability					Maintenance Service					Software and Technical Support				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
ADDS Regent 100	3	6	2.0	0	0	2	0	2.0	0	1	1	1	2.7	1	0	2	0	2.3	0	2	0	1	2.3	0	1	2	0	2.5	0	1	1	0	1.0	0	0	0	2
ADDS, other models	8	32	3.5	5	2	1	0	3.3	3	4	1	0	3.3	3	4	1	0	3.1	3	3	2	0	3.3	5	1	1	1	3.7	2	1	0	0	2.0	0	0	2	0
Subtotals	11	38	3.2	5	2	3	0	2.9	3	5	2	1	3.1	4	4	3	0	2.9	3	5	2	1	3.0	5	2	3	1	3.2	2	2	1	0	1.5	0	0	2	2
Beehive 100 Series	4	486	2.3	0	1	3	0	3.0	0	4	0	0	3.0	0	4	0	0	3.0	0	4	0	0	1.8	0	0	3	1	2.0	0	0	1	0	2.0	0	0	4	0
Beehive, other models	4	152	3.3	1	3	0	0	3.3	1	3	0	0	3.3	1	3	0	0	3.3	1	3	0	0	3.0	1	2	1	0	3.5	1	1	0	0	2.5	1	1	1	1
Subtotals	8	638	2.8	1	4	3	0	3.1	1	7	0	0	3.1	1	7	0	0	3.1	1	7	0	0	2.4	1	2	4	1	2.7	1	1	1	0	2.3	1	1	5	1
Bunker Ramo, all models	3	288	3.0	1	1	1	0	3.3	1	2	0	0	3.0	0	3	0	0	3.3	1	2	0	0	3.3	2	0	1	0	2.3	0	2	0	1	2.0	0	1	1	1
Burroughs TD 700/701	4	156	2.8	1	2	0	1	2.8	0	3	1	0	2.8	2	0	1	1	3.3	1	3	0	0	2.5	0	3	0	1	2.3	0	2	1	1	2.0	1	0	1	2
Burroughs TD 800/801/802	8	122	3.0	1	6	1	0	3.0	0	8	0	0	2.5	0	4	4	0	2.9	0	7	1	0	2.0	1	1	4	1	2.0	1	0	5	2	1.7	1	0	2	4
Burroughs TD 830	22	230	3.4	8	14	0	0	3.5	11	11	0	0	3.4	9	12	1	0	3.4	8	14	0	0	3.4	9	12	1	0	2.6	4	8	8	2	2.3	2	7	7	5
Subtotals	34	508	3.2	10	22	1	1	3.3	11	22	1	0	3.1	11	16	6	1	3.2	9	24	1	0	3.0	10	16	5	2	2.4	5	10	14	5	2.1	4	7	10	11
Computer Optics CO:77	6	129	3.2	1	5	0	0	3.5	3	3	0	0	2.8	0	5	1	0	3.2	2	3	1	0	3.0	1	4	1	0	2.7	2	1	2	1	2.2	1	0	3	1
Courier (ITT) E60/E65	3	321	3.7	2	1	0	0	3.7	2	1	0	0	3.3	2	0	1	0	3.0	1	1	1	0	3.3	2	0	1	0	3.3	2	0	1	0	3.3	2	0	1	0
Courier (ITT) E260	3	214	3.3	1	2	0	0	3.0	0	3	0	0	2.6	0	2	1	0	2.3	0	1	2	0	3.0	1	1	1	0	2.3	0	1	2	0	2.0	0	0	1	0
Courier (ITT) 270/2700/2750	24	1,206	3.3	9	12	2	0	3.5	13	9	1	0	3.5	11	12	0	0	3.2	8	12	2	1	3.0	6	13	2	2	2.7	5	10	5	3	2.7	4	6	10	0
Courier (ITT) 7700/7750	3	16	4.0	3	0	0	0	4.0	3	0	0	0	4.0	3	0	0	0	4.0	3	0	0	0	4.0	3	0	0	0	2.7	0	2	1	0	3.7	2	1	0	0
Courier (ITT), other models	3	31	3.0	0	3	0	0	3.3	1	2	0	0	3.0	0	3	0	0	3.0	0	3	0	0	2.7	0	2	1	0	3.0	1	1	1	0	3.0	1	0	1	0
Subtotals	36	1,788	3.4	15	18	2	0	3.5	19	15	1	0	3.4	16	17	2	0	3.1	12	17	5	1	3.1	12	16	5	2	2.8	8	14	10	3	2.9	9	7	13	0
Data 100, all models	6	80	3.2	1	5	0	0	3.3	2	4	0	0	3.0	0	6	0	0	3.3	2	4	0	0	3.2	2	3	1	0	3.0	1	4	1	0	3.0	1	4	1	0
Datamedia Elite 1520/1520A	5	142	3.6	3	2	0	0	3.8	4	1	0	0	3.8	4	1	0	0	3.6	4	0	1	0	3.4	3	1	1	0	3.0	0	4	0	0	2.0	0	1	0	1
Datamedia Elite 1521/1521A	3	153	3.7	2	1	0	0	3.7	2	1	0	0	3.7	2	1	0	0	3.3	2	0	1	0	3.3	2	0	1	0	3.0	0	2	0	0	2.5	0	1	1	0
Subtotals	8	295	3.6	5	3	0	0	3.8	6	2	0	0	3.8	6	2	0	0	3.5	6	0	2	0	3.4	5	1	2	0	3.0	0	6	0	0	2.3	0	2	1	1
Datapoint 3600/3601	6	104	3.0	2	2	2	0	3.5	4	1	1	0	3.5	3	3	0	0	3.3	3	2	1	0	2.7	1	3	1	1	2.8	2	1	3	0	3.2	3	0	2	0
Digital Equipment VT-50	4	21	2.8	0	3	1	0	2.8	0	3	1	0	2.3	0	2	1	1	3.0	0	4	0	0	3.0	1	2	1	0	1.8	0	1	1	2	2.5	0	2	2	0
Digital Equipment VT-52	11	63	3.1	3	7	0	1	3.4	4	5	2	0	3.1	3	7	0	1	2.8	3	5	1	2	3.1	4	5	1	1	3.0	5	2	1	2	2.8	2	5	0	2
Digital Equipment VT-100	3	13	3.3	2	0	1	0	2.7	1	1	0	1	2.7	1	1	0	1	3.0	1	1	1	0	3.5	1	1	0	0	3.3	1	2	0	0	3.5	1	1	0	0
Subtotals	18	97	3.1	5	10	2	1	3.0	5	9	3	1	2.8	4	10	1	3	2.9	4	10	2	2	3.1	6	8	2	1	2.8	6	5	2	4	2.8	3	8	2	2
Four-Phase IV/50	3	142	3.3	1	2	0	0	3.3	1	2	0	0	3.0	0	3	0	0	3.3	1	2	0	0	3.0	1	1	1	0	3.0	1	1	1	0	3.3	0	1	2	0
Four-Phase 5115	3	519	3.0	0	3	0	0	3.3	1	2	0	0	3.0	0	3	0	0	3.0	0	3	0	0	3.0	1	1	1	0	2.3	0	1	2	0	2.3	0	1	2	0
Four-Phase, other models	5	97	3.4	2	3	0	0	3.8	4	1	0	0	3.4	2	3	0	0	3.0	0	4	1	0	3.4	2	3	0	0	2.8	2	0	3	0	1.6	0	0	3	2
Subtotals	11	758	3.3	3	8	0	0	3.5	6	5	0	0	3.2	2	9	0	0	3.0	1	9	1	0	3.2	4	5	2	0	2.7	3	2	6	0	2.0	0	2	7	2
Genesis One (Wordstream) G77	22	299	3.6	13	9	0	0	3.6	14	8	0	0	3.7	16	6	0	0	3.3	9	12	0	1	3.5	12	10	0	0	3.0	8	11	3	0	3.0	3	13	3	0
Harris/Sanders 8170	3	70	3.7	2	1	0	0	4.0	3	0	0	0	4.0	3	0	0	0	3.3	1	2	0	0	3.7	2	1	0	0	3.0	0	3	0	0	3.0	1	1	1	0
Harris/Sanders, other 8000 Series	3	376	3.3	1	2	0	0	3.3	1	2	0	0	3.0	0	3	0	0	2.7	0	2	1	0	3.3	1	2	0	0	3.3	1	2	0	0	3.3	1	2	0	0
Harris/Sanders, other models	4	144	2.3	0	2	1	1	3.3	2	1	1	0	2.3	0	2	1	1	2.5	0	3	0	1	3.0	1	2	1	0	2.3	0	1	2	0	1.3	0	0	1	2
Subtotals	10	590	3.0	3	5	1	1	3.5	6	3	1	0	3.0	3	5	1	1	2.8	1	7	1	1	3.3	4	5	1	0	2.6	1	6	2	0	2.6	2	3	2	2
Hazeltine 1500 Series	6	112	2.8	0	5	1	0	3.0	2	2	2	0	3.0	2	2	2	0	2.7	1	3	1	1	2.2	0	1	5	0	2.2	0	3	1	2	2.0	0	2	3	0
Hazeltine 2000	8	80	2.8	0	6	2	0	3.3	3	4	1	0	3.0	1	6	1	0	2.8	1	4	3	0	2.4	0	4	3	1	2.3	0	4	2	2	2.5	0	5	2	1
Hazeltine Modular One	4	153	3.0	0	4	0	0	3.0	0	4	0	0	3.3	2	1	1	0	2.8	0	3	1	0	2.0	0	1	2	1	2.0	0	1	2	1	1.8	0	1	1	2
Subtotals	18	345	2.8	0	15	3	0	3.1	5	10	3	0	3.1	5	9	4	0	2.7	2	10	5	1	2.2	0	6	10	2	2.2	0	8	5	5	2.3	0	8	6	3
Hewlett Packard 2640 Series	10	53	3.9	9	1	0	0	3.3	3	7	0	0	3.8	8	2	0	0	3.7	7	3	0	0	3.9	9	1	0	0	3.8	7	2	0	0	3.4	3	4	0	0
Honeywell VIP 7700/7700R	5	134	3.2	2	2	1	0	3.2	2	2	1	0	2.3	0	1	3	0	3.6	3	2	0	0	3.0	2	2	0	1	3.4	2	3	0	0	2.8	0	4	1	0

LEGEND: Weighted Average (WA) is based on assigning a weight of 4 to each user rating of Excellent (E), 3 to Good (G), 2 to Fair (F), and 1 to Poor (P).

Users' ratings of alphanumeric display terminals

Plug-Compatibility with:	Number of User Responses	Percent of Responses
Teletype 33/35	51	15%
IBM 3270/3275/3277	131	38
IBM 2260/2265	17	5
Burroughs TD Series	9	3
Honeywell VIP Series	6	2
DEC VT-50/52	6	2
Univac Uniscope	1	0
Other emulations	16	5
No emulation	103	30

Clearly, the replacement of Teletype 33/35 teleprinters and IBM display units continues to dominate the replacement market segment, even though many displays having compatibility with other vendors' protocols have been introduced.

The users were also asked questions designed to determine the current usage patterns for display units.

Of the total 619 users, 371 reported having single display station configurations, and 291 reported having clustered display station configurations. (Users were counted in more than one category if they reported multiple types of usage.)

Single-station configurations can be summarized as follows:

	Remote Connection to Computer (via Communications Lines)	Local Computer Connection
Number of user responses	218	226
Percent of total user responses	35%	37%
Number of display stations	3,899	3,587
Percent of total display stations	20%	18%

Display Terminals—The Soft Copy Alternative to Teleprinters

Display Supplier and Model	No. of User Re-sponses	No. of Displays in Use	Weighted Averages and Response Counts																																		
			Overall Performance					Ease of Operation					Display Clarity					Keyboard Feel and Usability					Hardware Reliability					Maintenance Service					Software and Technical Support				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
IBM 2260	3	43	3.0	0	3	0	0	2.0	0	1	1	1	2.3	0	2	0	1	3.3	1	2	0	0	2.3	0	1	2	0	3.3	1	2	0	0	2.7	0	2	1	0
IBM 3275	26	839	3.3	9	14	1	0	3.3	10	12	2	0	3.2	8	13	3	0	3.1	5	16	2	0	3.3	9	14	1	0	3.4	11	11	2	0	3.2	7	12	3	0
IBM 3276	17	254	3.4	8	8	1	0	2.8	4	8	3	2	3.4	8	7	2	0	3.0	6	7	0	3	3.1	5	10	1	1	3.2	8	5	3	1	2.9	1	13	2	0
IBM 3277	116	6,032	3.4	52	57	5	0	3.4	49	59	8	0	3.2	34	66	16	0	3.4	53	54	7	0	3.5	61	50	5	0	3.2	40	60	15	1	3.1	27	69	16	1
IBM 3278	31	1,457	3.5	15	15	1	0	2.9	8	13	5	3	3.7	21	8	1	0	2.9	11	10	3	5	3.1	10	15	4	1	3.3	15	10	4	1	3.1	8	16	6	0
IBM 3270, other and unspecified	19	628	3.6	12	5	1	0	3.2	5	13	1	0	3.1	3	15	1	0	2.9	4	10	5	0	3.3	8	9	2	0	3.3	7	11	1	0	3.3	7	10	2	0
IBM, other models	6	77	3.8	5	1	0	0	3.8	5	1	0	0	3.7	5	0	1	0	3.7	5	1	0	0	3.7	5	1	0	0	3.2	3	1	2	0	3.2	3	1	2	0
Subtotals	218	9,330	3.4	101	103	9	0	3.2	81	107	20	6	3.2	79	111	24	1	3.2	85	100	17	8	3.4	98	100	15	2	3.2	85	100	27	3	3.1	53	123	32	1
Lear Siegler ADM-2	3	215	3.3	1	2	0	0	4.0	3	0	0	0	3.3	2	0	1	0	3.3	1	2	0	0	3.7	2	1	0	0	3.0	1	1	1	0	2.0	0	0	3	0
Lear Siegler ADM-3A	11	70	3.4	6	3	2	0	3.4	5	5	1	0	3.2	4	5	2	0	2.9	3	4	4	0	3.5	7	3	0	1	2.8	3	2	1	2	2.6	2	2	3	1
Subtotals	14	285	3.4	7	5	2	0	3.6	8	5	1	0	3.2	6	5	3	0	3.0	4	6	4	0	3.5	9	4	0	1	2.8	4	3	2	2	2.5	2	2	6	1
Memorex 1377	23	373	3.4	11	11	1	0	3.4	9	13	0	0	3.4	9	12	1	0	3.2	5	16	1	0	3.2	7	12	2	0	3.1	6	13	3	0	3.0	2	12	2	0
Microterm Act IV/Act V	3	18	2.3	0	2	0	1	3.3	1	2	0	0	3.0	0	3	0	0	3.0	0	3	0	0	2.3	0	2	0	1	2.3	0	2	0	1	2.3	0	2	0	1
NCR 796 Series	7	213	3.4	3	4	0	0	3.6	4	3	0	0	3.4	4	2	1	0	2.9	2	3	1	1	3.3	2	5	0	0	2.9	2	3	1	1	3.0	2	3	0	1
Omron 8025/8025A	5	79	2.8	0	4	1	0	2.6	1	1	3	0	3.2	2	2	1	0	2.8	1	2	2	0	2.4	1	1	2	1	2.0	0	2	1	2	2.3	0	2	1	1
Ontel, all models	3	99	3.3	1	2	0	0	3.3	1	2	0	0	3.7	2	1	0	0	3.3	1	2	0	0	2.3	0	1	2	0	1.5	0	0	1	1	2.5	0	1	1	0
Perkin-Elmer, all models	5	12	3.8	4	1	0	0	3.6	3	2	0	0	3.6	4	0	1	0	3.8	4	1	0	0	3.6	3	2	0	0	3.4	3	1	1	0	2.8	2	1	1	1
Raytheon PTS-100	5	188	3.0	2	1	2	0	2.8	1	2	2	0	3.0	1	3	1	0	2.4	1	1	2	1	3.4	3	1	1	0	2.8	1	2	2	0	1.8	0	1	2	2
Raytheon, other models	3	41	3.0	0	3	0	0	3.0	0	3	0	0	3.0	0	3	0	0	3.0	0	3	0	0	3.0	0	3	0	0	3.3	1	2	0	0	3.0	0	3	0	0
Subtotals	8	229	3.0	2	4	2	0	2.9	1	5	2	0	3.0	1	6	1	0	2.6	1	4	2	1	3.3	3	4	1	0	3.0	2	4	2	0	2.3	0	4	2	2
Soroc, all models	3	7	3.7	2	1	0	0	3.3	1	2	0	0	3.7	2	1	0	0	3.3	1	2	0	0	3.3	1	2	0	0	2.0	0	1	1	1	1.3	0	0	1	2
Sycor, all models	8	54	3.4	3	5	0	0	2.9	2	3	3	0	3.3	3	4	1	0	3.0	1	6	1	0	2.8	1	4	3	0	2.8	1	4	3	0	2.9	1	6	0	1
TEC, all models	4	63	2.8	0	3	1	0	3.3	2	1	1	0	3.5	2	2	0	0	2.8	0	3	1	0	2.5	0	2	2	0	2.5	0	2	2	0	2.7	0	2	1	0
Teleray, all models	6	58	3.6	3	2	0	0	3.6	3	2	0	0	3.2	1	4	0	0	2.6	0	3	2	0	3.2	1	4	0	0	3.3	1	3	0	0	3.5	1	1	0	0
Teletype 40 Series	12	246	3.2	2	9	0	0	2.9	1	8	2	0	3.2	3	7	1	0	2.6	2	3	6	0	2.9	1	9	0	1	2.8	1	8	1	1	2.4	0	5	4	1
Telex TC-277	15	525	3.5	8	7	0	0	3.4	5	9	0	0	3.3	7	4	3	0	3.5	8	5	1	0	3.4	5	7	2	0	2.9	3	6	5	0	2.7	2	5	6	0
Telex, other models	3	482	3.3	1	2	0	0	3.3	1	2	0	0	3.7	2	1	0	0	3.3	2	0	1	0	3.3	1	2	0	0	3.3	2	0	1	0	2.0	0	0	3	0
Subtotals	18	1,007	3.5	9	9	0	0	3.4	6	11	0	0	3.4	9	5	3	0	3.5	10	5	2	0	3.2	6	9	2	0	2.9	5	6	6	0	2.6	2	5	9	0
Trivex Plus 70 Series	8	17	3.6	5	3	0	0	3.3	2	6	0	0	2.9	2	3	3	0	3.3	2	6	0	0	3.5	5	2	1	0	3.3	4	3	0	1	1.7	0	2	0	4
Trivex, unspecified	3	109	3.3	1	2	0	0	3.0	0	3	0	0	3.3	1	2	0	0	3.0	0	3	0	0	3.0	0	3	0	0	2.3	0	2	0	1	1.0	0	0	0	1
Subtotals	11	126	3.5	6	5	0	0	3.2	2	9	0	0	3.0	3	5	3	0	3.2	2	9	0	0	3.4	5	5	1	0	3.0	4	5	0	2	1.6	0	2	0	5
Univac Uniscope 100	5	413	3.0	0	5	0	0	3.4	2	3	0	0	2.8	0	4	1	0	3.0	0	5	0	0	2.4	1	1	2	1	3.0	1	3	1	0	2.3	0	1	3	0
Univac Uniscope 200	8	159	3.0	1	6	1	0	3.4	3	5	0	0	3.5	4	4	0	0	3.0	2	4	2	0	3.3	3	4	1	0	3.1	3	3	2	0	2.6	2	2	3	1
Univac UTS-400	6	59	3.0	1	4	1	0	3.3	2	4	0	0	3.7	4	2	0	0	3.3	2	4	0	0	3.0	1	4	1	0	2.5	1	1	4	0	2.5	0	4	1	1
Subtotals	19	631	3.0	2	15	2	0	3.4	7	12	0	0	3.4	8	10	1	0	3.1	4	13	2	0	2.9	5	9	4	1	2.9	5	7	7	0	2.5	2	7	7	2
All others	40	742	3.1	11	21	7	0	2.8	10	23	6	0	3.0	10	19	10	0	2.9	9	19	8	3	3.0	12	15	9	2	2.7	10	13	6	7	2.5	5	11	11	4
GRAND TOTALS	619	19,716	3.3	243	318	44	4	3.3	234	316	51	8	3.2	227	304	72	6	3.1	198	317	70	20	3.2	231	274	81	20	3.0	177	255	114	40	2.8	102	253	137	49

LEGEND: Weighted Average (WA) is based on assigning a weight of 4 to each user rating of Excellent (E), 3 to Good (G), 2 to Fair (F), and 1 to Poor (P).

Users' ratings of alphanumeric display terminals (continued)

Cluster configurations are described below:

	Remote Connection to Computer (via Communications Lines)	Local Computer Connection
Number of user responses	184	168
Percent of total user responses	30%	27%
Number of controllers	1,164	444
Number of displays	7,132	4,744
Average displays per controller	6	11
Percent of total display stations	36%	24%

The final question in our survey concerned the strengths and weaknesses of the users' particular terminals. The following tables summarize their reactions.

	Number of responses			Percent of Total		
	IBM	Others	All	IBM	Others	All
Strengths						
Low cost	60	280	340	28%	70%	55%
Site restrictions	58	160	218	27	40	35
Physical size	46	177	223	21	44	36
Programmability	45	100	145	21	25	23
Weaknesses						
High cost	110	62	172	50%	15%	28%
Site restrictions	33	45	78	15	11	13
Physical size	73	87	160	33	22	26
Non-programmability	62	147	209	28	37	34

Thus, it's clear that the most important factor—assuming that a particular terminal meets the user's requirements in most other ways—is cost. And, as

Display Terminals—The Soft Copy Alternative to Teleprinters

compared with last year's survey results, the emphasis on cost is increasing. 42 percent of last year's user responses showed cost as an advantage compared to 55 percent this year. IBM's equipment is considered more cost-competitive by this year's users: 61 percent of last year's users felt that the cost of IBM's equipment was a disadvantage and 6 percent an advantage, compared with 50 percent (disadvantage) and 28 percent (advantage) this year. As in last year's survey, lack of programmability is considered a significant weakness; users clearly want more capability within the terminals they buy.□

User Programmable Terminals

Problem:

A terminal has been defined as a device which permits a human being to communicate with a computer and vice versa. The implication of this definition is that the computer can accept and process data and provide us with information we need, because it is a stored program machine. The terminal is simply the interface between a human being and the computer.

The idea of programming a terminal to process data or of a terminal's having its own random access storage device might be hard to accept, until you realize that since the above definition of a terminal was established, the physical size of a computer has shrunk hundreds of times and its manufacture has changed from the manual assembly of thousands of components, to a few tiny integrated chips which plug into a printed circuit board. It is entirely possible to enclose a computer inside of a terminal, and quite a few manufacturers have done just that. However the complete category of user programmable terminals ranges from fully portable teleprinter models, up to multi-station shared processor systems, with just about every configuration in between also included. This report describes how these machines evolved, how they vary in both capability and appearance, and what users think of them.

Solution:

The words programmable and terminal taken separately are basic to data processing and precisely definable. But, when put together to describe one class of equipment, the meaning becomes quite vague and even controversial.

In this report we present a discussion of Datapro's definition of user-programmable terminals; a review of the development of distributed processing equipment; a discussion of the basic characteristics of user-programmable terminals; consideration of the type of applications for which these terminals can be used; and the results of our user-programmable terminal user survey, including a summary of usage patterns and user ratings on equipment from over a dozen vendors.

THE DEFINITION CONTROVERSY

The single most difficult task involved in developing this report was the answering of a seemingly simple question: "What is a user-programmable terminal?" We started with this basic definition: a user-programmable terminal is any device that:

- Can communicate with a remote host computer; and
- Permits operator/host interaction via a keyboard/display or a keyboard/printer; and
- Permits local file updating; and

User Programmable Terminals

- Accommodates user-written applications programs.

As you can see, our definition provides a set of functional rather than physical characteristics. This point of view turned out to be surprisingly controversial. Numerous times throughout the preparation of the report, we found that our definition provided the focus for lively discussions with vendors who had different perspectives. In some cases, the definition as a whole conflicted with the traditional nomenclature a vendor used to identify his equipment; in other cases, one or another of the four qualifications was subject to various interpretation. In fact, only a small percentage of the vendors contacted actually market their equipment as "programmable terminals;" most others were not readily willing to classify their equipment as a user-programmable terminal and a number of vendors who *do* offer what they call "programmable terminals" were *disqualified* because their equipment does not fulfill our definition.

A summary of our discussions with the vendors serves as an excellent introduction to user-programmable terminals.

Take for example, our use of the word "terminal." A number of these devices are not ordinarily designated as terminals. In spite of the fact that this capability is offered, vendors frequently took issue with the use of the word terminal when applied to their minicomputers, key-to-disk systems, and other traditional classes of equipment. One vendor felt the solution might be to retitle the report "Remote Intelligent Workstations"; another suggested we should limit the report to "pure terminals" and exclude "hybrids", such as minicomputers equipped with communications interfaces.

In cases where whole classes of equipment might technically be qualified by our criteria, the decision to include or exclude a particular device in our vendor comparison charts was based on the practicalities of actual usage. We considered the frequency with which actual installed systems had been configured as user-programmable terminals, how strongly the vendor supports the capabilities, how actively the capabilities are marketed, etc. The process in these cases became one of elimination. For instance, of all key-to-disk systems that technically can accomplish the required functions of a user-programmable terminal, those in fact marketed exclusively as dedicated to a specific application (such as data entry) were generally not included. Minicomputers that provide a communications interface, but provide minimal vendor support for communication capabilities, were also eliminated. And, because full-blown computer systems can be programmed to do almost anything, including act like a user-programmable terminal, sheer size in some cases was used as a criteria for judging whether a device was beyond the scope of the report.

Controversy also developed over the meaning of the word "user-programmable." As we progressed, our basic "starter" requirement that a device be able to accommodate user-written applications programs evolved into a broader and more detailed set of guidelines by which we judged equipment proposed for inclusion in the charts. All of the devices selected have these elements in common:

- At least one programming language is offered, supported, and documented for development of programs by end-user customer personnel. In addition to high-level languages, such as COBOL, RPG, FORTRAN, and user-oriented vendor-proprietary languages, low-level assembly languages and machine languages (which generally require more programmer sophistication) were acceptable as long as the vendor provided the customer with documentation on how to program the device.

- Programs can be entered and revised via the terminal keyboard (though support for other methods may also be provided). We also included a few specialty terminal devices for which programs are developed by the user on a separate companion device for program development, as long as the terminal itself executes and stores the finished program, and programs are easily interchanged by the user.

- Programs can be stored and/or compiled at the host end, but must be executable locally at the terminal.

- Users are not restricted by the programming language to specific types of applications. For instance, we eliminated terminals with programming languages that supported only one application, such as data entry, or that permitted the user to "program" by choosing modules or parameters from a pre-existing set of programs. In addition, the programming language must support all of the operations generally expected for general purpose programming, such as arithmetic functions (add, subtract, multiply, and divide).

As a result, a number of devices marketed as "programmable" did not meet Datapro's criteria for user-programmability.

As we progressed in the development of the report, we began to accept the fact that few devices can be classified purely as user-programmable terminals, and that it was important to present products that matched our

User Programmable Terminals

definition regardless of traditional classification. As you will see, we have gathered together for your consideration a highly diverse collection of equipment that can function as user-programmable terminals under a broad range of circumstances. It is only you, as the prospective user, who can finally judge which devices in the group are suitable for the requirements of your particular application.

TERMINAL SYSTEM PERSONALITIES

As a group, user-programmable terminals tend to have rather unique "personalities," with stronger emphasis placed on some characteristics than on others. Like snowflakes or fingerprints, no two are exactly alike. A particular terminal system may offer an elaborate package of emulators and other communications features, but minimal off-line processing capability. Another may provide a wide range of applications software, but offer only basic batch transmission. A third may support an elaborate multi-user/multi-tasking system, but focuses software support on a specific application like data entry or word processing.

It is especially important, therefore, in choosing a user-programmable terminal that you take the time to consider the device's overall suitability to your current and future needs. Here are some questions you might want to ask yourself before you buy:

- Is the memory size large enough to handle the applications you intend to support?
- Is the level of software support provided by the vendor compatible with the level of sophistication of the persons on your staff who will be programming the terminal? (You might also check the feasibility of purchasing software from outside contractors.)
- Does the system have enough configurational flexibility to support system expansion if your initial needs increase?
- Are the peripheral devices easy to operate and featured to suit your needs?

If the system's "personality" is not quite compatible with your needs, continue shopping!

CHECK THE ANCESTRY

One of the strongest factors influencing a terminal system's "personality" is the source from which it has evolved. Only a few manufacturers' basic product lines began with user-programmable terminals; Datapoint, Incoterm, Raytheon, and Sycor might be considered representative of this group. Though their product lines have broadened to include other categories of equipment, the traditional orientation of their equipment probably comes closest to representing the classic user-programmable terminal. Most user-programmable terminals, however, find their roots in a manufacturer's other traditional product lines, and they tend to inherit the traits emphasized by those lines. For instance, the user-programmable terminals offered by traditional key-to-disk data entry equipment manufacturers, such as Inforex, Nixdorf (Entrex), and Mohawk generally incorporate those manufacturers' expertise in multi-station systems and strong data entry software support. The user-programmable terminals of manufacturers traditionally associated with minicomputers, such as DEC and Hewlett-Packard; CRT displays, such as ADDS, Beehive, Computek, and Ontel; and teleprinters, such as Texas Instruments and Computer Devices to some degree reflect the basic orientations of their respective roots. It is therefore important that you consider the source of your equipment, since that will determine, at least to some degree, the direction and emphasis of the support you will receive.

On-line Capabilities

Terminal communications tend to fall into two basic categories, distinguished by the types of functions to be performed: interactive and batch. Terminals supported for interactive communications permit engagement of the host in conversational sessions for inquiry and response, interactive data entry, and other applications that require back-and-forth interplay between the terminal and the host. Batch communications permit either of two capabilities: remote job entry (RJE), in which the terminal acts as a remote "console" for the central site computer to initiate execution of applications programs; and remote batch data communications, in which the terminal acts as a remote input/output device within an application program being performed at the host end.

Many user-programmable terminals are technically capable of performing any of these functions to one degree or another. However, the actual usage pattern for a particular terminal depends on what hardware components and software support are provided by the vendor and/or the user. For instance, for a device that is configured to be used as a remote batch terminal, only a small console display, with a screen capacity of 8 or 16 lines of 32 characters, may be provided for terminal/host interaction. Although the display is perfectly adequate for job initiation, status requests, and other conversational system-to-system communication, it is not intended for the more elaborate data entry functions generally associated with conversational terminals. By the same token, an interactive terminal programmed to communicate only in character-by-character mode may not provide the

User Programmable Terminals

most efficient means of transmission of large masses of data.

OFF-LINE PROCESSING

The key to *user-programmable* terminals is their off-line processing capabilities. All user-programmable terminals have user-accessible main memory for program execution. The memory capacity provided by minimum user-programmable configurations ranges widely, from as little as 1K bytes to as much as 128K bytes; when fully expanded, the upward size limit ranges from 16K bytes to 500K bytes. A typical system starts with about 32K bytes and is expandable to between 64K and 128K bytes.

User-programmable terminals also feature software support. The vendor, at the very minimum, normally provides a basic operating system and a language assembler, interpreter, or compiler. Depending on the amount of emphasis placed on software support, the vendor may also provide debugging and diagnostic subroutines; I/O utilities; additional programming languages; communications emulators; and general purpose (e.g., text editing or data entry) or industry-specific (e.g., payroll, general ledger, accounts receivable) applications programs. Some programmable terminals are available in turnkey configurations, which means that all programming necessary for a particular application is furnished by the vendor and the terminal is ready to use as soon as it is installed.

Most (though not all) user-programmable terminals also provide for local mass storage via a wide variety of vendor-supplied and supported peripheral devices such as floppy, hard, or cartridge disk or half-inch, cassette, or cartridge magnetic tape. Interface ports may also be available on the system for attachment of customer-supplied devices, though software supporting these devices may have to be written by the user.

With main memory, an operating system, a user programming language, mass storage, and file handling capabilities, these devices sound more like small computer systems than terminals, don't they? It's true. Many of these systems can function as stand-alone data processing units, independent of host support. And those that can't are still capable of operating upon a file to perform updating, computations, and other data processing tasks.

Whether users at the local site are actually participating in the programming of their terminals and autonomously directing the full processing operations performed by these devices is another question. In the user survey completed for this report, we found that although 77 percent of the users used programs written by persons on their own staffs for some or all of their processing, 26 percent of all programs were down-line loaded from the host computer. In addition, it is possible that diskette- and cassette-stored programs are being duplicated and distributed from a central site. And a large number of companies use vendor-supplied programs or software purchased from independent consultants. Maintenance of a full-time programming staff at each of multiple terminal sites seems unlikely to be efficient for all but the most elaborate terminal systems, and the ability of local operators to create their own programs, even in very high-level languages, is questionable. The point is that local processing does not necessarily imply local programming. You should not allow lack of data processing sophistication of local site personnel to keep you from considering the use of user-programmable terminals.

WHAT ABOUT PRICES?

As you might have expected, price is generally in proportion to capability. The single-quantity price for the minimum configuration of a single-workstation user-programmable terminal generally ranges from as low as $2,700 to over $17,000; when fully expanded to include the maximum main memory, mass storage, and peripheral options, the high end of the price can reach $30,000 to $40,000. Prices for multiple-workstation shared-processor systems go even higher. A typical system with 32K bytes of memory, a keyboard, a 1920-character display, a dual floppy disk drive, assembler language programming support, and one interface for a custom-supplied peripheral costs about $8,200; when an additional 32K bytes of memory, a second dual floppy drive, and three extra I/O ports are added, the price increases to about $13,700.

DISTRIBUTED DATA PROCESSING

One of the most prominent and frequently discussed concepts within the data processing community today can be summarized by three words: distributed data processing. If the trade press and market analysts are to be believed, DDP equipment and services constitute some of the fastest growth areas in the marketplace. Though the concept has yet to be defined in formal fashion, the basic idea is that because costs of processing components have fallen so drastically, economies of scale realized with use of large central processing systems no longer necessarily outweigh the advantages of processing data at the place where it originates and is used.

Historically, this concept was first developed in the early 1970's. The term "distributed data processing" was coined by key-to-disk manufacturers such as Entrex and CMC to describe remote multistation key-

User Programmable Terminals

to-disk systems that had been enhanced to include local file storage and updating capabilities. Although programs were generally written at the central site and down-line loaded for local processing, high-level data entry languages were also developed for use with these systems.

Slowly other segments of the marketplace, specifically minicomputer and terminal vendors, also added to their basic product lines features that supported the distributed processing concept. As industry pressure grew, even IBM to some degree began to embrace the DDP concept via enhancement of its 3790 Communication System. This highly complex and costly system, however, was one of the few IBM products which did not enjoy typical IBM success, and did not relieve industry pressure for support for distributed processing from IBM.

"Official" industry sanction (i.e., IBM support) for the DDP concept didn't come until October 1978, when IBM released its 8100 Information System. This remote multiple workstation shared-processor system provides full support for the concept of distributed data processing. Its advent implies that IBM can be counted on to add credibility to—and provide education for—distributed processing. Minicomputer and distributed terminal vendors should now be able to fine-tune their product lines and marketing approaches in relation to the 8100 System to provide users with a wide variety of choices. (Whether they do or not is another question.)

Despite strong promotion by the trade press and market analysts of distributed processing, there is some question as to whether significant market penetration has yet been made. For although there is very strong evidence that distributed data entry has become commonplace, Datapro has yet to find any great enthusiasm among all but a handful of the vendors we contacted for complete dedication to the concept of distributed data processing. (One possible exception, besides the key-to-disk manufacturers previously mentioned, may be in the area of remote batch terminals, where full off-line processing support has in recent years become standard.) As we noted above, some vendors whose equipment has the capability, were reluctant to be included in this report. In addition, each time Datapro has initiated user surveys on user-programmable and intelligent terminal equipment (the results of the most recent appears later in this report), the response from our subscribers does not indicate widespread usage.

Though we offer no explanation for this seeming discrepancy between the industry forecasts and the conservatism of most vendors and (Datapro subscriber) users, it is probably safe to say that the "jury is still out" on whether distributed data processing will become dominant, or even accepted, by the data processing community.

USER EXPERIENCE

To assess the current level of user satisfaction with user-programmable terminals, and to determine the patterns of usage of these terminals, Datapro conducted a subscriber survey. A Reader Survey form was included in the June 1979 supplement to DATAPRO REPORTS ON DATA COMMUNICATIONS and the July 1979 supplement to DATAPRO 70, and mailed to all subscribers. By August 15, usable responses had been received from 60 users with a total of 1375 terminals installed at 603 separate terminal sites.

Because some of the users reported on more than one model of user-programmable terminal, the user replies generated a total of 82 individual equipment ratings and profiles. Overall, the average number of terminals per user was 23, while the average number of terminals per response was 19. The average number of sites per user was 10, and the average number of terminal systems per site was 2.

In reply to the question about terminal system configuration, a total of 34 responses, or 41 percent, indicated that the terminals were configured with a single keyboard/display or keyboard/printer workstation; 47 responses or 57 percent, indicated multiple configurations, with an average of 4 workstations per system.

The users were asked to rate the overall performance, ease of operation, ease of programming, manufacturer's software, hardware reliability, and maintenance service for each terminal system by assigning a rating of excellent, good, fair, or poor. The resulting ratings for equipment of over a dozen manufacturers are summarized in the accompanying table. Any model or category that received more than two user responses is identified by manufacturer: models, categories, or manufacturers receiving only one or two responses were categorized as "other."

Because of the small sample size for some of the models, it would be unwise to draw firm conclusions from the indicated ratings. Rather, the ratings should be used as guides to potential strengths and weaknesses that may call for further investigation in selecting the most suitable equipment for your needs.

To put the raw counts into a form more readily grasped, Datapro calculated a weighted average for each rating category. Each user response was assigned a weight of one, and the ratings were weighted on a

User Programmable Terminals

Terminal Supplier and Model	Number of User Responses	Number of Terminals in Use	Weighted Averages and Response Counts																													
			Overall Performance					Ease of Operation					Ease of Programming					Manufacturer's Software					Hardware Reliability					Maintenance Service				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
Burroughs, all models	3	27	3.7	2	1	0	0	3.3	2	0	1	0	2.3	0	2	0	1	1.7	0	0	2	1	3.7	2	1	0	0	1.3	0	0	1	2
Datapoint 1100 Series	4	55	3.5	2	2	0	0	3.5	2	2	0	0	2.7	0	2	1	0	3.0	1	2	1	0	2.5	1	1	1	1	2.8	0	3	1	0
Datapoint 1500	3	6	3.0	0	3	0	0	3.3	1	2	0	0	2.5	0	1	1	0	3.0	1	1	1	0	3.3	1	2	0	0	3.0	0	3	0	0
Datapoint 4000 Series	4	5	3.3	2	1	1	0	3.0	0	4	0	0	3.3	1	3	0	0	3.0	0	4	0	0	4.0	4	0	0	0	4.0	4	0	0	0
Datapoint 5500 Series	4	47	3.8	3	1	0	0	3.8	3	1	0	0	3.3	1	3	0	0	3.3	2	1	1	0	3.8	3	1	0	0	3.5	2	2	0	0
Datapoint, other models	3	12	3.3	2	0	1	0	3.7	2	1	0	0	3.0	0	3	0	0	3.3	1	2	0	0	3.0	2	0	0	1	3.3	2	0	1	0
Datapoint subtotals	18	125	3.3	9	7	2	0	3.4	8	10	0	0	3.0	2	12	2	0	3.1	5	10	3	0	3.3	11	4	1	2	3.3	8	8	2	0
Digital Equipment, all models	3	12	3.7	2	1	0	0	3.0	1	1	1	0	3.3	1	2	0	0	3.0	0	3	0	0	3.3	1	2	0	0	3.0	1	1	1	0
Four Phase, all models	8	92	3.3	2	6	0	0	3.3	3	4	1	0	3.0	2	3	2	0	2.9	0	5	2	1	3.4	3	5	0	0	2.8	0	6	2	0
IBM, all models	8	36	3.0	3	2	3	0	3.3	4	2	2	0	2.6	0	5	3	0	3.0	4	0	2	1	3.4	5	1	2	0	3.1	2	5	1	0
Ontel OP-1	3	395	3.3	1	2	0	0	3.7	2	1	0	0	2.7	0	2	1	0	2.3	0	1	2	0	3.7	2	1	0	0	2.3	0	1	2	0
Raytheon PTS 100/1200	4	53	2.3	0	2	1	1	2.8	0	3	1	0	2.3	1	0	2	1	2.3	1	1	0	2	1.3	0	0	1	3	1.5	0	0	2	2
Sycor 350/351	4	42	3.3	2	1	1	0	3.3	1	3	0	0	3.0	1	2	1	0	3.3	1	3	0	0	3.0	2	1	0	1	3.0	2	1	0	1
Sycor 440	4	15	3.0	1	2	1	0	3.0	1	2	1	0	2.5	1	0	3	0	2.8	1	1	2	0	2.8	1	2	0	1	3.0	1	2	1	0
Sycor subtotals	8	57	3.1	3	3	2	0	3.1	2	5	1	0	2.8	2	2	4	0	3.0	2	4	2	0	2.9	3	3	0	2	3.0	3	3	1	1
Texas Instruments Silent 700 Series	7	186	3.6	4	3	0	0	3.4	4	2	1	0	3.1	3	2	2	0	3.0	2	3	2	0	3.6	4	3	0	0	3.4	3	4	0	0
Univac UTS 400	5	277	3.4	2	3	0	0	2.4	2	2	1	0	2.2	0	2	2	1	2.2	0	1	4	0	3.4	2	3	0	0	3.0	1	3	1	0
All others	15	115	3.4	7	7	1	0	3.3	5	10	0	0	3.2	5	9	0	1	3.1	4	9	1	1	3.2	5	7	2	0	2.8	2	7	3	1
GRAND TOTALS	82	1,375	3.3	35	37	9	1	3.3	33	40	9	0	2.9	16	41	18	4	2.8	18	37	20	6	3.2	38	30	6	7	2.9	20	38	16	6

LEGEND: Weighted Average (WA) is based on assigning a weight of 4 to each user rating of Excellent (E), 3 to Good (G), 2 to Fair (F), and 1 to Poor (P).

User's ratings of user-programmable terminals

conventional scale of 4, 3, 2, and 1 for excellent, good, fair, and poor, respectively. The data is presented as an additional information source, and not as the final word on the merits of the equipment represented. Individual vendors ratings are tabulated in the accompanying table.

Datapro is unable to draw any conclusions as to why user response was so low. It may at least partially stem from the traditional nomenclature associated with a particular system (minicomputer, key-to-disk system, teleprinter, etc.) and the "double-duty" functions that these products often serve. Few of the devices covered in this report are actually designated as "user-programmable terminals" by the manufacturer, and the communications capability may often be one of the less visible features in a multifunction system. Consequently, it is possible that many users who have the equipment installed, do not think of it as a "terminal" and simply did not respond.

In addition to basic configuration data and equipment ratings, the survey requested details on equipment usage patterns and other related information. The users were asked what types of functions their user-programmable terminals were performing, and what percent of operational time was allocated to each type of function. Most users indicated that their terminals were not dedicated to a single function, but rather were used for more than one purpose. Their usage patterns can be summarized as follows:

Type of Usage	Number of Responses	Percent of Operational Time
RJE/Batch	51	30.5%
Data Entry	58	58.4
Interactive with host	30	50.6
Distributed processing/ local file maintenance	41	28.5

Datapro also asked which terminals these users' equipment emulates, if any; some responses indicated more than one type of emulation was being used. A summary of the results, presented in decreasing frequency of occurrence, is as follows:

User Programmable Terminals

Terminal Emulation	Number of Responses	Percent of Total Responses
IBM 2780/3780	35	43%
Teletype 33/35	10	12
IBM 360/20 HASP	9	11
IBM 3270	8	10
Burroughs TD	3	4
Honeywell G115	3	4
Univac 1004	3	4
Other Univac	4	5
Other and unspecified	6	7
No answer	11	13

When asked what programming languages were being used with these terminals, the response was as follows, again presented in decreasing frequency of occurrence:

Programming Language	Number of Responses	Percent of Total Responses
Assembler	19	23
BASIC	17	21
COBOL	16	20
DATABUS (Datapoint)	13	16
TAL (Sycor)	5	6
RPG	3	4
FORTRAN	3	4
DATA IV (Four Phase)	3	4
TPL (Texas Instruments)	3	4
GP 300 (Burroughs)	2	2
3790 (IBM)	2	2
MACROL (Raytheon)	2	2
TICOL II (Texas Instruments)	2	2
Other	9	11
None	7	9

Use of more than one language was indicated on quite a few of the 82 responses, and about 38 percent indicated use of proprietary languages for at least part of the programming needs. Also, 9 percent of the responses stated that the users were not programming their terminals themselves at all; of these, most indicated that they were using ready-made vendor-proprietary software packages.

Datapro asked the users to identify the sources for their applications programs, with these results; again, multiple answers were quite common:

Source	Number of Responses	Percent of Total Responses
Custom-written by in-house personnel	63	77%
Custom-written by vendor's personnel	7	9
Custom-written by independent consultant	3	4
Proprietary/ready-made software package from vendor	31	38
Proprietary/ready-made software package from independent consultant	10	12

These users were also asked how their applications programs were loaded. Their responses were as follows:

Method Used	Number of Responses	Percent of Total Responses
From diskette	3	4
Downline loaded from host	21	26
From cassette tape	17	21
From disk	16	20
Other	15	18

The average amount of main memory per terminal was also requested, with these results:

Amount of Memory	Number of Responses	Percent of Total Responses
Less than 8K bytes	18	22%
8K to 32K bytes	44	54
Over 32K bytes	17	21
No answer	3	3

Datapro also asked these users to indicate what types of peripheral devices were included in their terminals' configurations. Their responses follow:

Type of Peripheral	Number of Responses	Percent of Total Responses
CRT display	59	72%
Serial printer	44	54

User Programmable Terminals

Diskette drive	39	48
Disk drive	31	38
Line printer	27	33
Tape cartridge or cassette drive	22	27
Half-inch magnetic tape drive	16	20
Card Reader	3	4
Other and unspecified	5	6

Finally, we asked these users how they acquired their user-programmable terminal equipment, with these results:

How Acquired	Number of Responses	Percent of Total Responses
Rent/lease from manufacturer	47	57%
Outright purchase	34	42
Lease from third party	1	1

In analyzing the results of this user survey, we became curious as to what types of companies utilize user-programmable terminal equipment. We were able to identify 30 responses or 37 percent as belonging to manufacturers. A few others could be identified as utilities, government and transportation. The majority were not identifiable by industry.

The wide variation in usage patterns and in the types of equipment reported by these users attests to the highly diverse characteristics of this class of equipment. There seem to be no standards for configuration or usage of user-programmable terminals, and there seems to be a great opportunity for users to "do their own thing." If and when the concept of distributed data processing becomes more acceptable to a large proportion of the data processing community, the "rules" may become more formalized. Until that time, apparently, almost anything goes.□

Remote Batch Terminals

Problem:

Up until recently, it was very common for a company, even a large company, to have only one centralized computer site. Full utilization (24 hours a day) of the computer was a major goal because of the computer's high cost. Such utilization was difficult to achieve because the data entry activity could not prepare (manually key) sufficient data during the prime shift to keep the CPU busy for three shifts.

This led to decentralizing data entry (preparation) to outlying plants and the introduction of batch terminals to transmit this data to the central site in "batches" at the end of the day. For companies with many outlying locations this practice frequently provided the additional data volume needed to load the central CPU.

In spite of the amount of press coverage given to the more exotic interactive transmission, there is still a large amount of data transmitted today in batch mode, but the terminal configurations are changing, as this report states.

Solution:

A quiet revolution has taken place within the last decade in batch data communications and remote job entry equipment. At one time most of the devices in this category were dedicated to the batch/RJE function and operated only under the direction of the central host computer. Now, however, many vendors offer the batch/RJE function simply as a feature of multi-functional systems, instead of providing a dedicated device. These multi-functional systems generally go well beyond the confines of batch/RJE terminal communications to include user-programming via a wide variety of programming languages, off-line data processing, file storage and maintenance, and sometimes interactive (in addition to batch) communications support. Because the characteristics of these more sophisticated terminal systems are much broader in scope than dedicated batch data communications and remote job entry, Datapro has recently introduced a new report, User-Programmable Terminals, that provides comprehensive coverage of these multi-functional devices. To avoid duplication, these devices were subsequently removed from this report.

Batch Terminals now provides coverage only on those more simplistic devices that are designed to transmit and receive large amounts of data, and (on some units) to perform data entry/editing and certain other limited off-line functions, but to do nothing else. If your requirements include some of the more sophisticated capabilities of the user-programmable batch terminals, such as local file updating and storage and local program execution, you should refer to the report on User-Programmable Terminals.

EVOLUTION OF THE RBT

The development of the earliest batch terminals originated from the concept of decentralization, which promoted use of the central computer to process jobs entered at remote locations; i.e., remote

Remote Batch Terminals

job entry (RJE). This represented a dramatic departure from the traditional use of computers in which all jobs were keypunched, entered, processed, and output at the central computer site. At the time, the punched card was the principal medium used for conveying data to the computer, and many of these remote locations had punched card accounting equipment installed. Thus, the provision for punched card input/output by the remote batch terminals lent itself directly to the existing card-oriented computer environment.

The concept of remote job entry gained momentum and the batch terminal market witnessed steady growth, with the major mainframe vendors introducing new products and an ever-increasing number of independent vendors introducing replacements for the mainframers' products. Nearly all of these early batch terminals shared a common characteristic: they were card-oriented and included a card reader, and printer in their basic configuration. Viewed in the light of today's trend toward application-oriented terminals, this "classic" architecture tended to severely limit the effectiveness of the remote batch concept. Another limitation of the early batch terminals was their inability to communicate with computers other than the one they were designed for. Users with more than one computer system, as well as those that employed remote computing services, were restricted to communicating with the specific computers or services with which their terminals were transmission-compatible.

This basic limitation was eliminated with the second phase of remote batch terminal development. Dramatic price reductions in the semiconductor market, including microprocessors, prompted minicomputer vendors to provide communications support for their products. With appropriate software support, their products could now emulate the communications protocol and operation of virtually any other terminal. And users could communicate with any mainframe for which suitable software was provided. Many users can now be equipped with two or more interchangeable terminal emulators, which gives them substantially improved operating flexibility over hard-wired terminals.

More recently, vendors of key/disk data entry systems have also introduced the capability to transmit prepared data to the remote computer, eliminating the need of physically transporting the prepared data to the computer site for processing. This alternative approach leads to increased operating flexibility and faster response. Some of these vendors have gone a step further to support their systems for stand-alone off-time data processing.

Placing intelligence at remote locations has opened the door for a significant new concept referred to as "distributed processing." This current "buzz word" simply means processing data where it is generated instead of batching it and transmitting it to a central computer for processing—another important step in decentralization. Distributed processing means that branch and regional locations can maintain their own files and merge them periodically with the corporate files located at the central site. This concept has necessitated changes in the architecture of the classic batch terminal configuration to arrangements that are more conducive to distributed processing environments. Disk or diskette storage is necessary for file storage and maintenance. CRT display units provide convenient access to the files and facilitate control of the remote system. Other devices may be required to satisfy expanded applications or operating contingencies. For additional information on intelligent batch terminal devices, please refer to Report C09-021-101.

THE BATCH CONCEPT

In producing and handling nearly anything, there are efficiencies in working with batches (i.e., groups with sufficient common characteristics to permit a single way of working with them).

As in any technical discipline, a jargon surrounds computer activities and often obfuscates the simpler meanings of words. For example, when grouping data in batches on magnetic tape, it's called blocking. When grouping data on moving-head disk units, it's called putting it in cylinders. When master files are organized sequentially and input transaction data is stored sequentially, the grouping is referred to as batch processing. Group the transaction data without sorting it before input, and you have real-time or random processing. Allow the program and data segments to be grouped in blocks of equal but arbitrary length, and you are ready for virtual processing. Transmit data over a communications line in any arrangement other than a character at a time, and somebody will refer to it as batch transmission.

Batch transmission is simply the grouping of data into specific blocks for transmission in individual spurts. A batch terminal, then, is a device that includes provisions for handling the transmission and/or reception of data in blocks. All data communications terminals actually possess this capability to some degree; otherwise, they could only transmit or receive one character per connection. However, convention dictates that the batch terminal designation shall be applied only to those devices that can accumulate a block of data separate from the input/output media and the transmission line; this is

Remote Batch Terminals

a rather wordy way of identifying a buffer. The buffer can be high-speed (e.g., semiconductor memory) or low-speed (e.g., magnetic tape cassette or floppy disk). Throughout this report, we will stick to this simple definition: a batch terminal is a buffered terminal.

This functional definition, of course, covers a much broader range of products than those that are categorized as dedicated batch/RJE terminals. Multifunctional systems, such as minicomputers, small-to-medium-scale general purpose computer systems (such as an IBM 360/20 or a Univac 9200), key-to-disk shared processor data entry systems, key-to-tape data entry units and systems, clustered and stand-alone intelligent CRT systems, and typewriter terminals with cassette or other I/O accessories frequently provide for batch data transmission. However, all of these categories of equipment are designed to fulfill additional functions, to which the batch/RJE function is subsidiary.

The dedicated batch/RJE terminal, to which this report is devoted, classically consists of some combination of a line printer, card reader, and/or card punch, plus communications interfacing. In some cases high-speed paper tape perforators and/or readers are included in the configuration; and tape, disk, or diskette drives may be provided for storage of data and of control and communications software.

Some systems may also support data entry and editing via CRT displays and keyboards, and even provide a limited data entry programming language through which the user can design data entry formats, field descriptors, and validation parameters. A few offer support for higher level programming languages, such as BASIC, RPG II, or COBOL, so that users may program additional emulators or otherwise modify functions related to the data entry and communications functions, but no general purpose applications programs (inventory control, payroll, etc.) can be programmed or executed at the terminal site.

In line with our earlier discussion, we will stipulate that these dedicated batch terminals are buffered. In general, buffer size ranges from 80 characters to about 512 characters and transmission speed range of 1200 to 19,200 bits per second. Still higher speeds, of up to 56,000 bps, or larger blocks, up to 4000 characters, are sometimes used.

As the block size increases, there is a trend towards higher efficiency of transmission because fewer line reversals are required (in half-duplex operation) and fewer control code sequences are required. But, since retransmission is still the only error correction technique that is commonly employed, block size can reach a point where the probability of encountering at least one error is very high for every block transmitted, thus reducing throughput to essentially zero. Optimum block length is highly dependent on the quality of the line. Under normal conditions, efficiencies due to increased block length do not fall off until the block size reaches several thousand characters.

RBT VS. RJE

We have talked about the equipment types available for performing the remote batch communications function, but have not talked about the function itself. Now is the time to introduce a descriptor, remote job entry or RJE, that is frequently used to describe the same equipment that we have previously referred to as remote batch terminals.

The two terms—remote job entry and remote batch communications—are functional descriptions, not equipment descriptions. The functional differences lie in the host computer's programming. If the terminal is acting as a peripheral device within an application program, then you have remote batch data communications. On the other hand, the RJE function employs the remote terminal as a "console" device to be used to initiate the execution of applications programs and usually to supply one of the input streams. The remote batch transmission function performs under the control of an applications program.

The RJE function operates under the control of the operating system software. While an entry keyboard and reply display or printer are not absolute necessities for the RJE function, they do make the operator's job easier, particularly for making job status requests.

Thus, remote batch communications treats the remote points as tributaries for input data and as destinations for outgoing data; remote sites are data-oriented. An RJE location, on the other hand, is job-oriented, controlling as it does, the programs that are executed on the host computer and the destinations of the results. The RJE location is, in effect, a remote console, input source, and output source for operating a computer. Typically, a job entered by an RJE terminal is just one of several jobs the host computer is processing at any one time.

THE COMPATIBILITY QUESTION

Up until the late 1960's, only a few companies were producing off-the-shelf batch/RJE terminal equipment, and only a few more were willing to put together a unit on a custom basis. Sensitive line

Remote Batch Terminals

USER RATINGS OF BATCH TERMINALS

Manufacturer & Model	No. of User Responses	No. of Terminals Represented	Weighted Averages and Response Counts*																													
			Overall Performance					Ease of Operation					Hardware Reliability					Maintenance Service					Software					Technical Support				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
DEDICATED BATCH/RJE TERMINALS (Not user-programmable)																																
Control Data, CDC UT 200	3	3	2.7	0	2	1	0	2.7	0	2	1	0	2.7	0	2	1	0	2.7	0	2	1	0	3.0	0	2	0	0	2.7	0	2	1	0
Harris—																																
1600 Series	10	21	3.3	5	3	2	0	3.6	6	4	0	0	3.2	4	4	2	0	2.7	3	3	2	2	3.2	4	5	0	1	2.5	0	7	1	2
Other and unspecified	3	10	3.3	1	2	0	0	3.7	2	1	0	0	3.0	0	3	0	0	3.0	0	3	0	0	3.3	1	2	0	0	2.7	0	2	1	0
Subtotals	13	31	3.3	6	5	2	0	3.6	8	5	0	0	3.2	4	7	2	0	2.8	3	6	2	2	3.2	5	7	0	1	2.5	0	9	2	2
IBM—																																
3741	5	98	3.4	2	3	0	0	3.4	2	3	0	0	4.0	5	0	0	0	3.2	1	4	0	0	2.7	1	1	2	0	2.8	1	2	2	0
3776	26	85	3.3	10	15	1	0	2.9	4	16	5	1	3.3	10	14	2	0	3.0	7	14	4	1	2.8	2	9	5	0	2.8	5	9	8	1
3777	22	62	3.2	6	14	2	0	2.9	4	12	6	0	3.3	8	12	2	0	3.0	6	11	4	1	2.8	1	14	5	0	2.9	3	14	4	0
3780	6	15	3.0	0	6	0	0	3.2	2	3	1	0	2.8	2	1	3	0	3.0	1	4	1	0	2.3	0	1	2	0	3.2	2	3	1	0
Subtotals	59	260	3.3	18	38	3	0	3.0	12	34	12	1	3.3	25	27	7	0	3.0	15	33	9	2	2.8	4	25	14	0	2.9	11	28	15	1
Northern Telecom Systems (formerly Data 100)—																																
70	5	19	3.4	2	3	0	0	3.4	2	3	0	0	2.6	0	3	2	0	2.6	0	3	2	0	2.6	0	2	1	0	2.4	0	2	3	0
74	10	20	3.2	2	8	0	0	3.2	3	6	1	0	2.7	2	4	3	1	2.9	4	2	3	1	2.4	1	3	5	1	2.4	1	3	5	1
76	24	111	3.0	3	19	1	1	3.0	3	18	3	0	2.5	1	13	8	2	2.5	1	12	9	2	2.9	2	12	4	0	2.7	1	14	6	1
77	4	12	3.2	1	3	0	0	3.2	1	3	0	0	3.2	1	3	0	0	2.7	0	3	1	0	3.2	1	3	0	0	3.0	0	4	0	0
78	12	138	3.4	5	7	0	0	3.4	5	7	0	0	3.1	3	7	2	0	3.1	4	5	3	0	2.9	3	4	4	0	2.8	3	4	3	1
Others and unspecified	5	27	3.0	0	5	0	0	3.0	1	3	1	0	2.8	0	4	1	0	2.6	0	3	2	0	3.2	1	3	0	0	2.7	1	1	2	0
Subtotals	60	327	3.2	13	45	1	1	3.2	15	40	5	0	2.8	7	34	16	3	2.7	9	28	20	3	2.8	8	27	14	1	2.7	6	28	19	3
Unitech, all models	5	12	3.4	2	3	0	0	3.4	2	3	0	0	3.0	1	3	1	0	2.4	0	2	3	0	3.0	1	3	1	0	2.6	0	3	2	0
All others	8	23	3.3	3	4	1	0	3.4	3	5	0	0	3.0	2	4	2	0	2.8	1	5	1	1	2.9	2	3	1	1	2.8	2	3	2	1
Totals	148	656	3.2	42	97	8	1	3.1	40	89	18	1	3.0	39	77	29	3	2.8	28	76	36	8	2.9	20	67	30	3	2.7	19	73	41	7

*LEGEND: WA—Weighted Average, E—Excellent, G—Good, F—Fair, P—Poor. The weighted averages are based on assigned weights of 4, 3, 2, and 1 for Excellent, Good, Fair, and Poor, respectively, for each response. Cross totals may not sum to the number of responses because not all responses included ratings for every category.

protocols were established for the major computer systems and remote computing services based on the specific batch terminals available then. These were principally the IBM 2780, the Control Data 200 User Terminal, the Univac 1004, the Univac DCT 2000, and, a little later, the IBM System/360 Model 20 as a multileaving HASP terminal. With the telecommunications software already established in host computers to accommodate one or more of these terminals, other manufacturers who wanted to get into the expanding data communications market for high-performance terminals (i.e., remote batch) found it easier to imitate the existing equipment and take advantage of existing software rather than face the software development and educational problems that would have accompanied the introduction of innovative communications techniques.

This pressure for plug-compatibility has not diminished. In fact, the products and procedures established in the 1960's by IBM, Burroughs, Control Data, Honeywell, and Univac have become defacto standards for batch communications, not only for dedicated batch/RJE devices, but for the more sophisticated multifunctional terminal systems as well. It is not uncommon on today's equipment for a vendor to provide a variety of interchangeable emulation packages from which the user can pick and choose those required for each application to be supported.

Some terminal vendors have implemented a particular vendor's protocol, such as IBM-developed BSC or SDLC, without committing themselves to complete emulation of one of that vendor's terminals. To the buyer, the choice between total emulation of a specific terminal and transmission compatibility with a particular protocol depends on the purpose of the new equipment. If you wish to replace existing terminals with no change in your host software or operating procedures, then direct emulation is required. If you wish to add new capabilities as you replace existing equipment or expand your configuration to include new terminals, transmission compatibility may well be adequate, since host software will require revision to accommodate new functions anyway.

PROTOCOLS

Probably the most confusing aspect of data communications is the link (line) protocols. All information, including both data and control codes, is transmitted within the same serial bit stream.

Remote Batch Terminals

USER RATINGS OF BATCH TERMINALS (Continued)

Manufacturer & Model	No. of User Responses	No. of Terminals Represented	Weighted Averages and Response Counts*																													
			Overall Performance					Ease of Operation					Hardware Reliability					Maintenance Service					Software					Technical Support				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
USER-PROGRAMMABLE BATCH TERMINAL SYSTEMS																																
Burroughs—																																
B761	3	7	3.0	0	3	0	0	3.3	1	2	0	0	2.7	0	2	1	0	2.3	0	1	2	0	2.7	0	2	1	0	2.3	0	1	2	0
Others	4	5	3.0	0	4	0	0	3.2	1	3	0	0	2.7	0	3	1	0	2.7	0	3	1	0	3.0	0	4	0	0	2.5	0	2	2	0
Subtotals	7	12	3.0	0	7	0	0	3.3	2	5	0	0	2.7	0	5	2	0	2.6	0	4	3	0	2.7	0	5	2	0	2.9	0	3	4	0
Datapoint—																																
1100 Series	4	14	3.5	2	2	0	0	3.7	3	1	0	0	2.7	1	2	0	1	3.2	1	3	0	0	3.2	2	1	1	0	2.5	1	1	1	1
1500	3	7	3.3	1	2	0	0	4.0	3	0	0	0	3.3	1	2	0	0	3.3	1	2	0	0	3.0	1	0	1	0	3.0	1	1	1	0
Others	3	10	3.3	1	2	0	0	3.7	2	1	0	0	3.0	0	3	0	0	2.7	0	2	1	0	3.0	0	3	0	0	2.0	0	1	1	1
Subtotals	10	31	3.4	4	6	0	0	3.8	8	2	0	0	3.0	2	7	0	1	3.1	2	7	1	0	3.1	3	4	2	0	2.5	2	3	3	2
Four Phase, all models	7	19	2.6	1	4	0	2	2.3	0	4	1	2	2.4	1	3	1	2	1.8	0	2	2	3	2.7	1	3	2	1	2.1	1	1	3	2
Harris, 1200 Series	8	12	2.9	1	6	0	1	2.6	2	2	3	1	2.5	0	5	2	1	2.9	2	3	3	0	2.7	3	2	1	2	2.6	2	3	1	2
IBM—																																
3774	3	14	3.3	1	2	0	0	3.3	1	2	0	0	3.3	1	2	0	0	3.0	0	3	0	0	3.0	0	1	0	0	3.3	1	2	0	0
3775	3	6	3.0	0	3	0	0	2.0	0	0	2	0	3.0	0	3	0	0	3.5	1	1	0	0	—	0	0	0	0	2.5	0	1	1	0
Other	8	20	3.4	3	5	0	0	3.1	1	7	0	0	3.3	2	6	0	0	3.3	2	6	0	0	3.0	0	8	0	0	3.1	1	7	0	0
Subtotals	14	40	3.3	4	10	0	0	3.0	2	9	2	0	3.2	3	11	0	0	3.2	3	10	0	0	3.0	0	9	0	0	3.1	2	10	1	0
Mohawk, all models	5	28	3.2	1	4	0	0	3.0	0	5	0	0	3.2	1	4	0	0	3.0	1	3	1	0	3.0	0	4	0	0	3.0	0	5	0	0
Sycor, all models	4	314	3.7	3	1	0	0	3.7	3	1	0	0	3.5	2	2	0	0	3.5	3	0	1	0	3.5	2	2	0	0	2.7	1	1	2	0
Univac, all models	3	3	3.0	0	3	0	0	3.3	2	0	1	0	3.0	1	1	1	0	3.3	1	2	0	0	2.3	0	1	2	0	2.7	0	2	1	0
All other	10	30	2.6	0	7	2	1	2.8	2	4	4	0	2.5	1	6	1	2	2.9	2	6	1	1	2.6	1	5	3	1	2.7	1	6	2	1
Totals	68	489	3.1	14	48	2	4	3.1	21	32	11	3	2.9	11	44	7	6	2.9	14	37	12	4	2.8	10	35	12	4	2.7	9	34	17	7
GRAND TOTALS	216	1145	3.2	56	145	10	5	3.1	61	121	29	4	3.0	50	121	36	9	2.9	42	113	48	12	2.9	30	102	42	7	2.7	28	107	58	14

*LEGEND: WA—Weighted Average, E—Excellent, G—Good, F—Fair, P—Poor. The weighted averages are based on assigned weights of 4, 3, 2, and 1 for Excellent, Good, Fair, and Poor, respectively, for each response.Cross totals may not sum to the number of responses because not all responses included ratings for every category.

Certain conventions or rules are required to be able to separate control information from data. Simple, low-speed terminals require simple control procedures, and such protocols have few rules. High-speed terminals, often with several separately addressable components and with extensive controllable functions (such as retransmission for error control) require many more rules to permit orderly transmission of data and control of the many functions.

Early protocols were developed around the use of unused data-character bit patterns and character sequences to identify control functions. This method used more bits than were absolutely required. Thus, bit-oriented protocols were introduced in which the lengths of control fields were not tied directly to eight-bit character lengths. Typical of the character-oriented protocols, and the one most used for batch terminals, is IBM's Binary Synchronous Control (Bisync or BSC). Typical of the bit-oriented protocols is IBM's Synchronous Data Link Control (SDLC), which is gaining favor slowly.

An international standard has been developed for bit-oriented protocol (ISO's High-Speed Data Link Control, or HDLC), which is a super set of SDLC. Most computer vendors have adopted a subset of HDLC as the protocol for their new terminals; frequently, older protocols are also supported to permit use of new terminals with existing networks.

In addition, higher-level protocols are being developed, such as CCITT's X.25 for access to a public packet switching network. Still higher-level protocols have been introduced for network control and management, including IBM's SNA/ACF and equivalent network protocols from other major computer vendors, such as Univac's Distributed Communications Architecture (DCA) and Digital Equipment's DECnet. A network protocol, although it incorporates one or more specific link protocols, goes far beyond the needs of transferring data between a terminal and a host computer; it incorporates rules to permit alternate connections, access to multiple host computers, access to multiple applications within one computer, etc. Even higher-level protocols, which one day will simplify application-oriented communications, are being developed.

POINT-TO-POINT OR MULTIPOINT?

There are two basic "network" arrangements: point-to-point and multipoint or multidrop. Point-to-point

Remote Batch Terminals

is just what it sounds like; i.e., a direct connection between the terminal and the host computer or another terminal. All transmission links established over the public telephone network (DDD) are point-to-point for obvious reasons. A multipoint arrangement shares one line among several terminals operating one at a time. There are many techniques for implementing a multipoint arrangement to ensure that each terminal sooner or later will have its chance on the line; the techniques range from computer-directed polling to unequal time-outs at each terminal bidding (contending) for the line. Because of the added control functions and contention among the terminals associated with sharing the line, multipoint operation can reduce the overall transmission efficiency.

The transmission mode—half or full duplex—has a large bearing on transmission efficiency. Half duplex is two-way transmission, alternately. It takes up to 250 milliseconds or so to perform each line turnaround. (In certain situations using adaptive modems for line equalization to attain high transmission rates, the turnaround time can be significantly higher, requiring a full duplex facility in order for transmission to be effective.) Full duplex is two-way transmission, simultaneously. Facility (line) costs are typically no higher for a full duplex leased voice-band line, but full duplex does add to the terminal costs and to the complexity of line control procedures.

Some interesting approaches have been taken to try to get around the limitations of half-duplex (performance) and full duplex (cost and complexity) operation. A popular approach is performed by software—IBM's HASP multileaving facility. This uses half-duplex transmission, but instead of just returning a positive or negative acknowledgement to a data block, another data block is returned along with the acknowledgement. Naturally, some provisions in the terminal have to be made to accommodate simultaneous data streams in and out and simultaneous operation of peripheral devices, even though the transmission is half-duplex. Many batch terminal devices now provide IBM HASP multileaving emulation.

DON'T OVERLOOK THE SOFTWARE

Discussion of remote batch transmission, or any data communications topic for that matter, would be incomplete without mentioning software. In today's world of data processing, there are two principal points at which software can impinge on your data communications activities: in the host computer (and possibly a programmable front end for it), and in the terminal itself.

Discussion of host computer software is really outside the scope of this report, but it centers around provisions of the operating system and perhaps an additional telecommunications control package for the physical control of each character of each line through the system, as well as message assembly and routing to the appropriate applications program or operating system module. Do not let this brief summary fool you; the functions mentioned can represent a major chunk of the host computer's memory capacity and processing time. Alternatively, a majority of this processing load can be wished off on a programmable front-end communications processor.

At the terminal end, software, if any, can be divided into emulation programs and other facilities. Emulation has been discussed above. The "other" category is only mildly flippant, as it can represent a grab-bag of capabilities ranging from shared-processor key/disk data entry to COBOL compilation. Careful analysis is required on the user's part to determine whether the data communications or the "other" functions are of primary concern, so that selection of equipment can be based on the proper priorities. The reader is again referred to User-Programmable Terminals for those batch terminals which provide high level sophistication in software support.

USER EXPERIENCE

An RJE/Batch Terminal Reader Survey Form was included in the August 1979 supplements to DATAPRO 70 and DATAPRO REPORTS ON DATA COMMUNICATIONS and mailed to all subscribers. By September 26, usable responses had been received from 140 users with a total of 1145 installed remote batch terminals.

Because some users reported on more than one model of terminal, the user replies generated a total of 216 responses or individual equipment ratings and profiles. The users were asked to rate the overall performance, ease of operation, hardware reliability, maintenance service, software, and technical support for each batch terminal by assigning a rating of excellent, good, fair, or poor.

The users' ratings for all the popular terminal models or families are summarized in the accompanying tables. Groupings by manufacturer only are presented where ratings for individual models were too few to present separately or where the model was not speciifed. Prospective buyers should note that the small sample sizes for some of the models and manufacturers make it unwise to draw firm conclusions from the indicated ratings.

Because the definition of user-programmability is a difficult one, users were asked to supply information

Remote Batch Terminals

on all devices being used for batch data communications or remote job entry, regardless of whether they were user-programmable. For your convenience, we have separated the responses into two categories: Dedicated Batch/RJE Terminals covered by this report, and User-Programmable Batch Terminal Systems.

As has been Datapro's practice, weighted averages are also presented to assist you in comparing units with differing numbers of responses. These averages were calculated by assigning the same weight to each user response and assigning weights of 4, 3, 2, and 1 to ratings of excellent, good, fair, and poor, respectively. This technique tends to prevent the comments of a large user from overshadowing those of a small user, and is essentially the same weighting that a prospective buyer would instinctively use when talking to existing users. Since there are many different ways in which the responses could be weighted to emphasize particular aspects, the averages presented in the tables should not be interpreted as final or absolute, but just as another piece of useful information to be correlated with other information.

In addition to the equipment ratings, other questions were asked to determine usage patterns, such as number of terminal sites, length of time installed, host computer, principal application (of the terminal), transmission speed and code, communications facility, transmission arrangement and line discipline, emulation of other terminals, transmission volume, method of acquisition, and plans for replacement.

This survey was meant to determine usage patterns (in addition to equipment ratings) rather than to develop rigid statistical results. Responses included some users who reported on more than one type of equipment, some who gave multiple answers to individual questions, and some who did not answer all questions. To provide a consistent basis, we have presented the usage patterns in terms of total number of mentions and as a percentage of the total number of user responses. In some cases, the percentages, therefore, total more than 100. All usage pattern figures include both dedicated batch/RJE terminals and multifunctional systems being used for batch data communications.

Among the 140 users responding, 1145 batch terminals were installed at 859 separate sites, for an average of 6.1 sites per user and an average of 1.3 terminals per site. The length of time that these terminals have been installed was reported as follows:

Time Installed	Number of Mentions	Percent of User Responses
Less than two years	68	31
Two to five years	125	58
Over five years	20	9

When asked whether they had active plans to replace these terminals, these respondents answered:

	Number of Mentions	Percent of User Responses
Yes, within 12 months	48	22
Yes, within 24 months	39	18
No active plans for replacement	109	50

The host computer or computers were identified on virtually all 216 responses. A total of 216 computers were mentioned, including some respondents who indicated use of multiple hosts and a few gave no response to the question. Multiple host computers per response can reflect large users reporting on several computer systems all using the same type of terminal, as well as users who connect to several different remote computing services. The overall distribution of host computer models had the following pattern:

Host Computer Model	Number of Mentions	Percent of User Responses
IBM 360/20, 30, & 40	4	1.9
IBM 360/50 & larger	7	3.2
IBM 370/115, 125, 135, & 145	44	20.4
IBM 370/155 & larger	73	33.8
IBM 303X	34	15.7
Other IBM	3	1.4
Amdahl 470 V/6 & V/7	11	5.1
Burroughs B 6700/7700	8	3.7
Burroughs B 2805/2815	2	0.9
Control Data 6000/7600/ Cyber 70 Series/Omega 480	12	5.6
Honeywell 66/68	1	0.5
Honeywell 600/6000 Series	1	0.5
Itel AS/5 & AS/6	4	1.9
Univac 90/60	3	1.4
Univac 1100 Series	9	4.2

The distribution of the principal applications for each terminal is summarized in the following table:

Principal Use	Number of Mentions	Percent of User Responses
Dedicated batch	183	84.7
Key/disk data entry	35	16.2
Stand-alone processing	11	5.1
Others	6	2.8

Many users indicated that they used more than one transmission speed, reflecting different speed usages

Remote Batch Terminals

by different terminals of the same type and/or the use of lower-speed DDD back-up procedures for leased lines. The distribution had the following pattern:

Transmission Speed	Number of Mentions	Percent of User Responses
2000 bps	7	3.2
2400 bps	47	21.8
4800 bps	117	54.2
9600 bps	88	40.7
19,200 bps	10	4.6

The overwhelming majority of respondents reported the use of EBCDIC code—173 mentions or 80.1 percent. Mentions from users of ASCII code totaled 30 (13.9 percent).

The users were asked to identify the type of communications facility used. As expected, the public telephone network (DDD) and leased voice-grade lines accounted for the great majority of the responses. Many users mentioned both DDD and leased facilities, indicating different usage with different terminals of the same model and also the use of DDD for backing up a leased facility. The combinations of multiple terminals, speeds, and facilities identified in individual responses prevented a correlation between these parameters. The pattern of distribution was:

Communications Facility	Number of Mentions	Percent of User Responses
DDD (telephone network)	72	33.3
Leased voice-band line	129	59.7

The users were asked to identify any terminals that were emulated by their equipment. The distribution was as follows:

Terminal Emulated	Number of Mentions	Percent of User Responses
IBM 3780	78	36.1
IBM HASP terminal	82	38.0
None	20	9.3

Other emulation noted by the users included IBM 2780 (4 responses); Univac NTR, 1004, or DCT 100 (4 responses); Honeywell GRTS (2 responses); Control Data 200 UT (2 responses); IBM 2770 (1 response); and Burroughs B761 (1 response).

The users were also asked to identify the transmission arrangement and line discipline used with each terminal. The pattern of distribution was:

Transmission Arrangement	Number of Mentions	Percent of User Responses
Point-to-point	191	88.4
Multipoint	16	7.4
Binary Synchronous protocol	164	75.9
SDLC protocol	24	11.1

The users were also asked for typical transmission volumes for each terminal. Although not all users provided these figures; the results for those who did are as follows (because a handful of users with unusually high volumes tended to skew an "average" figure, we felt it more accurate to show a median usage, i.e. half of all users responding transmit or receive less than the figure shown, and the other half more than that figure):

	Number of Mentions	Median Vol. Per Mention
Records sent per day	108	2,400
Print lines received per day	112	45,000
Records received per day	38	3,000

Finally, the users were asked by what means they had acquired their terminals, with these results:

	Number of Mentions	Percent of User Responses
Rent/lease from manufacturer	166	76.9
Outright purchase	30	13.9
Lease from third party	9	4.2□

Integrated Point of Sale (POS) Systems

Problem:

Point of Sale (POS) terminals are not necessarily communications-oriented devices because many companies tend to use them in the closed environment of a store without linking them into a network. If they are used locally, then the computer also has to be on site, and that is not usually practical for a large company with dozens or even hundreds of retail outlets. Two usage trends are developing. Rapidly falling computer costs, especially for special-purpose microprocessors, are gradually making it feasible to consider on-site computers; and less expensive communications facilities are making it equally attractive to link POS terminals to a central site that can handle all of a company's outlets. The direction a company takes will reflect largely how the company is organized—whether it operates with strong central control or delegates all but the highest management responsibilities to its outlets.

Regardless of the company's choice, the communications planner/designer/manager must have a working knowledge of POS terminals because they will either be communicating terminals, or they will be indirect considerations as inputs to the stores' on-site computers. The on-site computers will function as communications concentrators or front-end processors for the company's main computer.

This report is offered to familiarize you with POS techniques and equipment. It also explains some subtle but important differences among the concepts of POS terminals, POS systems, and integrated POS systems.

Solution:

The acquisition of an integrated POS system involves considerable expenditure. The system may not be fully cost justified unless a complete merchandise control system that takes advantage of the data captured is implemented. An integrated POS system is not just a sales transaction system; it's also a computer data entry system. Not every retailer needs or can afford such a system, but many are finding that an integrated system can provide significant cost savings in many facets of retail operation.

It is important to note that the terms POS system and integrated POS system are not synonymous and that several basic types of modern POS equipment are available to meet various requirements and budgets. This report emphasizes the concept of POS as it represents the broad, all-encompassing viewpoint of a complete retail-oriented information system.

The integrated POS system represents the automation of all point of sale functions in one hardware/software configuration. Such a system is a full-fledged data processing system in itself, designed to meet sales transaction and merchandise control requirements. As will be seen, the major problem lies in attempting to perform all such functions at a lower overall cost than in present non-computerized sales transaction systems.

INTEGRATED POS—PRO AND CON

Integrated POS systems provide the means for combined data entry and completion of sales transactions (i.e., cash registers or cash register replacements), the means for processing the data, and the means for presenting reports to management. To date, the emphasis on the data entry aspect has all too frequently overshadowed the other aspects of POS. The user who acquires the front end of a modern POS system had better decide exactly what he wants to do with the resulting data if he hopes to justify the cost of the equipment.

The retailer's nightmare about the possibility of having the complete system inoperative and being unable to handle any sales transactions almost blew the whole idea of integrated POS systems. The need for high system reliability has led to three designs: (1) self-contained terminals that can operate independently of the computer, (2) a dual processor arrangement, and (3) a combination of both that offers a third level of redundancy. Unfortunately, these increased reliability provisions have also increased the system complexity to a point at which the cost per terminal exceeds that of most cash registers. However, a direct comparison of equipment costs overlooks the significant potential savings made possible by the advantages of integrated POS systems. The advantages are summarized as follows:

- Ease of use. The various integrated POS systems offer the salesperson significant assistance in completing the transaction, up to the ultimate point where only a few keys need to be pressed. The keyboards of the POS terminals are designed for easily entering transaction functions and data. Furthermore, merchandise tag readers can just about eliminate the need for operator data entry. In each system, a display shows the entered data, and a detailed sales slip is produced.

- Accuracy. The critical problem of salesperson errors is virtually eliminated by the integrated POS systems, which perform extensions, sales tax computations, and split package computations; credit store and vendor coupons; and compute trading stamps, food stamps, and change. In other words, the need for mental arithmetic is eliminated.

- Speed. The speed advantage of any electronic device over a mechanical one is well known. When one adds the ability to perform sales calculations and the use of a tag reader for automatically entering data, the speed advantage becomes even greater. Tests have shown the electronic terminals to be about 50 percent faster without tag readers and 75 percent faster with tag readers.

- Merchandise data entry. The system also permits the entry of accurate merchandise data in addition to the customer transaction data. The timely availability of this data opens up an entire new cost savings area for the merchandiser that may supersede his floor-level savings.

- Credit authorization. The equipment used for sales data entry can also perform the credit authorization and credit sales functions simultaneously. Therefore, separate terminal and communications costs for the credit functions are eliminated.

- Operating cost savings. Significant savings in overall operating costs can be gained by reducing personnel requirements. (Personnel costs typically run about 8 percent of store sales.) Considering that most stores rely heavily on part-time personnel and have a high turnover, a natural by-product of an integrated POS system is a better organized selling staff.

COMPONENTS OF INTEGRATED POS SYSTEMS

The two basic components of an integrated POS system are the electronic register/terminals and the system controller. Other components frequently include merchandise tag readers, ticket printers, manager terminals, and (in supermarkets) electronic produce scales. The system controller may also perform the "backroom" data processing function, or the data may be transmitted to a larger computer either within the store or at a central location. The functions and general characteristics of each of these system components are described in the paragraphs that follow.

Electronic Register/Terminals

From the retailer's point of view, the electronic register/terminal is the most critical element in the system. It is here that the sale is completed and the customer/store interface exists. Whereas the conventional cash register was merely a tool for completing a cash transaction, the electronic register/terminal is designed to complete both cash and credit transactions and do so more accurately and quickly, in addition to performing computer data entry functions.

Integrated Point of Sale (POS) Systems

The key facilities of the POS register/terminals are:

- Keyboard. The basic keyboard design usually consists of a numeric panel and specific function keys. However, the number and arrangement of these keyboards can vary significantly among the manufacturers.

  The keyboards used in supermarket POS terminals tend to be even more function-oriented, with the operator control keys being specific department keys, and coupon, tender, and price look-up keys.

- Display. Operator displays in most register/terminals indicate the transaction being performed as well as the dollar amount and item number. Many systems provide a display for the customer as well, either on the opposite side of the panel or as a separate module.

- Printing. Many of the electronic registers have multiple print stations for producing customer receipts, validating the sales forms, and printing a sales journal. Printer speeds typically vary between 40 and 90 characters per second, and many of the units can print alphanumeric as well as numeric data, thereby producing descriptive register tapes. Facilities for handling sales forms of various sizes and types are often provided.

- Merchandise tag readers. With the advent of the electronic terminal and the integrated POS approach has also come a heavy stress on merchandise tag reading. The reasons quite obviously are: (1) the need for the manufacturer to demonstrate significant operational savings to justify the high system cost, (2) the awareness that salesperson errors must be minimized, and (3) the need for accurate merchandise control data to meet rising costs. The merchandise tag reading problem is so important and has so many facets that it is discussed separately later on in this report.

- Credit checking. In addition to entering merchandise data, each POS register/terminal can function as a credit card authorization and credit sale terminal. The credit card account number may be entered either via the keyboard or via an optical or magnetic reader—preferably, but not necessarily, the same reader that is used to capture merchandise data.

- Output data. The resultant sales transaction and merchandise data may either be recorded on a magnetic tape cartridge or reel and transmitted or carried to the computer system at the end of the day, or be transmitted directly to the controller where it is collected and sent to the backroom computer. Credit card authorization data must be sent and received in real time, regardless of the manner in which the sales and merchandise data is handled.

- Terminal Interactivity. The ability for terminals to operate interactively has become a key feature in the general retail segment of point of sale—so much so that most manufacturers claim terminal interactivity, although each may define it differently. While complete terminal interactivity denotes on-line real-time access to the headquarters computer, many manufacturers define interactivity as being on-line with the merchandise and customer data base, which can be implemented in individual store minicomputers. Semantics aside, each retailer must determine the precise on-line information he needs and the cost and reliability of the method for obtaining that information. Don't be fooled by the word "interactivity."

Controllers

The system controller, which is usually a minicomputer, may function on a store or multi-store basis, as we will see in the "Putting It All Together" section that follows. The controller's functions vary depending on whether it is being employed with intelligent or non-intelligent (i.e., programmable or non-programmable) terminals.

When used with non-intelligent (or "dumb") register/terminals, the controller must perform all cash register activities for each sales transaction. The critical problem of total system failure is usually alleviated by a dual minicomputer arrangement in which one processor is for back-up use. The positive side of this arrangement is a lower system cost, especially for larger stores, and the ability for the terminal functions to be easily customized through software in the controller.

The system controller is the focal point for sales, merchandise, and credit data. At a minimum, data from the register/terminals is checked for transmission errors and reformatted for transmission to the backroom computer. However, the controller may serve as much more than a data collection unit; it may also perform sales analysis, price look-up, and credit and check authorization functions at the store level. Price look-up is especially important for the supermarket industry in connection with the Universal Product Code; it can also yield savings by checking for incorrect merchandise pricing.

The controller also collects terminal sales data and can supply the manager with an up-to-the minute picture, which can be significant in the present competitive environment. Also, the manager can access his organization's central data base through on-line data communications links.

Integrated Point of Sale (POS) Systems

Technique	Advantages	Disadvantages
Print-Punch Tags	Already in widespread use; minimizes equipment expense	Data capacity of tags is limited; tags are easily lost from merchandise
Magnetically Encoded Tags	High data density; stick-on labels are not easily removed from merchandise; tag readers can also read credit cards; encoders can perform read-after-write checking	Encoder is comparatively expensive; tags cannot be changed for mark-downs
Optical Bar Codes	Can read labels inside packages; encoder and system costs are generally lower than for magnetic tags	Cannot read most credit cards; tags cannot be changed for markdowns
Universal Product Code	Accepted by the food industry; packages are source-marked by the vendors	Labels do not include prices; high-cost system is required for price look-up; reliability is not yet assured; supplementary in-store encoding may be required
Optical Character Recognition	Tags are economical to produce and directly readable by customers and salesclerks; tags can be changed for markdowns if handprint recognition is available	Readers are comparatively expensive; reliability of handprint recognition is questionable

Table 1. Merchandise tag reading techniques

Customer credit and check authorization may be performed at either the store or backroom level. In the former case, the controller may perform either negative or positive credit authorization.

Today, all of the manufacturers of fully integrated POS systems employ minicomputer or microprocessor controllers at the store level.

Manager Terminals

The manager or administrative terminal is usually a CRT display terminal, teletypewriter, or printer that allows the manager to access the store-level controller or backroom computer for sales, merchandise, and credit information. The printing terminals (or an associated printer with a CRT) may permit the manager to request hard-copy reports, an excellent system feature.

Merchandise Tag Readers

Devices that can automatically capture the information recorded on merchandise tags or labels are available for use with most of the integrated POS systems currently on the market. Their use is certain to grow rapidly because of the reductions in labor cost and error rates they promise. In addition, these devices make it economically practical to capture complete, accurate merchandise data.

Five basic techniques for reading merchandise tags are currently being promoted. The Table 1 summarizes the principal advantages and disadvantages of each of the five techniques.

Print-Punch Tags. Most of the merchandise tags currently in use are of the print-punch variety. Therefore, some manufacturers have opted to use readers for such tags in their POS systems and thereby avoid additional expense. Unitote/Regitel, for example, offers a handheld unit that does not require the tags to be removed from the items. The print-punch tag, however, still has the severe limitations of the small amount of data it can hold, the possibility of its being lost from the merchandise, and its inability to be placed on certain types of merchandise (e.g., cans).

Magnetically Encoded Tags. The placing of magnetically encoded data on the merchandise tags has been supported by Sweda, Singer, and IBM. Such tags need not be removed from the merchandise and offer a reliable reading approach in which the tag data can't be changed. This latter advantage, however, can also be a disadvantage for handling markdowns. The key attraction of this approach is the capability it offers for reading many different credit cards, since the magnetic stripe is becoming a universal standard for credit cards.

The expensive part of the system is the tag encoder that magnetically encodes the tags, performs a check on the recorded data, prints data in human-readable form, and counts the tags. The data is usually

Integrated Point of Sale (POS) Systems

entered via a keyboard, although at least one system, from Litton, can accept punched card data. IBM's system can encode up to 60 characters per tag and read up to 500 tags per minute.

The reading is accomplished by a magnetic wand reader, and reading is omnidirectional. Readers of this type are relatively inexpensive, with the IBM reader priced at $350.

Optical Bar Code. An alternative to magnetic encoding is the use of an optical bar code technique in which the data is printed in the form of specific combinations of bars and spaces to denote each character. The value of the optical bar code over the magnetic technique is its lower printer cost ($2,000 to $10,000). The tag readers typically cost about $700 each, and the overall system cost, in most cases, is lower than with magnetically encoded tags. However, credit cards must be specially prepared when optical bar codes are used, or else separate readers for the magnetic-stripe credit cards must be provided.

In the retail industry, each manufacturer has developed his own special bar code, which only adds to the confusion. NCR has a three-color coding system (black, green, and white), in which the codes are based on the order in which the colors are sensed. While this approach leads to a highly reliable system, the tag reading cost is relatively high. A much lower-cost and simpler approach, developed by Pitney Bowes, uses black and white bars and represents each digit by four black bars and three spaces. Bars and spaces can be narrow or wide.

Universal Product Code (UPC). The grocery industry has selected a standard optical bar code to be printed on food containers or labels and scanned at the checkout counter. This bar code is a 10-digit code, expandable to 30 digits, and consists of a black and white code in the form of bars and spaces to represent numbers. Human-readable OCR-B numerals are printed under the bar code. The object is to have the manufacturers or distributors of all products, rather than the stores, place the labels on the merchandise, and have one standard, universally acceptable code.

With supermarkets having traditionally small profit margins, the use of tag reading is a virtual necessity for the justification of an integrated POS system, considering the high system cost. For example, McKinsey & Company projects a before-tax saving through UPC tag reading of 1 to 1.5 percent of a store's sales; e.g., a $40,000-per-week store could save $27,000 before taxes each year. However, a substantial investment of about $120,000 for such a store is needed, which means a five-year breakeven point.

The Grocery Industry Universal Product Code presents product but not price information. Prices are stored within the POS system and retrieved on the basis of the product information read from the labels. Consumer reaction against the lack of item pricing has been strong enough so that certain state legislatures have passed laws requiring item pricing. Furthermore, the elimination of item pricing virtually necessitates a duplex minicomputer system to assure continuous system operation. With UPC scanning coupled with price marking, both increased customer satisfaction and a less costly system can result. The cost of marking prices in this situation should be less than present marking costs, and overall system savings should still be significant.

Amid some early skepticism, and in spite of its relatively recent implementation, the UPC already has become a well-established standard. The source marking of goods by packagers has reached 80 percent or more of all items, and this has helped to spur a recent increase in scanning installations. In fact, the long-heralded use of scanners on a large scale seems finally to be at hand. Proven hard cost savings in existing scanner-equipped stores aided by the high incidence of source-marked goods, is attracting all major chains to this new method of checkstand operation.

Optical Character Recognition (OCR). The National Retail Merchants' Association has selected an OCR A font, called NRMA Voluntary Retail Identification, as the standard for merchandise reading. (It should be noted that the OCR A code is incompatible with the OCR B font selected by the supermarket industry, although there have been indications from NRMA that this problem may be close to a solution.)

The OCR technique offers the most straightforward approach to tag reading and permits both machine and human readability. The technique permits the use of a low-cost printer as compared to the other tag techniques. However, at present the OCR tag reader cost is higher than that of the optical bar-code or magnetic readers.

Through NRMA adoption of the standard, the industry is attempting to induce vendors to source-mark their merchandise so as to eliminate the need for store marking. However, the retail industry is likely to run into some of the same problems experienced by the supermarkets. In fact, the problems are likely to be compounded because of the greater numbers of vendors and products involved.

Some of the larger retailers, such as Sears and J.C. Penney, have already ordered OCR tag readers and are persuading their vendors to do source marking. Still, wide industry acceptance is uncertain and will depend on three major factors: 1) vendor marking, 2) lower reader prices, and 3) compatibility with the supermarket code.

The print-punch, magnetic encoding, and optical bar code techniques for reading merchandise tags are in widespread use; however, the success of the NRMA standard could replace these methods with OCR-A marking.

Ticket Printers

The ticket printers used to prepare the merchandise tags or labels vary widely in complexity and cost (from about $200 to $10,000), depending on the type of tags employed. Merchandise data can be keyed or automatically entered via punched cards or magnetic tape. At present, the ticket printers are usually placed in each store. In the future, it is hoped that universal merchandise coding will be printed on virtually all items by the manufacturers.

Electronic Produce Scales

An important optional component of integrated POS systems for the food industry is an electronic scale attached to the register/terminal to accurately weigh produce and record the price. While clearly reducing labor in the produce section, these units can have a serious effect on productivity at the checkout stations, slowing throughput while the checkers put the produce on the scales. This factor suggests that the optimum position for the electronic scale is in the produce section, where it will simultaneously weigh the produce, attach the code, and record the weight.

PUTTING IT ALL TOGETHER

On the previous pages, we have described the components of the integrated point of sale systems; in this section we will attempt to fit the pieces together. It should be restated that an integrated POS system is far more than a cash register and can in fact serve as a total retail information system. Consequently, our terminology relates to POS equipment that not only performs the sales transaction function but also enables sales and merchandise analysis.

The most basic type of POS system consists of free-standing individual devices, such as a specialized intelligent terminal having its own processing capability for sales transactions and an optional computer-readable output medium (e.g., magnetic tape) for storing sales and merchandise data prior to transmittal to the backroom computer. The terminal contains sufficient logic for arithmetic operations, extensions, discounts, tax calculations, operator guidance, and, in some cases, price look-ups. However, credit authorization cannot be performed at free-standing terminals. As the number of terminals increases, this approach proves uneconomical and cumbersome.

A system configuration that retains the basic advantages of the free-standing concept but results in a lower cost per terminal is the POS central collection approach (Figure 1). In this configuration, each terminal operates independently but transmits the data to a central collection unit in each store. The collection unit records the data on tape for later transmission to the backroom computer via a polling technique. Again the data is captured and transmitted in batches. While having the advantage of a lower cost per terminal than the separate free-standing units, this configuration cannot perform credit authorization nor supply the store manager with detailed, on-line information about his store operations.

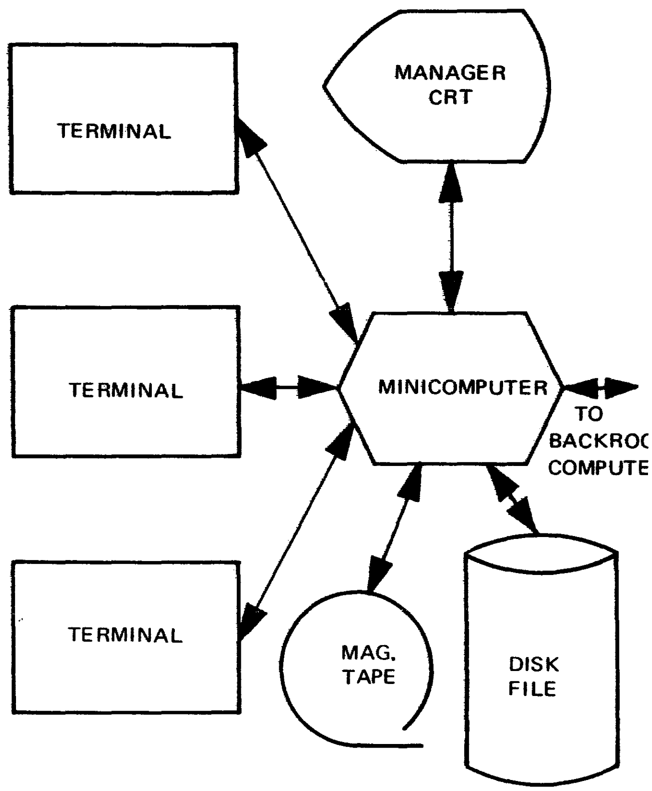


Figure 1. In-store intelligent terminal/shared-processor configuration

The central collection unit's early economic advantages over a minicomputer-based controller have declined with the constant price reductions in minicomputers. Replacing the central collection unit with one or more minicomputers (Figure 2) opens up new horizons and permits efficient execution of all the functions of an integrated POS system. In fact, these minicomputer-oriented system designs are superseding the central collection units and can, in most

Integrated Point of Sale (POS) Systems

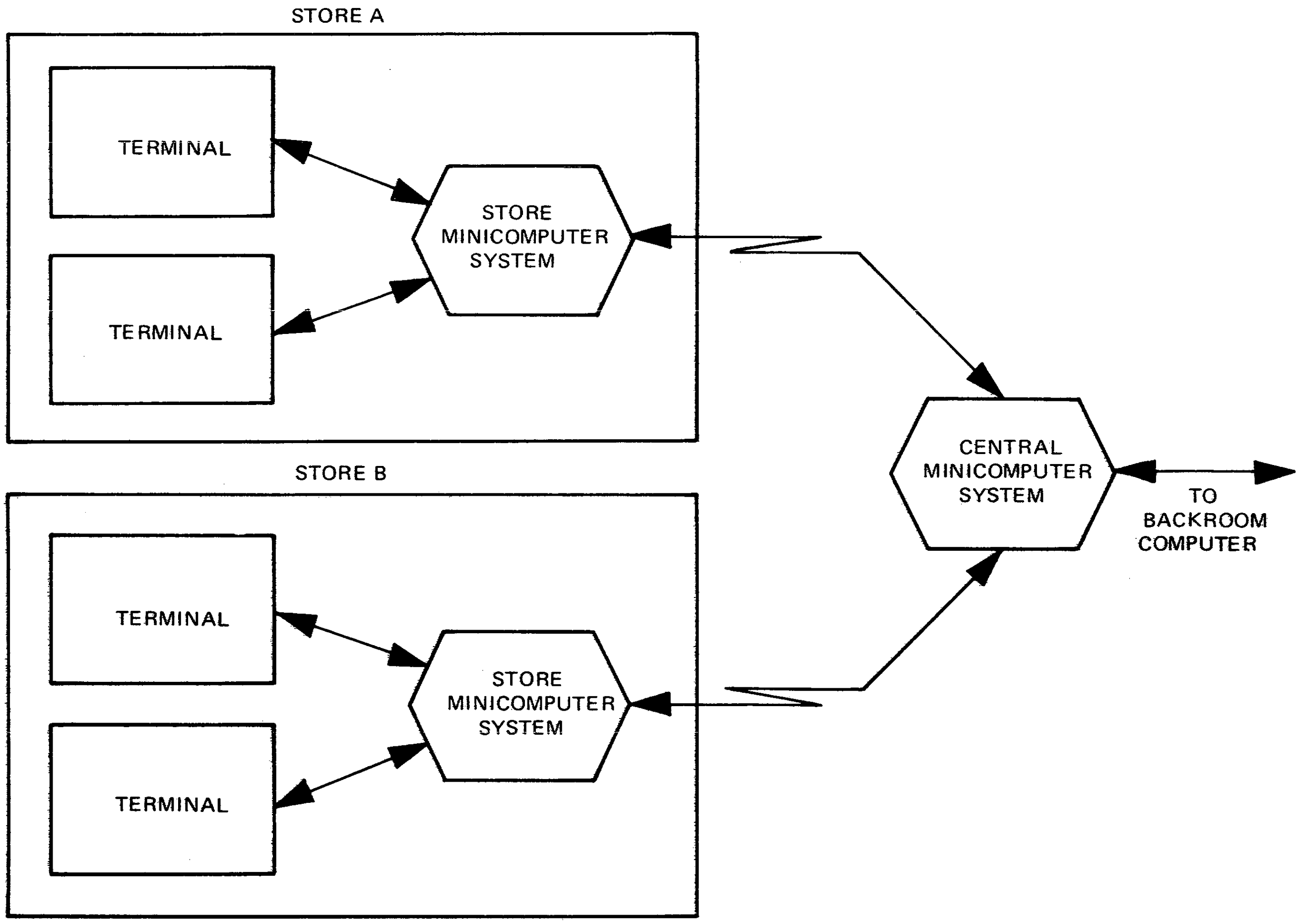


Figure 2. Multi-store intelligent terminal/shared-processor configuration

cases, be easily upgraded from central collection configurations.

In these shared-processor configurations, while the individual terminals still contain the necessary sales transaction logic, the minicomputer at the store or multi-store level provides store sales and merchandise analysis for store management. Therefore, while the corporate management receives company-wide reports via the backroom computer, each store manager is "on-line" to his own store's operations. The minicomputer can also perform credit authorization, which may be negative or positive, at the store level. With present communications problems and costs, this facility may prove to be both desirable and economical.

The in-store system configuration (Figure 1) represents virtually all of the current supermarket integrated POS systems and the recently introduced discount store systems. The concept of a complete in-store system has received increased interest as the prices of minicomputers continue to fall while their capabilities increase.

The system architecture of these intelligent terminal/shared-processor configurations is quite modular in that the terminal operations can be performed independently of the minicomputer. In most systems, modifications or even breakdowns of the minicomputer will not greatly affect the terminals. The offsetting disadvantage is that such systems do not represent the least costly approach to performing the tasks of an integrated POS system.

A lower-cost arrangement (Figure 3), can be called the dumb terminal/shared-processor configuration. This arrangement becomes progressively more economical as the number of terminals increases, since all the register logic is in the processor and is shared by all the terminals. Other advantages are that the terminal functions can be easily changed by software in the processor, and that the terminals themselves are relatively simple and easy to service. To protect against the serious problem of total system failure, a dual-processor arrangement is used that yields a system essentially as reliable as the intelligent terminal approach. However, it is hard to convince many retailers of this fact, and so there is a negative

Integrated Point of Sale (POS) Systems

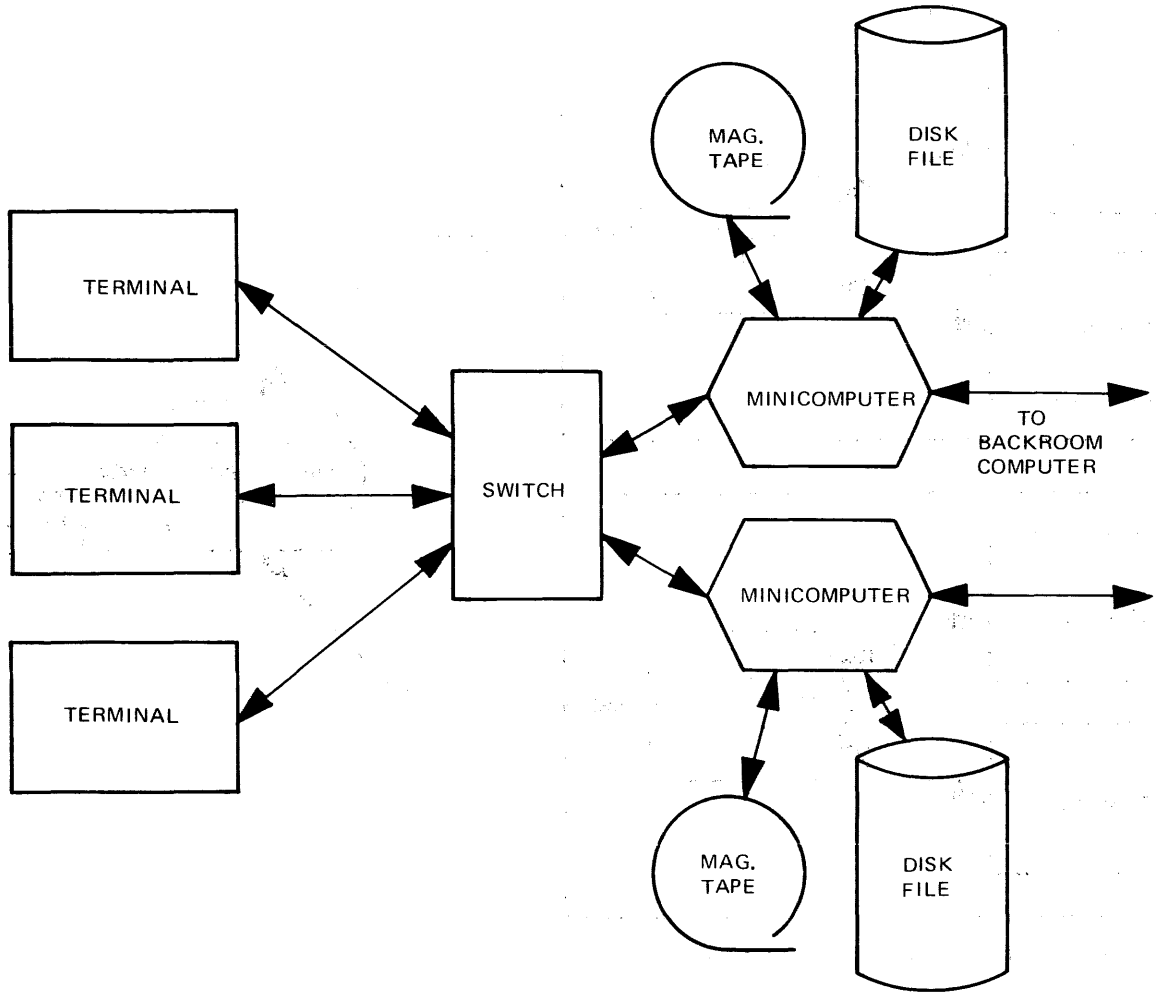


Figure 3. Dumb terminal/shared-processor configuration with duplex processors

psychological factor. Also, having to put all the logic in the minicomputer militates against the modularity concept and can make system modifications and reconfigurations more costly.

The retailer who bases his system selection on the lowest initial equipment cost may be in for greater costs than he bargained for. Centralized data processing has often suffered from putting everything into one unit.

While each of the basic system configurations offers certain attractive features, the modular, highly flexible system approach taken by NCR and Singer, and to a lesser extent by IBM, has substantial merit in facilitating system upgrading and in permitting independent terminal operation while providing a total POS system. However, for the experienced large store that desires to enter POS with a substantial system (e.g., over 30 terminals), the dumb terminal/shared-processor arrangement undeniably offers a lower-cost system.

Finally, prospective buyers should keep in mind that personnel savings are more substantial than equipment costs in the long run, and that the desirable end result is an effective total retail information system.□

Voice Response Data Terminals

Problem:

A "terminal" in its broadest possble sense can be literally anything that supports a dialogue between you and a computer or between you and another terminal. The dialogue can also be literally anything you want it to be from a heavily one-sided stream of information with the equivalent of an occasionally grunted response to a highly complex interchange of partially synthesized information (inter-database conversations through intelligent terminals). We have examined several of the dialogic alternatives in earlier reports in this segment, all of which have used some form of visual contact between the users and the machines. The other possibility for man-machine dialogues is some kind of aural medium, which is just as large a part of the daily human information input/output experience as the visual media. Aural is too broad a term because it qualifies any heard sound, so we will use the terms "audio" or "voice" in the narrower sense of modulated sound that carries the intelligence of written words.

At this point in the technology of voice-based man-machine communications, the computer has the advantage because it can talk to us in understandable form, but we can't talk to it directly except in some very limited and special cases. The computer-to-man path is called voice response; this report demonstrates the maturity of voice-response technology. The man-to-machine path is called voice recognition. It is a fledgling technology, at best, but will undoubtedly become very significant by about the mid-1980's. The one-sided nature of voice-response systems is not all that bad nor inconvenient, especially when you consider that any handy telephone can be easily made into a voice-response terminal.

Solution:

Of the three major aspects of information handling—collection, processing, and dissemination—only the area of processing has seen really effective solutions. The problems of data collection and dissemination have not fared nearly so well. Lack of results does not, in this case, denote a lack of effort. Just look at the nomenclature that has risen around the efforts to solve these problems: on-line, real-time, random processing, source data collection, key entry, telecommunications, data bases, timesharing, OCR, etc.

Invariably, the first thought is how to get the people with the information directly in touch with the computer and the computer in touch with the people who need the information. (The U.S. mail, or anybody else's mail for that matter, is hardly the answer.) The central theme of approaches to this problem is data communications—electrical connections between the central computer facility and the points where the information is created and used. At this time, we have a theoretical solution but not a practical one. These connections between the computer and outlying points (which may be on the shop floor less than a hundred yards away or across the country) involve what has been euphemistically referred to as man-machine interfaces. To most of us, they are terminals. (Often, cost considerations mean that terminals represent the termination of an investigation.)

It was no surprise, then, when the common telephone, with its intrinsic connection to just about everywhere, was proposed as a terminal. No incremental cost was involved for these terminals because every office had to have one anyway. But early attempts to use the telephone as a data entry device met with difficulty. The rotary dial system of circuit interruption may be fine for making telephone exchanges work, but it is unsatisfactory for data entry, both from the standpoint of user convenience and signal interpretation by data processing equipment. With the advent of the Touch-Tone system, a whole new era was born for use of the telephone as a terminal. The only catch was that the only possible output, without extra-cost equipment, was voice.

WHY NOT VOICE RESPONSE?

People react more immediately to the human voice than to any other means of communication. True, the printed word may have a more lasting and long-range effect, but for immediate reaction, you just can't beat the human voice.

Agreed that the sounds available from existing voice response units are hardly silver-tongued oratory, but the visual appeal of a CRT doesn't compare with great art either. The point is that people are keyed to responding quickly to the spoken word.

The primary difference between human speech and voice replies generated by existing equipment is information content. The subtle shadings of intonation used in everyday speech cannot conveniently, be reproduced by machine at this time. One of the reasons is that it is virtually impossible at this time to fully codify intonation.

The speed of information transfer using voice deserves some mention. A typical speaking rate is 125 words per minute, or about 12.5 characters per second. Beyond about 250 words per minute, it not only becomes very difficult to form the words fast enough, but most people have difficulty in understanding what is said. By contrast, typical reading speeds range from about 400 words a minute up to several thousand for people with special training. Clearly, voice is not a rapid means for transferring information. Compounding this problem is the fact that one cannot develop an effective scanning capability for vocal communications. If you record a conversation, playing it back at the original speed is the only satisfactory way to refresh your memory. Visual scanning and abstracting, on the other hand, is a highly developed technique.

One of the best ways to think about voice response in terms of information handling is to regard it the same way you would regard someone reading tabular information to you. This quickly gives you a feel for the relatively small amount of information that can be transferred before loss of attention interferes.

The whole case for voice response, then, rests on two points:

- It permits use of low-cost, readily available terminals—telephones.
- People are traditionally keyed to react to spoken communication.

JUST WHAT IS VOICE RESPONSE?

The heart of a voice response system is a voice or speech generation module. In many respects, a voice response system is like any other data communications system. But instead of transmitting the digitally coded characters comprising the reply message to the remote terminals for interpretation and printing, recording, or displaying, a series of characters is transferred to the voice generation module to specify what words, and in what order, will be "said" to the person listening at the other end of the line. Thus, the messages that can be generated depend entirely on the words that have been previously recorded and placed in the voice generation module. This group of words is referred to as the "vocabulary." Later in this report, some of the more esoteric details of voice generation will be covered.

The subtleties of working with a fixed vocabulary can best be illustrated by example. The normal way of saying 1,429 is "one thousand four hundred twenty nine." This requires six different words. A couple of words can be saved by saying "one four two nine." Indeed, any number can be said by using only the ten basic numeric digits. To say numbers up in millions in the normal way (first example), a vocabulary of some 24 words is required. But probably even more important, it is much more difficult to code a program to generate numbers spoken in the conventional fashion than to simply output a series of digits. In general, the programming difficulties involved in making use of a large vocabulary can outweigh the

Voice Response Data Terminals

cost of the extra vocabulary as the deciding factor in determining the vocabulary to be used in an application. New developments impacting this "classical" approach are discussed later. (However, at least one vendor offers programming support for conversational -style number reporting.)

WHY LET THE COMPUTER TALK?

The three traditional areas of application for voice response systems are:

- Bank account balance information or, more generally, central file inquiry.
- Retail store credit checking.
- Status reporting, usually checking work in progress or inventory.

Of these areas, banking applications currently account for the vast majority of installations.

The areas above don't include your interest? Then let's cover some general guidelines about application areas. There are two situations for which voice response seems naturally suited:

- To answer "how about" questions.
- To implement low-cost data entry from the originating source.

The "how about" question typically takes the form of "how many widgits are in stock, so I will know whether I can fill an order immediately?" Or, "what is the status of project xyz?" These are questions you typically ask someone now and get a verbal reply, frequently over the telephone. You could almost as easily make an inquiry to a computer system, using the same telephone, and get back the same kind of reply you've been getting all along. Provided that is, that the computer system has been properly programmed and that the data base has been established and maintained. And these are not necessarily minor details. Careful attention needs to be paid in such a system to how and when data gets into the system so that correct replies can be generated.

Low-cost data entry carries two aspects. One is the obvious use of the telephone as the entry point, with its wide availability. The second is the use of voice response in data entry applications. An example best illustrates how voice response is used in data entry applications. A major retail firm in Canada permits direct catalog ordering from home Touch-Tone telephones by depressing appropriate buttons. Voice response is used to repeat the order for customer verification.

WHERE'S THE CATCH?

So far, voice response has sounded like a pretty good thing. You're probably thinking that there are some drawbacks somewhere. And you're right, there are. Before we get into specific limitations, let's take a look at some of the difficulties the voice response concept has experienced in the past.

Voice response became a practical reality in the early to mid-1960's and enjoyed a mild spurt of popularity. Some of the convenience of use was dimmed by the limited availability of Touch-Tone telephones. This meant that some sort of attachment was required to enter data. Though not particularly expensive, such an attachment seemed a nuisance. But much more damaging was the lack of vigorous marketing by the makers of the equipment. Other ways of handling problems suitable for voice response were supported more strongly, and a user could get more help in implementing these approaches. IBM was about the only vendor offering any kind of systems support, so it is no wonder it dominates the installations now, in spite of high equipment costs.

To computer buyers in general, voice response appeared more as a custom-designed system than a standard product. In addition, other techniques for data entry and dissemination were surfacing about this time—key-to-tape recorders, OCR, CRT's—and voice response nearly got lost in the shuffle, mainly because applications of the concept were difficult to visualize compared to other emerging techniques. Potential users were often disappointed when confronted with the limitations of working with a small vocabulary. Making the computer talk didn't seem nearly so exciting when it was discovered that it didn't even have the vocabulary of an idiot. There was, and is, much resistance to working without a hard copy of the information received. In addition, performance of a voice response system is a difficult thing to measure objectively. You can't point to a stack of printed output and say "that's what the computer did." The resurgence of interest now is due largely to the availability of lower-cost equipment with more flexibility, larger, more conversational vocabularies, software support, and better education of potential users by the vendors.

Generalities aside, what are the drawbacks of voice response?

For applications that involve data entry from originating sources, there are no major drawbacks—as long as you are entering numeric information. Sure, there are terminals for entering alphabetic information easily, but for more costs. And there are techniques for entering alphabetic information from a Touch-Tone telephone, but for any sizable amount of

Voice Response Data Terminals

alphabetic information, say a name and address, they're slow and error-prone. In addition, there is no hard-copy confirmation of what data was entered. Agreed, read-back can be used for operator confirmation of the data entered, but what about audit trails? Actually, this situation with voice response is no different from any other input technique except possibly OCR: the source documents are at the remote location. No better care will be taken of the source documents if the data is entered via a CRT than via the keys of a telephone. It is the same old story. Nothing happens automatically. You have to plan for it.

For "how about" inquiries, many of the same comments apply. The information received is limited by your memory or your ability to take notes. A programmed repeat function is often helpful to assist you in cases where your attention is diverted at the critical moment. Again, if you think of voice replies as someone reading you a tabulated report, you can get a pretty good idea of the amount of information that can be conveyed effectively in reply to an inquiry.

SOME RECENT VOICE RESPONSE DEVELOPMENTS

There have been several developments during the past few years that will have major impact on the future of voice response.

Perhaps the most important of these developments is the growing use of minicomputer-based voice response systems for controlling non-voice terminals in addition to the voice generation and response functions. Pioneered by companies like Periphonics and Wavetek, this capability promises some economies that will be welcomed by data processing users. It is one of those situations that sounds so logical and right as soon as someone thinks of it. The processing capability of the minicomputer can be just as effectively used to manage communications with conventional terminals as it can to manage voice terminals. This concept may do for voice response what the mixed-media approach (OCR combined with keyboard data entry stations) promises to do for OCR. Voice response is no longer an esoteric discipline standing off by itself. It now becomes just another tool to be considered and used where applicable, in much the same manner as you would choose between teleprinters and display terminals.

Discussion of any phase of data processing is incomplete without a look at what IBM is doing. The venerable IBM 7770 Audio Response Unit forms the basis for the great majority of present-day voice response systems. But the independents have made inroads. Of late, IBM has been proposing and installing special voice response systems built around its System/7 minicomputers. So far, all of these systems are based on special, non-standard hardware and software components and can accommodate up to 15 lines. Voice generation is done by frequency synthesis, digital patterns are stored on disk. A vocabulary of up to 1000 words goes with the system. Monthly rental is in the range of about $2,300. What future standard IBM products, if any, will emerge from this activity is anybody's guess.

Mention of frequency synthesis brings up another "new" topic. The quotes are required because the concept and working hardware are relatively old. One IBM unit, the 7772, used this technique years ago; the 7772 did not sell well and was dropped.

The Votrax Division of Federal Screw Works markets the Votrax system. This unit synthesizes phonemes, which are basic particles of speech. Theoretically, any word can be constructed from a fixed set of phonemes. In practice, a particular set will give rise to a particular accent. (Linguistic students, please forgive this simplistic approach to a very complex subject.) Votrax has adopted a set of just over 60 phonemes. This permits identification with six binary bits of information. Add two bits to permit identification of four stress (pitch) levels, and you arrive at the magic 8-bit byte. Roughly speaking, there are about as many phonemes in a word as there are letters.

One of the approaches that Votrax is taking is that it is wasteful to fully occupy a telephone with the low rate of information transfer represented by voice output. Leasing the voice reply generator at remote points permits several low-speed lines transmitting 8-bit commands digitally to be multiplexed on a single voice-grade line. Sounds like a time-sharing application, doesn't it?

We have listened to the Votrax system. The startling thing about it is the range of vocabulary available. Typically, the vocabulary is stored at the host computer as a series of byte commands for each word. Some users, along with Votrax, are working on algorithms to convert directly from the character form to the spoken form, which could reduce the amount of storage required for the vocabulary. The quality of the generated voice replies is difficult to judge. At its best, it is amazingly realistic, but without the full depth of the typical human voice.

Two applications notes will round out this discussion of voice response current events.

ASI Teleprocessing, Inc., of Watertown, Massachusetts, got into voice response through developing a system to permit text editing by blind people. The Nucleus 4000 is built around a DEC PDP-11 minicomputer. Integral to this system is software that will

Voice Response Data Terminals

take a word, as delimited by spaces, and search for a match in the stored vocabulary. In use in this application, the software permits a blind person to type a section of text, which is then literally read back to him. If the system cannot find a match for a word, it is spelled out letter by letter. Thus, spelling errors can be detected by the author. Text editing routines permit corrections. While not the only application for this system, the text editing function illustrates the unique nature of voice response.

Computone Systems, Inc., of Atlanta, Georgia, provides a time-sharing service for insurance salesmen and financial planning representatives. Using one of two special audio terminals, a representative can conveniently access the host computer in Atlanta. One of the terminals uses a 21-key pad and a digital display. The other uses multiple rotary switches to set up all data fields before transmission. A wide range of actuarial and financial planning routines can be accessed through the terminals. Computone manufactures the terminals, develops the supporting software, and operates the computer. The service is called KEYPACT, and Computone states that it has over 3000 current users.

A BRIEF SAMPLING OF VENDORS

For those areas of the country whose telephone companies do not provide Touch-Tone service or a keypad, and for those users who want a portable keypad or expanded terminal, a number of companies market key-input voice-output equipment. The following paragraphs discuss the relevant product lines of eight companies to give you a representative idea about what is available. (The addresses and phone numbers of these companies are listed in the Suppliers section of this service.)

ASI TELEPROCESSING, INCORPORATED offers a terminal with full typewriter keyboard. The ARTS-70 terminal was originally designed for use by blind people using time-sharing services for text editing in conjunction with a voice response system. (See the comparison charts for the company's voice response system.) The terminal utilizes the standard 3-of-14 coding technique, which provides 98 characters; alternatively, another model uses a 3-of-15 technique in conjunction with data sets provided by ASI Teleprocessing to provide a 125-character set. The terminals include an internal speaker and sell for $650 arranged for connection to a telephone line through a DAA; an acoustic coupler is available for $100 additional.

CHANCELLOR INDUSTRIES, INC. produces the portable PRT-102 terminal, which includes a 16-key pad and internal speaker. It sells for $600 or rents for $25 per month.

DYMO BUSINESS SYSTEMS, INC. has sold its product line and business inventory of Audacall Automatic Card Dialers and Automatic Dialing Telephones to North Supply Company. Early models manufactured by the Audac Division of Dymo include the Audacall II and III, which were complete telephone instruments including the automatic dialers for rotary and Touch-Tone exchanges, respectively. The Audacall IV and V which are add-ons to existing rotary and Touch-Tone telephone instruments, replaced the earlier models in the company's product line. The newer models permit convenient connection to multi-line telephones. The Audacoder automatic dialing card is coded with a black felt-tip pen rather than holes. In addition to the basic function of holding up to three independent telephone numbers, a card can be used to hold up to 66 characters of constant information.

ELCOM INDUSTRIES, INC. markets the AT-100 terminal, which is specifically oriented toward credit authorization and electric funds transfer applications. The terminal includes lever switches for entering 5 or 6 digits (or more on special order), a reader for magnetic-striped credit cards, and provision for automatic dialing of a telephone number (remote computer) by the depression of a single key. It can be acoustically attached to a telephone to gain access to the telephone network. The terminal has been certified by the FCC for direct connection to the public telephone network without a separate DAA. A 12-key pad is included, and the terminal can function as an ordinary manual Touch-Tone telephone. A 12-digit terminal identification number is stored internally and transmitted automatically. Up to 40 characters of information, in the standard ABA format, can be read from the credit card. The card read capability can also be used to make the terminal function as a card dialer. The basic terminal costs $550 and can be leased for as little as $14.10 a month (including maintenance) on a 5-year lease. Quantity discounts are available on the purchase of more than 9 units.

INTERFACE TECHNOLOGY, INC. produces a line of 12- and 16-key pads with and without digital displays. The 720 is portable, battery-powered, 12-key pad that sells for $79.50; the 721 is a 12-key pad that includes an acoustic coupler and sells for $82.50. Interface Technology states that over 2000 of the 720/721 terminals are in use. The 722 is a 16-key pad that sells for $125. The 727 is a 16-key pad designed for numeric and limited alphanumeric data entry; it sells for $99.50. All 720 Series units are equipped with an acoustic coupler as standard.

Interface Technology's 731 is a desktop model that provides an 8, 12, or 16-digit display in addition to a 16-key pad. Four backlighted displays are optional features that light when one of the four function keys

Voice Response Data Terminals

is depressed. The 731 is DAA-coupled to a telephone handset. It costs $450. The 736 is an interactive data entry terminal with a 19-key keyboard with an operator's guide and LED display. It is equipped with an EIA RS-232 or current loop interface as standard. It sells for $600.

TRANSCOM offers the Audioport 160, a 65-key audio terminal that includes an external speaker and can be acoustically coupled to a telephone or connected to a DAA, depending on the model. Models are available for operation on AC power, rechargeable batteries, or a combination. The keyboard is arranged in a block layout rather than in the staggered arrangement of a typewriter. The standard keyboard arrangement groups letters and numbers in sequential fashion, but other layouts are possible. A representative price for the Audioport 160 is $545, and for the portable Audioport 116, $160.

Transcom has recently introduced two new Touch-Tone/audio response terminals, the 12-key Audioport 012 and the 16-key Audioport 016. Both are battery-powered Touch-Tone keypads that attach to the mouthpiece of a telephone via an acoustic adapter. They are designed to provide Touch-Tone transmission from Touch-Tone and rotary-dial telephones and receive voice responses through the telephone earpiece. The 12-key unit sells for $75 and the 16-key unit for $99. Quantity discounts are available.

WAVETEK offers a pair of audio response terminals. The 510 is a 12-key pad, and the T500 is a 48-key unit with standard typewriter layout. Either can be connected to the telephone network through an acoustic coupler or directly through a DAA. Both contain an external speaker and operate from AC power.

HOW TO ASSESS VOICE RESPONSE SYSTEMS AND COMPONENTS

The first order of business, is to resolve the terminology of voice response versus audio response. Audio response is perhaps the most widely used term, because that's what IBM uses. However, audio encompasses the tones generated by the Touch-Tone telephone as well as voice. Those few who really care about the difference insist on including tone responses within the category of voice response because they can be used to drive printers or other recording devices. As a matter of fact, if you exclude the very high-speed broad-band data communications, all data communications falls within the audible range of frequencies; and, if you wanted to stretch a point, conventional data communications could be included under the heading of audio response. In this report, voice response is used only to highlight the specific subject area, not because of nit-picking about precision of definitions.

Voice response equipment offerings fall into four categories: complete peripheral subsystem, module for a computer communications subsystem, free-standing system, and voice generation module.

A complete peripheral subsystem includes facilities for interfacing and controlling communications lines and the voice response module. A module for a particular computer communications subsystem implies that the regular communications facilities of the computer can be adapted for line interfacing and control, while the subject equipment supplies the voice generation capability. A free-standing system includes facilities for interfacing and controlling the communications lines, the voice generation module, and additional processing and peripheral device capabilities suitable for implementing applications directly without requiring a host computer. A voice generation module is just that; it includes no capabilities for line interfacing or processing.

In some cases, the manufacturer offers several configurations, each intended for a different mode of use. All supported modes are listed. A remote subsystem is a special case of the peripheral subsystem; instead of being located physically adjacent to the host computer, it transfers data to and from it via a communications line.

Communications Input

Number of lines refers to the number of individual trunks or ports into the equipment and normally identifies the number of simultaneous inquiries that can be handled.

The **form** of communications input refers to the nature of the data signals the equipment works with. In general, three types of signals are important in voice response situations: tones, serial and parallel. Tones are generated by the Touch-Tone telephones and the 10-, 12-, and 16-key terminals produced by independent manufacturers. Serial signals are the most common form from conventional data communications terminals; usually there is a Bell System 100 or 200 series or equivalent modem connected to the terminal. The parallel form is rarely used by conventional terminals; usually there is a Bell System 400 series or equivalent modem connected to the terminal (although this does not guarantee that the parallel mode is being used, because some 400 series data sets can be adapted for serial operation). Tones are actually a form of parallel transmission, but they form an important separate category because the generating telephones do not require a data set and parallel, nontelephone terminals do.

The importance of the form of input is coupled with the decoding capability of the equipment; together

Voice Response Data Terminals

they determine what terminals and arrangements (data sets) are required at the central processing site. Generally, a Bell System 403D3 data set is required at the processing site to convert the tones into signals the equipment can handle.

Decoding refers to the capability to interpret the code used to represent numbers and letters. Three codes are important in voice response. These are the 2-of-8 (AB), 3-of-12 (ABB), and 3-of-14 (ABC). The letters in parentheses are alternate names for the codes. The 2-of-8 code is used by Touch-Tone telephones. It yields 16 different code combinations. Some independent manufacturers of pads provide 16 keys; at present the Touch-Tone telephones have 12, but a 16-key arrangement is likely to become available at some future time. The 3-of-12 code is used by the Transcom Audioport 160; special adapters are usually required to receive this code properly. The 3-of-14 code is used by IBM 1001 terminals and other parallel transmission devices.

Most manufacturers identify specific terminals from which the equipment can receive data. Naturally, equivalent devices are also accommodated. For example, systems that can handle the 12-key Touch-Tone telephones can also handle 12- and 16-key pads from independent manufacturers. Some systems can also accommodate conventional serial terminals such as teleprinters and CRT's. These capabilities are independent of voice response and tone decoding capabilities.

Processing Capabilities

Manufacturers use various ways to identify the capabilities included in their equipment and those that must be supplied by the user. Cut to the barest minimum, the following three areas of processing capabilities must be provided: line control, message control, and applications. Line control involves maintaining order among the several simultaneous data paths represented by several lines being active at once. Message control involves assembling incoming data characters into whole messages and disbursing the programmed reply words in the proper order. Applications processing is the capability for interpreting incoming data and determining what the reply message should be.

If the system is a free-standing one, then all of these processing capabilities must be included. Typically, a peripheral subsystem would include line and message control facilities, with the host computer handling the applications processing. The facilities in the voice response system can be implemented through specialized, hardwired logic or, in accordance with the current trend, via a minicomputer. If a minicomputer is included, it can sometimes be programmed to do some of the application processing.

Software indicates the scope of support the manufacturer provides. The more he provides, the less the user will have to do. Be careful to clarify the exact pricing arrangements for software; i.e., is it priced separately and, if so, how much?

Speech Generation

Vocabulary type identifies the basic building blocks from which reply messages can be constructed. The most frequently used vocabulary units are words and phrases. A phrase consists of two or three or more words, depending on the manufacturer. A phrase differs from several independent words in that if a phrase is addressed, it all comes out. A closely allied feature provided by some manufacturers is that of programmed synthesis, discussed later, which can effectively expand the vocabulary beyond the recorded words.

Vocabulary size is the number of individual vocabulary units that can be recorded.

Typical speech rate is a more meaningful way of phrasing the usual milliseconds-per-word or milliseconds-per-track specifications quoted by the manufacturers. Normal speech rate is about 125 words per minute, corresponding to 480 milliseconds per word. Depending on the generation technique (discussed later), the speech rhythm may or may not correspond well to natural spoken words.

Vocabulary storage plays an important role in how the output will sound. There are two aspects: nature of the signal in stored form, and type of storage device. The signal can be in analog or digital form. In addition, one company, Periphonics, has developed a proprietary technique which is somewhere-in-between.

An analog signal is what you hear yourself; a tape recording is a recording of analog signals. A rotating device is generally required to store an analog signal. The most common approaches are magnetic sleeves mounted on drums and film sleeves mounted on drums. In the first case, the signal intensity (loudness) is related to the degree of magnetization. In the second case, the sleeves are divided into tracks. Each track corresponds to the length of a word or phrase. The difficulty with this approach is that it forces words into constant lengths, whereas natural words have different lengths. This causes interference in the rhythm of speech. Particularly long words or phrases can sometimes be spread over two tracks—with a consequent reduction in vocabulary size.

Though these techniques may require a lot of diddling to get natural-sounding speech if long phrases are required, they are well suited to the

Voice Response Data Terminals

spell-it-out approach usually taken with small vocabularies. In all fairness, one of the best-sounding voice response systems heard by the Datapro staff was such a unit, and it included a rather long phrase. (The staff could not determine exactly how it was accomplished, but the understanding is that some rather careful recording of the vocabulary, splitting up the phrase over several tracks, was involved.)

Digital signals involve determining and assigning a binary (digital) value to the amplitude of the analog waveform at fixed intervals and storing the digital values. Knowing the sampling intervals, the digital values can be retrieved and the original waveform (spoken word) can be reconstituted. Any conventional device for storing digital information, such as magnetic drums or disks or core storage or solid-state components, can be used as the vocabulary storage device.

The above discussion is very basic. There are many problems to overcome in either analog or digital storage. Either can be effective and either can be a flop, depending entirely on the care taken in design. For example, analog recording on either a magnetic or film sleeve includes the problem of variations in output amplitude (loudness) not present in the original recording. For digital recording, the sampling rate and amplitude resolution effect the fullness of the output voice. In either method, the circuitry used affects the shape of the output waveforms and consequently the tonal characteristics of the voice. Fortunately, there's an easy way to check for tonal quality; listen to it. However, the unit you get may vary from the demonstration unit. Be sure to make some arrangements for hearing your equipment before acceptance.

Generation technique refers to the manner of taking the recorded signals and converting them to spoken words again. For analog recording, this is simply a matter of reading the signal and amplifying it. For digital recording, some kind of tone synthesis is required. This merely means putting tones together to get the final sound (spoken word).

Programmed synthesis is handy if you want to expand your vocabulary without taking up more storage or buying more units. In essence, it is a method of taking parts of words and putting them together to form words not in the vocabulary. An example would be to take the "fif" sound from fifteen and combine it with the "ty" sound from forty, to get fifty. Another example is the Votrax approach discussed earlier.

Programmed intonation is the capability to vary the sound of a recorded word to give it a rising inflection, falling inflection, or changed emphasis (volume). Human speech is more flexible than this, but these three form the basic capability. With rising and falling inflection, for example, a single vocabulary word could be used to end a statement, end a question, or appear in the middle of a sentence. The advantage of programmed intonation is to add to the understandability of the voice reply; it could be done with an expanded vocabulary if you wanted to go to the trouble. Only limited capabilities of this type are provided in any of the units now on the market.

Vocabulary multiplexing is the simultaneous output of the same word on multiple lines. Without this, unprogrammed pauses would be inserted in a voice replay while another line was using the next word to be output on this line. Single-line units naturally have no need for this capability.

Message capacity is a throughput measure. Usually, in a multi-line system, some line control logic is shared among the various lines. Therefore, calls on some lines must wait until the line control logic can get to them. This decreases the number of calls that can be handled on each line. Some very esoteric mathematics comes into play to properly discuss this subject. Queuing theory is definitely beyond the scope of this report. Basically, throughput varies also with the quality of service, as defined by the incidence of busy signals inquirers get.

More to the point, however, is a discussion of the average length of a call. The number talked about most frequently is 30 seconds. In a single-inquiry situation with little data entry, read-back verification, security checking, etc., such a call length is quite reasonable. And that is the situation most systems are in now. But as the applications for voice response expand, longer calls will be the order of the day. For this reason, throughput for 1-minute and 3-minute calls was included wherever the information was available.

USER BACKTALK ABOUT COMPUTER BACKTALK

A Reader Survey Form on voice response was included in both the April 1978 supplement to DATAPRO 70 and the May 1978 supplement to DATAPRO REPORTS ON DATA COMMUNICATIONS. By the editorial cutoff date of August 1978, a total of 28 usable replies had been received. Anticipating a comparatively small number of replies, Datapro designed the questionnaire to gain as much information as was reasonable to request. Consequently, the forms took a little longer than usual to fill out, and we heartily thank those of our subscribers who did respond. Though few in number, the re-

Voice Response Data Terminals

sponses represent considerable communications activity. These 28 responses represented over 77,000 calls per day from over 14,000 terminals.

Among the 28 responses, only 4 manufacturers received a sufficient number of mentions to warrant individual treatment. These four were IBM, Periphonics, Votrax, and Wavetek. They accounted for 26 of the 28 responses. ASI Teleprocessing received one response, and the other response did not specify the system in use.

Table 1 summarizes the characteristics of the systems reported on in the Reader Survey Forms. Of particular note are the average number of calls per day and the total number of terminals served. The large numbers reported here, in comparison with conventional data communications, would seem a characteristic of voice response as it is used today. It is not possible to achieve this volume of data communications using any other types of terminals at an economic price.

The subscribers were also asked to identify their applications for voice response. The 28 subscribers who responded to this question replied as follows:

	Number of responses	Percent of total responses
Data Entry:	13	46
With voice reply verification of data entered	11	39
Without voice reply verification of data entered	2	7
Inquiry Response:	25	89
Account balance	17	61
Status	11	39
Credit verification	3	11
Other Applications	5	18

The other applications listed included assignments for substitute teachers, memo posting, and teaching aids for elementary students.

Several users indicated that their data entry was for control purposes such as placing and releasing holds and cautions and activation deactivation of lines and terminals.

Of the 28 respondents, 17 (61 percent) were banks or bank service centers.

Table 2 presents the users' ratings of the equipment. Overall, the table shows an unusually high degree of satisfaction. This is borne out by the individual ratings, as well as by the unusually high marks earned by all the equipment vendors in the "bottom-line" category of overall satisfaction. Further substantiation is given by the general question about whether the system lived up to the manufacturer's promises; 54 percent of those responding indicated that their systems fulfilled the promises immediately, and the others said they fulfilled the promises eventually.

Of the 28 users with a total of 54 voice response systems who responded to our survey, 3 systems were free-standing; 22 were attached to IBM System 370's; and 29 were attached to other manufacturers' mainframes, including Honeywell and DEC. None of the systems reported on were attached to IBM System/360 computers.

Quantitative information about the terminals used was a little harder to pin down. The figures regarding the number used in Table 1 are conservative. Several users indicated a number as "approx. 300," for example; we used the figure directly. Several users identified the terminals in use as something other than Touch-Tone telephones or keypads. These users had Transcom 160's. Bell System Transaction Telephones, Interface Technology 720's, and Votrax equipment.

In general, the users rated the Touch-Tone telephone very high, with an occasional complaint about promptness and/or quality of maintenance. We could not determine the origin or many of the keypads in use; most seemed to be from the telephone company. The keypads rated a little lower in every category than the telephones. Overall, of the 28 users, 13 used Touch-Tone telephones only; 6 used both Touch-Tone telephones and keypads; and 5 used some other terminal.□

Voice Response Data Terminals

	IBM*		Periphonics*	Votrax*	Wavetek*	Others*	Total or Average, All Responses*
	7770	System/7					
Number of responses	5	4	5	3	9	2	28
Number of systems	5	8	5	25	9	2	54
Average number of telephone lines per system**	16	11	16	6	17	18	114
Vocabulary size:							
10-35 words	0	0	0	0	0	0	0
36-75 words	1	3	0	0	5	1	10
76-150 words	4	1	4	0	2	0	11
Over 150 words	0	0	1	2	2	1	6
Average number of calls per day:							
0-1000	0	2	1	1	4	1	9
1001-2500	1	1	2	0	3	0	7
2501-5000	2	1	0	0	1	1	5
5001-7500	0	0	1	1	1	0	3
7501-10,000	1	0	0	0	0	0	1
Over 10,000	0	0	1	0	0	0	1
Total number of terminals served:							
Number of responses	4	3	5	1	5	1	19
Number of systems reported on	4	7	5	9	5	1	31
Number of terminals	1238	3245	7308	9	2255	200	14,255
Average length of call:							
0-1 minute	5	1	4	0	6	1	17
1-2 minutes	0	0	1	0	2	1	4
2-5 minutes	0	0	0	1	1	0	2
Over 5 minutes	0	3	0	2	0	0	4
Average response time:							
1-5 seconds	4	4	4	3	8	2	24
5-10 seconds	0	0	1	0	0	0	1
10-15 seconds	0	0	0	0	1	0	1
Over 15 seconds	0	0	0	0	0	0	0

*All information, except number of systems, terminals, and average number of lines per system, consists of response counts.
**Average for users responding with both number of lines and number of systems identified.

Table 1. Overall system characteristics

Voice Response Data Terminals

	IBM 7770	IBM S/7	Periphonics	Votrax	Wavetek	Others	All
Overall Satisfaction—							
Wtd. Avg.	3.6	3.0	2.8	3.0	3.4	3.5	3.3
Excellent	3	1	0	1	4	1	10
Good	2	2	4	1	5	1	15
Fair	0	1	1	1	0	0	3
Poor	0	0	0	0	0	0	0
Ease of installation—							
Wtd. Avg.	3.4	2.3	2.4	3.7	3.4	4.0	3.1
Excellent	2	0	0	2	5	2	11
Good	3	2	2	1	3	0	11
Fair	0	1	3	0	1	0	5
Poor	0	1	0	0	0	0	1
Throughput/response time—							
Wtd. Avg.	3.2	3.5	3.4	3.7	3.4	3.5	3.4
Excellent	1	2	2	2	5	1	13
Good	4	2	3	1	3	1	12
Fair	0	0	0	0	1	0	1
Poor	0	0	0	0	0	0	0
Voice reply quality—							
Wtd. Avg.	3.4	3.5	3.4	3.0	3.3	2.5	3.3
Excellent	2	2	2	2	5	0	13
Good	3	2	3	0	4	1	12
Fair	0	0	0	0	0	1	2
Poor	0	0	0	1	0	0	1
Hardware reliability—							
Wtd. Avg.	3.8	4.0	2.8	2.0	3.4	3.5	3.3
Excellent	4	4	2	1	5	1	16
Good	1	0	1	0	3	1	7
Fair	0	0	1	0	1	0	2
Poor	0	0	1	2	0	0	3

	IBM 7770	IBM S/7	Periphonics	Votrax	Wavetek	Others	All
Promptness of maintenance—							
Wtd. Avg.	3.2	3.8	2.2	2.0	2.7	4.0	2.8
Excellent	2	3	0	1	1	1	8
Good	2	1	3	0	5	0	11
Fair	1	0	0	0	2	0	3
Poor	0	0	2	2	1	0	5
Quality of maintenance—							
Wtd. Avg.	3.4	3.8	2.4	2.5	3.0	4.0	3.2
Excellent	2	3	0	1	3	1	10
Good	3	1	3	0	4	0	7
Fair	0	0	1	0	1	0	2
Poor	0	0	1	1	1	0	3
Mfr's. software—							
Wtd. Avg.	2.6	2.3	2.8	—	2.0	3.0	2.6
Excellent	0	0	0	—	1	0	1
Good	3	2	4	—	0	1	10
Fair	2	1	1	—	2	0	6
Poor	0	1	0	—	0	0	1
Mfr's. technical support—							
Wtd. Avg.	2.8	2.0	2.0	2.5	3.3	3.0	2.5
Excellent	1	0	0	1	2	0	4
Good	2	1	2	0	1	1	7
Fair	2	2	1	0	1	0	6
Poor	0	1	2	1	0	0	4
System worked as promised:							
Immediately	4	2	0	2	6	1	18
Eventually	1	2	3	0	2	0	6
Never	0	0	0	0	0	0	0

Figures presented opposite the ratings of Excellent, Good, etc. are the number of responses. Wtd. Avg. means weighted average and is calculated based on a 4, 3, 2, 1 weighting assignment.

Table 2. Users' ratings of voice response systems

Modems: Terminal—Line Linking Devices

Problem:

A telephone line can more than adequately handle low-frequency analog signals, of which voice is an easily recognizable example, but it gets very poor marks for handling the high-speed pulse-type signals spewed out by digital equipment like computers and most terminals. But short of getting into some still very expensive alternative transmission lines, the telephone line can be made to do a reasonably decent job of transmitting digitally-encoded data (not necessarily the actual pulses themselves) by using interface devices to convert the digital pulses from the transmitter into analog signals for the telephone line and then reconverting from analog signals back to digital signals at the receiver. The basic tradeoff is the convenient availability and low cost of the telephone line in exchange for the inherently very high speed of purely digital traffic. The devices that do this little bit of digital-to-analog-to-digital electronic magic are called modulators/demodulators, or simply modems.

Most modems work within the limitations of the ordinary voice-grade telephone line, which means, roughly, that any frequency below about 300Hz and above about 3000Hz is lost to the receiver even though it may have been produced by the sender. Modem designers have really done remarkably well within these severe frequency constraints; the newest modems can push digital data into an ordinary telephone line at a rate of 9600 bits (1200 8-bit characters) per second (bps). Some modems can push the telephone lines even harder, but the error rate begins to rise alarmingly above 9600 bps. True, this 9600 bps restraint slows the mighty computer to an ignominious crawl—an unleashed computer can easily send data 1000 times faster—and much of the current technological pressure in communications is focused directly on the problem of how to break that restraint while holding on to the convenience and low cost of the telephone line. But until we get there—maybe sometime in the mid-1980's—we would strongly advise you to get acquainted with modems.

Solution:

The year 1979 did not have the merger/acquisition activity of other recent years and the technology advances and pricing also appear to have slowed down. The use of 16-bit microprocessors within modems has begun with the expanded instruction set being used for more and more diagnostic capability.

Modems: Terminal—Line Linking Devices

AT&T and IBM have announced new microprocessor based modems with impressive diagnostic capability for data transmission of 2400 to 9600 bps. The Codex subsidiary of Motorola began deliveries of its 2400 full-duplex dial-up modem and has announced that this device will be incorporated into a voice privacy terminal that includes a voice digitizer for secure voice applications. Another Motorola subsidiary, Universal Data Systems, has announced a new 103-compatible modem with an interesting wrinkle; it does not require AC operating power—it derives it from the telephone line (entirely legal and energy conservation minded even though the end result of power consumption is approximately the same).

Definitions

There are a number of different transmission devices on the market. In some cases, the differences are obvious but in others, they are more subtle. The criteria used by Datapro to classify transmission devices in one of the listed categories is generally accepted by the industry and is explained in the following paragraphs.

A modem is a device that accepts data from a computer or terminal device in the form of a digital signal and transforms the data into a form more suitable for transmission over a communications facility; another modem at the other end of the line, restores the data to its original form. The transformation and restoration process is called modulation and demodulation, which is contracted to modem.

The Low-, Medium-, and High-Speed modems listed in the charts operate over 3002 (or equivalent) voice grade lines at essentially unlimited distances. The Wideband category consists of two classes: those that operate over group or supergroup voice multiplexed links at 19.2K bps and up and those that perform an inversed multiplexing function. In the latter, a high speed serial interface (e.g., AT&T 300 Series) from the business machine equipment (i.e., terminal, front end, DTE) is segregated into individual bit streams that are analog modulated and applied to multiple 3002 facilities (e.g., one 19.2K bps stream inverse multiplexed to two 9600 bps signals) and then recombined at the distant end to provide transparent communications. This generally requires some overhead for synchronization (e.g., the 9600 bps throughput may operate at 10.8K bps).

The Telephone Coupler section covers devices that are acoustically or inductively coupled to the line through a standard telephone handset without any hard-wired connection. The Parallel Interface category covers those modems that have a parallel digital input, a parallel analog output, or both. These devices are generally used with paper tape readers, punched card readers, magnetic card readers, or with touch-tone signalling in such applications as "tellerless" banking.

The Limited Distance Data Set section covers three types of transmission elements: short haul modems, line (cable) drivers, and modem eliminators. The short haul category employs analog modulation and as such can be routed over standard TELCO facilities that include amplifiers, equalizers, compandors, etc. These machines use low-cost equalizers and as such are limited in the number of frequency translations (up/down microwave radio) that they can tolerate. A distance of 50 miles at 4800 bps is not uncommon. The line (cable) driver is a device that requires DC continuity (no amplifiers, etc.) and many times cannot tolerate loading coils that are placed in the circuit to enhance voice transmission. This low-cost equipment generally employs "baseboard" modulation, usually DC phase shifted, with or without encoding. These drivers are, therefore, not modems in the strict sense of the definition because they do not produce analog outputs. The last category is that of modem eliminators. This equipment is aptly named in that one such device replaces two modems. The equipment usually is located midway between two DTE equipments and serves to amplify data and control signals as well as providing the crossover function (e.g., XMT/REC data; XMT/REC synchronous clock; RTS/CTS turnaround, etc.) necessary to make the DTE's believe that they are operating through modems.

Data Access Arrangements

With the Carterphone decision in 1969, users were permitted to connect foreign (non telephone company) modems to the public switched network. However, the operating companies required the use of a data access arrangement (DAA) connected between the foreign modem and the network. This was required based on the phone company's position that the network must be protected from abnormally high signal power and also electrical protection from short circuits, etc. The independent modem manufacturers complained bitterly that the DAA was anti-competitive because it added cost ($2 to $10/month) on their devices, which the telephone company's modems did not require. The independents took their case to the FCC and 1975 that body decided that a certification program should go into effect whereby each manufacturer would submit a sample modem to an independent testing laboratory that would verify that the design of the machine was suitable for direct connection to the DDD. Effectively, the DAA circuitry was incorporated into the modem. This was a major defeat for the telephone companies and to make matters worse, their own modems were required to be certified as well.

Modems: Terminal—Line Linking Devices

When the Bell system certified the last of their modems, they "grandfathered" the DAA and no longer actively leased them after July 1979. This does not present a problem to the user except where they have purchased a non-certified modem that requires a DAA and a physical move is required (such as relocating a computer center or expanding a time sharing network). There are a number of independent manufacturers of DAA equivalents, which have been certified. In some cases, the DAA will only operate with that firm's modems; in other cases, complete Bell replacements are offered.

Automatic Calling Units (ACU's)

The Bell system leases an ACU for dial pulse central offices (801A) and also for touch tone offices (801C). In its present design, the ACU interfaces with a CPU front end via RS-366 parallel interface. One front end port is required for each ACU and one ACU is required for each modem to automatically establish a connection to a phone number stored in the front end. The applications for these devices are myriad and include off hour (night) data retrieval for order entry, inventory, etc. Using the Bell ACU's can be a costly operation because a port is required for each ACU and it can only control one modem. Independent vendors supply 801A/801C equivalents and many of them can control more than one modem. This is accomplished by establishing the connection and passing control to the modem; thereafter, the ACU goes off-hook on another business line and establishes another connection for another modem. Many of these independent devices can also accept a serial RS-232-C input in lieu of the parallel RS-366 interface. This allows a dialer to be located in a city distant from the front end over a multiplexer link and can save a considerable amount of money by allowing less expensive WATS bands or reduced toll charges.

Optical Transmission

Communications using light as the transmission vehicle have probably been around almost as long as the sun. Practical application however has not been forthcoming until recently. Using air as the medium, light transmission is limited to line-of-sight applications, requires very accurate alignment, and is severly limited by weather conditions. The development of the laser in the early '60's provided the design engineer with an accurate and powerful light source but the perfection of a transmission medium lagged. In 1970, fiber optics using low-loss glass strands was perfected and a wide range of applications opened up including:

- Long and short haul telephone circuits;
- CATV;
- Process control;
- Military;
- Broadband wiring within buildings; and
- Data communications.

This technology is now out of the laboratory and has been implemented in several Bell and GT&E operating telephone companies, in shipboard and aircraft military applications; and in inter-/intra-building communications for large firms. The most common light sources used are Light Emitting Diodes (LED's) and lasers. The receivers are generally standard photo-electric diodes. These devices have been around for some time but it remained for the transmission medium to be perfected before feasible communications could occur. The original optical fibers were constructed of plastic but the attenuation introduced was too great for all but a handful of applications. Practical telecommunications became possible in 1970 when Corning developed a glass fiber with attenuation of less than 20 db per kilometer. Present day fibers introduce less than 10 db/km. The principal advantages of light communications include:

- Large bandwidth—depending on the fiber used, bandwidth as high as 1 billion Hz can be achieved.
- Low loss—systems with repeater separation of up to 10 km are operational and the theoretical limit could be as high as 50 km.
- Operating speeds up to 20M bps with error rate of 1 x 10^9 or better.
- Immunity to electromagnetic interference—fibers do not radiate interfering signals and are not susceptible to magnetic radiation.
- Security—it is all but impossible to "tap" fiberguide without being detected.
- Electrical isolation—no electronic common ground is required.
- Small size, light weight, low power requirement, extended environmental operating conditions.

The state of the fiber optics art is in considerable flux and one of the areas that must be addressed is that of standards. Various manufacturers use different definitions and parameters in specifying hardware capabilities and it is difficult to make comparisons. In general, LED transmitters are less expensive than lasers but the laser has a wider bandwidth, a higher operating rate, and higher output power.

Modems: Terminal—Line Linking Devices

Policy Changes

In addition to the "grandfathering" of the DAA, several other important changes to AT&T offerings have been implemented. Advanced Communications Service (Report C31-046-801) has been further delayed by what has been identified as "software problems". Effective in late 1979, Bell has been released from the restraint of providing international data communications and can now provide data as well as voice service from the states to overseas locations. A drastic change in WATS and MTS service rates is anticipated for an effective date about the time this report is published. In a radical departure from previous rules, Bell has allowed "interpositioning" of foreign hardware between "end points" of TELCO provided facilities; i.e., formerly, if any modem on a leased line was of non-TELCO manufacture, all modems on the line had to be of non-TELCO manufacture. Similarly, if a TELCO supplied terminal (e.g., Dataspeed 40) was leased, a TELCO-supplied modem was required to interface the terminal with the communications facility. In an apparent effort to secure a higher percentage of the market, Bell has relaxed both of these rules and now allows "mixed" modems on a leased line and "foreign" modems interfaced to their terminals.

The FCC has again rejected the rates for Digital Dataphone Service (DDS) and Digital Switched Dataphone Service (DSDS). Wideband 5000 Series channels have been effectively discontinued and the fate of the lower speed 8000 channels remains in question.

Another change that has been implemented rather drastically is the role of Bell's Purchased Products Division. Under pressure from the FCC, Justice, and others, AT&T has established a policy of going to independent vendors for some items as opposed to buying in-house from Western Electric. Several items come to mind including the LADS limited distance data set (Codex manufacture), the Netcon-5 Network Management Center (General DataComm Industries), nodal processors for ACS (Digital Equipment Corp. reportedly in the driver's seat), and the COMSTOR II intelligent storage system (Sykes Datatronics). In addition to supplying these and other equipment of independent manufacture, competition has stimulated the design/development of a number of new AT&T offerings. Among them are the Dataphone II Series of diagnostic modems, the 212A full-duplex 1200 bps modem (in response to the Racal-Vadic device), and improved versions of the entire data set line along with the overdue retirement of many older devices.

The Need for Diagnostics

The successful operation of a data communications network depends on the ability of those in charge to quickly spot failures and to diagnose the trouble accurately. Equipment for diagnosing data transmission problems is available in the form of integral modem features, portable test units, or central-site monitor and diagnostic units for large networks. Diagnostic capabilities range from basic functions such as local and/or remote loopback to very sophisticated forms of testing that analyze the quality of the received signal. The degree of sophistication for any specific network must be determined by the user. Obviously, cost tends to increase in proportion to the level of testing capability.

The most common form of diagnostic testing is the loopback function, which allows the user to perform a complete test of the data link, including the local modem, remote modem, and interconnecting telephone facilities, all from one site. The test is performed by returning or "looping back" the output of the transmitting device to its input via the local modem interface (local digital loopback), the local modem itself (local analog loopback), the opposite end of the communications line (remote analog loopback), or the remote modem (remote digital loopback).

Other, more sophisticated measures for diagnosing communications failures include such features as a self-test capability for determining the operational status of the modem before connecting it into the data communications network, and random pattern generation and detection for diagnosing errors resulting from random noise, harmonic distortion, and phase jitter—all specific transmission problems that influence the quality of the transmitted signal.

Central-site diagnostic equipment is available that will automatically run on-line diagnostic tests of communications facilities, remote modems, and terminals without operator intervention at the remote sites. Intertel manufactures one such device that can accommodate as many as 12 multipoint lines with 40 points on each line.

Various types of portable test equipment are available. One in particular, manufactured by Infotron, is connected between a modem and terminal. Test results such as error counts, turnaround time, and percent distortion are presented via a light-emitting diode (LED) display. Racal-Milgo and others also market portable diagnostic equipment that provides a complete range of network diagnostic capabilities.

Modems: Terminal—Line Linking Devices

Microprocessor-Based Modems

Modems built around a microprocessor offer distinct advantages over other modems, including improved equalization capability, increased reliability, and added flexibility. By using the powerful capabilities of a microprocessor, more sophisticated equalization algorithms can be implemented, allowing equalization to be performed automatically on changing line conditions. This overall improvement in equalization minimizes the errors due to line distortions. The reliability of these new units is improved by eliminating some of the circuitry now required in conventional designs. In addition, microprocessor-based modems can be reconfigured to handle changing operating requirements, thus improving the flexibility of the units. This flexibility is an advantage to the vendors as well as the users because of the increased longevity of the product line and the lower production costs. (In quantity purchases, 8-bit microprocessors can now be obtained for less than $10 and 16-bit devices for approximately $12 each.)

Integral Modems

For some time, it has appeared that the neatest solution to the whole modem problem is simply building the modems into the business machine equipment. Technically, it is not difficult. Some savings over present, separate arrangements should be available because of elimination of some duplicated or unnecessary circuitry and cabinetry. Some terminals, notably Teletype's Model 43 typewriters, currently offer integral modems, and the increasing demand for single-source suppliers should convince other terminal makers to adopt the built-in modem approach. But the growth rate has been slower than expected.

Feelings among the modem makers vary as to the importance of the built-in modem concept in future years. No small part of the problem is the necessity to generate a special design for each different terminal. General-purpose, free-standing modems, by adhering to standard interface specifications, are assured of wide application. But there are no standards for logic-level interfaces, which is what the modem maker normally has to work with if he builds an integral modem. Thus, each design would have limited application.

The next logical question is that if it's such a good idea to build modems into terminals, then why don't the terminal manufacturers take the initiative and develop their own? After all, it can't be a terribly difficult design problem when scores of other companies have already solved it. Part of the answer may be simply that the terminal makers don't want to dilute their efforts in any way; in the economic situation of today, many are already stretched painfully thin as a result of the emphasis on leasing and nationwide marketing and service. A more fundamental reason is that built-in modems would raise the base price of the company's equipment. And price is all-important today. Look at price first and features second; this is the guiding philosophy in many markets, not just the computer field. A company's position is generally weakened in a competitive situation if it has to talk about overall considerations, such as comparing competitors' products plus the cost of an independent modem with its own integrated package.

USER EXPERIENCE AND RATINGS

In November 1979 Datapro sent its subscribers a reader survey form on modems. By our editorial cut-off date, December 17, we had received usable replies from 272 users of modems. Because most users reported on multiple models, these 272 user replies added up to a total of 786 individual responses on a total of 26,702 modems.

The ratings assigned by the users to the various modem models they reported on are summarized in the accompanying pages of "Users' Ratings" charts. Where the number of responses permitted, the entries are separated by model number. For many of the modem suppliers, a general entry, "Other and unspecified," is included; these entries include models not listed individually as well as responses in which the model could not be exactly identified.

The overall ratings obtained in this year's survey are virtually identical to those obtained in last year's survey. As was the case last year, the ratings obtained for Diagnostic Capabilities were considerably lower than those obtained for the other three categories.

Included on the reader survey form were questions regarding the usage of various modem features, operating speeds, communications facilities, carriers, and problem areas encountered by the users. The results obtained in this area are also similar to those obtained in last year's survey.

The various features used and the corresponding number of responses received for each are presented in the following:

Feature	Number of Responses
Reverse channel	49
Secondary channel	35
Multispeed	160
Split stream	79
Multipoint	193
Automatic equalization	245

Modems: Terminal—Line Linking Devices

Feature	Number of Responses
Alternate voice/data	92
Auto answer	255
Auto dial	41
Local diagnostics	446
Remote diagnostics	165

The various operating speeds that were utilized are as follows:

Speed (bits per second)	Number of Responses
0-300	129
600-1800	143
2000/2400	189
3600	8
4800	207
7200	19
9600	167
19,200	7
Other	3

The number of responses totals more than the overall figure of 786 because some modem models were used at several different speeds.

Users were also asked to identify the primary type of communications facility over which they were operating, and the source (carrier). The results are as follows. (The number of responses totals more than the overall figure of 786 because in some cases more than one communications facility and/or source was used.)

Communications Facility	Number of Responses
Dial Network (DDD):	
Telephone company DAA	217
Independent DAA	23
Certified modem	111
Acoustic coupler	27
Leased voice line:	
Unconditioned	262
C1 conditioning	34
C2 conditioning	55
C3 conditioning	8
C4 conditioning	1
D1 conditioning	61
D2 conditioning	6
Other	60

Communications Carrier	Number of Responses
Bell System	702
Western Union	13
MCI	6
SPC	10
Independent telco	43
Private network	23
Other	23

Users were also asked to list the conditions that caused major difficulties; the results are as follows:

Areas of Major Difficulties	Number of Responses
Line quality	287
Line outages	210
Hardware reliability	72
Equalization	30
Mulitpoint operation	37
Transmission errors	101
Diagnosis of problems	213
Other	14

512 users indicated that they had leased their modems, while 279 users indicated that they had purchased them.

Some users with multiple units had purchased some and also leased some, accounting for the additional number of responses exceeding the overall figure of 786 responses. A total of 654 users indicated that the vendor serviced their modems, while 94 stated that they performed their own service, and 44 indicated that they received service via a third party. Some serviced their own equipment occasionally even though it is also covered under a maintenance contract; this accounts for the additional responses.□

Modems: Terminal—Line Linking Devices

USERS' RATINGS OF DATA COMMUNICATIONS MODEMS

Modem Manufacturer and Model	Number of User Responses	Number of Modems in Use	Users' Ratings: Overall Performance WA	E	G	F	P	Diagnostic Capabilities WA	E	G	F	P	Hardware Reliability WA	E	G	F	P	Maintenance Service WA	E	G	F	P
Anderson Jacobson—																						
242	6	75	3.0	1	4	1	0	2.0	0	2	0	2	3.0	0	6	0	0	3.0	0	5	0	0
1200 Series	4	6	3.8	3	1	0	0	2.0	1	0	1	2	4.0	4	0	0	0	2.5	0	1	1	0
Others & unspecified	5	10	3.8	4	1	0	0	1.5	0	1	0	3	3.8	4	1	0	0	3.2	1	4	0	0
Subtotals	15	91	3.5	8	6	1	0	1.8	1	3	1	7	3.5	8	7	0	0	3.0	1	10	1	0
Astrocom, all models	8	113	3.1	2	5	1	0	2.8	2	1	3	0	3.0	2	3	0	1	2.8	1	3	0	1
Bell System—																						
103A	5	9	3.8	4	1	0	0	2.4	1	2	0	2	3.6	4	0	1	0	3.2	2	2	1	0
103J	14	71	3.5	8	5	1	0	2.8	4	4	3	2	3.6	8	6	0	0	3.3	5	7	1	0
Other & unspec. 103	5	358	3.4	2	3	0	0	1.8	0	1	2	2	3.4	2	3	0	0	2.6	1	2	1	1
103 subtotals	24	438	3.5	14	9	1	0	2.5	5	7	5	6	3.5	14	9	1	0	3.0	8	11	3	1
113A	5	30	4.0	5	0	0	0	2.8	0	3	1	0	4.0	5	0	0	0	3.4	3	1	1	0
113B	7	180	3.3	3	3	1	0	2.0	1	2	0	4	3.4	3	4	0	0	3.4	3	4	0	0
113D	5	14	3.6	3	2	0	0	3.0	2	1	2	0	3.2	1	4	0	0	3.4	2	3	0	0
113 subtotals	17	224	3.6	11	5	1	0	2.5	3	6	3	4	3.5	9	8	0	0	3.4	8	8	1	0
201A	6	41	2.8	1	4	0	1	1.7	0	1	2	3	3.3	3	2	1	0	2.2	0	2	3	1
201C/2400	71	881	3.5	37	31	2	1	2.6	11	29	19	10	3.5	42	27	1	1	3.1	25	32	7	5
Other & unspec. 201	4	202	3.0	0	4	0	0	2.0	0	1	2	1	3.3	1	3	0	0	2.8	0	3	1	0
201 subtotals	81	1124	3.4	38	39	2	2	2.5	11	31	23	14	3.5	46	32	2	1	3.4	25	37	11	6
202C	3	11	3.7	2	1	0	0	1.0	0	0	0	3	3.7	2	1	0	0	3.3	1	2	0	0
202S	15	249	3.5	9	5	1	0	2.3	1	5	7	2	3.7	11	4	0	0	3.1	4	8	3	0
202T	12	770	3.4	5	7	0	0	2.8	1	7	4	0	3.5	6	6	0	0	3.2	4	6	2	0
Other & unspec. 202	3	18	3.7	2	1	0	0	2.7	1	0	2	0	3.7	2	1	0	0	3.3	1	2	0	0
202 subtotals	33	1048	3.5	18	14	1	0	2.4	3	12	13	5	3.6	21	12	0	0	3.2	10	18	5	0
208A	41	599	3.5	20	20	1	0	2.6	6	16	14	5	3.5	22	19	0	0	3.0	11	23	5	2
208B/4800	40	348	3.5	20	18	1	0	2.6	3	18	14	3	3.6	22	17	0	0	3.1	14	16	5	2
Other & unspec. 208	10	354	3.3	4	5	1	0	2.4	2	2	4	2	3.2	5	3	1	1	2.9	3	4	2	1
208 subtotals	91	1301	3.5	44	43	3	0	2.5	11	36	32	10	3.5	49	39	1	1	3.1	28	43	12	5
209A/9600	38	405	3.5	22	14	2	0	2.8	7	18	9	3	3.5	21	14	2	0	3.2	16	14	7	1
212A/1200	17	100	3.8	13	4	0	0	3.2	6	8	3	0	3.7	12	5	0	0	3.5	9	8	0	0
Other Bell modems	13	8270	3.7	9	4	0	0	2.8	2	5	2	1	3.5	6	4	1	0	3.3	5	4	2	0
Bell subtotals	314	12,910	3.5	169	132	10	2	2.6	48	123	90	43	3.5	178	123	7	2	3.1	109	143	41	13
Burroughs, all models	10	35	2.9	0	9	1	0	2.7	0	7	1	1	2.9	0	8	1	0	2.9	0	8	1	0
Codex—																						
LSI 48/4800	18	256	3.4	11	4	2	1	3.1	5	11	1	1	3.3	9	6	2	1	2.7	4	7	5	2
LSI 72/7200	3	6	3.7	2	1	0	0	3.7	2	1	0	0	4.0	3	0	0	0	3.7	2	1	0	0
LSI 96/9600	22	362	3.6	15	6	1	0	3.0	6	9	4	1	3.7	14	5	1	0	3.0	6	9	4	1
8200	12	385	3.6	7	5	0	0	2.8	3	4	3	1	3.6	7	4	0	0	2.9	2	5	3	0
Others & unspecified	10	195	3.6	8	1	0	1	2.9	2	2	3	0	3.4	4	2	1	0	3.1	3	2	2	0
Subtotals	65	1204	3.6	43	17	3	2	3.0	18	27	11	3	3.5	37	17	4	1	2.9	17	24	14	3
ComData, all models	4	17	3.3	3	0	0	1	2.3	1	0	1	1	3.3	3	0	0	1	4.0	2	0	0	0
Gandalf—																						
LDS 120	4	82	3.5	3	0	1	0	2.0	0	0	4	0	3.3	2	1	1	0	3.5	1	1	0	0
LDS 309	6	86	4.0	6	0	0	0	3.2	2	3	1	0	3.8	5	1	0	0	3.8	4	1	0	0
Others & unspecified	10	259	4.0	10	0	0	0	2.6	0	5	3	0	3.9	9	1	0	0	3.8	7	2	0	0
Subtotals	20	427	3.9	19	0	1	0	2.7	2	8	8	0	3.8	16	3	1	0	3.8	12	4	0	0
General DataComm—																						
LDM-1	3	32	3.7	2	1	0	0	3.0	1	1	1	0	4.0	3	0	0	0	4.0	1	0	0	0
103	3	1061	2.7	0	2	1	0	2.7	0	2	1	0	2.3	0	1	2	0	2.7	0	2	1	0
201	5	72	3.8	4	1	0	0	3.0	2	2	0	1	3.2	2	2	1	0	3.2	1	4	0	0
202	4	63	3.8	3	1	0	0	2.5	0	3	0	1	3.8	3	1	0	0	2.5	0	3	0	1
208	4	171	3.5	2	2	0	0	3.0	1	2	1	0	2.8	1	1	2	0	2.5	0	2	2	0
Others & unspecified	8	169	3.8	6	2	0	0	2.9	2	4	1	1	3.6	5	3	0	0	2.5	0	3	3	0
Subtotals	27	1568	3.6	17	9	1	0	2.9	6	14	4	3	3.3	14	8	5	0	2.7	2	14	6	1
GTE/Lenkurt, all models	5	28	3.8	4	1	0	0	2.6	1	2	1	1	3.8	4	1	0	0	2.6	0	4	0	1
IBM—																						
3603	3	72	4.0	3	0	0	0	1.7	0	1	0	2	4.0	3	0	0	0	2.0	1	0	0	2
3872	7	106	3.0	1	5	1	0	2.4	1	3	1	2	3.1	2	4	1	0	2.9	1	4	2	0
3875	3	10	3.7	2	1	0	0	2.0	0	0	3	0	3.3	2	0	1	0	2.7	0	2	1	0
Others & unspecified	4	14	3.3	1	3	0	0	2.5	1	1	1	1	3.3	1	3	0	0	3.0	1	2	1	0
Subtotals	17	202	3.4	7	9	1	0	2.2	2	5	5	5	3.4	8	7	2	0	2.7	3	8	4	2

LEGEND: E—Excellent; G—Good; F—Fair; P—Poor; WA—Weighted Average based on a weighting of 4 for Excellent, 3 for Good, 2 for Fair, and 1 for Poor.

Modems: Terminal—Line Linking Devices

USERS' RATINGS OF DATA COMMUNICATIONS MODEMS

Modem Manufacturer and Model	Number of User Responses	Number of Modems in Use	Users' Ratings: Overall Performance					Diagnostic Capabilities					Hardware Reliability					Maintenance Service				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
Intertel, all models	5	418	3.4	2	3	0	0	2.8	1	3	0	1	3.0	1	3	1	0	1.8	0	0	3	1
Multi-Tech, all models	4	10	3.0	0	4	0	0	2.0	0	0	1	0	3.3	1	2	0	0	2.0	0	0	1	0
Paradyne—																						
MP48/4800	6	238	3.7	4	2	0	0	3.4	3	1	1	0	3.8	4	1	0	0	3.2	1	4	0	0
LSI 96/9600	13	142	3.3	6	6	0	1	2.8	2	6	3	1	3.2	6	3	2	1	2.5	2	5	1	3
M96	5	52	3.6	4	0	1	0	3.6	4	0	1	0	3.8	4	1	0	0	2.8	1	3	0	1
Others & unspecified	8	83	3.1	2	5	1	0	2.9	2	2	3	0	3.1	2	4	1	0	2.9	2	3	1	1
Subtotals	32	515	3.4	16	13	2	1	3.0	11	9	8	1	3.4	16	9	3	1	2.8	6	15	2	5
Penril—																						
PSH 24/48/72/96	4	18	3.8	3	1	0	0	2.3	0	1	3	0	3.5	2	2	0	0	2.3	1	0	2	1
Others & unspecified	9	22	3.6	6	2	1	0	2.6	2	1	3	1	3.5	5	2	1	0	3.1	4	2	1	1
Subtotals	13	40	3.6	9	3	1	0	2.5	2	2	6	1	3.5	7	4	1	0	2.8	5	2	3	2
Prentice—																						
SLD	4	608	3.5	2	2	0	0	3.0	0	3	0	0	3.7	2	1	0	0	2.7	0	2	1	0
Others & unspecified	11	577	3.5	5	6	0	0	2.3	1	3	3	2	3.3	3	8	0	0	2.5	0	5	6	0
Subtotals	15	1185	3.5	7	8	0	0	2.5	1	6	3	2	3.4	5	9	0	0	2.5	0	7	7	0
Racal-Milgo—																						
24 LSI	11	1059	3.5	7	3	1	0	3.8	7	2	0	0	3.3	5	2	2	0	3.0	3	5	1	1
24 LSI Mark II	7	198	3.6	5	1	1	0	3.0	1	5	1	0	3.6	5	1	1	0	3.3	2	5	0	0
MPS 48	16	336	3.4	6	10	0	0	3.0	6	6	2	2	3.3	5	10	1	0	2.9	4	8	3	1
96/9600/MM	22	294	3.6	14	8	0	0	3.5	14	6	2	0	3.5	10	12	0	0	2.9	6	9	4	2
2200/24	3	120	3.3	2	0	1	0	2.7	1	1	0	1	3.3	2	0	1	0	3.0	0	3	0	0
33/36	4	57	3.8	3	1	0	0	2.8	1	1	2	0	3.5	3	0	1	0	2.8	1	2	0	1
4500/48	8	77	3.5	4	4	0	0	2.4	2	0	5	1	3.6	5	3	0	0	2.8	1	4	3	0
4600/48	5	61	3.8	4	1	0	0	2.6	2	0	2	1	3.8	4	1	0	0	3.4	3	1	1	0
4800/72	7	27	3.4	3	4	0	0	1.9	0	1	4	2	3.6	4	3	0	0	3.0	0	5	0	0
5500/96	4	26	3.5	2	2	0	0	2.3	0	2	1	1	3.5	2	2	0	0	2.0	1	0	0	2
Comlink II	5	103	3.2	1	4	0	0	2.4	0	2	3	0	2.8	1	2	2	0	2.6	0	4	0	1
Comlink III	5	186	4.0	5	0	0	0	3.0	1	3	1	0	4.0	5	0	0	0	3.3	1	3	0	0
Others & unspecified	12	119	3.3	6	5	0	1	3.1	4	6	1	1	3.4	7	4	0	1	3.0	4	5	2	1
Subtotals	104	2602	3.5	58	42	3	1	3.0	37	35	22	8	3.4	54	39	8	1	2.9	23	53	13	9
Racal-Vadic—																						
317	7	1891	3.9	6	1	0	0	3.3	4	1	2	0	3.6	4	3	0	0	3.2	1	4	0	0
Other 300 Series	4	103	3.3	2	1	1	0	2.8	1	1	2	0	2.8	1	2	0	1	2.7	0	2	1	0
1200 Series	3	319	3.7	2	1	0	0	3.3	1	2	0	0	3.7	2	1	0	0	4.0	1	0	0	0
3455	6	61	3.8	5	1	0	0	2.5	0	2	2	0	3.7	4	2	0	0	3.0	1	3	1	0
3467	9	101	3.7	6	3	0	0	3.0	4	0	2	1	3.4	5	3	1	0	3.1	2	4	1	0
Other 3400 Series	5	181	3.8	4	1	0	0	3.0	0	5	0	0	3.6	3	2	0	0	2.8	0	4	1	0
Others & unspecified	3	1722	4.0	3	0	0	0	3.7	2	1	0	0	4.0	3	0	0	0	4.0	3	0	0	0
Subtotals	37	4378	3.7	28	8	1	0	3.1	12	12	8	1	3.5	22	13	1	1	3.1	8	17	4	0
Rixon (Sangamo)—																						
201	5	54	3.8	4	1	0	0	2.5	1	1	1	1	3.6	4	0	1	0	2.7	0	2	1	0
208	4	80	3.5	2	2	0	0	2.8	0	3	1	0	3.5	2	2	0	0	2.8	0	3	1	0
Others & unspecified	9	148	3.8	7	2	0	0	2.7	1	5	2	1	3.7	6	3	0	0	2.9	1	6	2	0
Subtotals	18	282	3.7	13	5	0	0	2.6	2	9	4	2	3.6	12	5	1	0	2.8	1	11	4	0
Spectron, ME 81	4	7	4.0	4	0	0	0	1.0	0	0	0	1	3.5	2	2	0	0	2.5	0	1	1	0
Syntech—																						
201	8	23	3.0	3	3	1	1	2.6	2	2	3	1	2.9	3	3	0	2	2.7	2	2	2	1
208	6	8	2.3	1	2	1	2	2.7	1	2	3	0	2.2	1	1	2	2	2.8	1	3	2	0
Others & unspecified	7	110	3.4	4	2	1	0	2.7	3	1	1	2	3.1	4	1	1	1	3.6	5	1	1	0
Subtotals	21	141	3.0	8	7	3	3	2.7	6	5	7	3	2.8	8	5	3	5	3.1	8	6	5	1
Teledynamics, all models	4	90	3.5	2	2	0	0	2.8	0	3	1	0	3.5	2	2	0	0	2.7	0	2	1	0
UDS, all models	4	35	3.8	3	1	0	0	2.5	0	2	2	0	3.3	1	3	0	0	3.0	0	2	0	0
Ven-Tel, all models	4	35	3.5	2	2	0	0	2.0	1	0	1	2	3.0	2	1	0	1	1.5	0	0	1	1
Western Union, all models	3	7	3.3	1	2	0	0	2.5	0	1	1	0	3.5	1	1	0	0	3.0	0	3	0	0
All others	33	362	3.4	20	8	3	2	2.4	8	5	6	10	3.5	21	8	3	1	3.1	13	8	3	4
GRAND TOTALS	786	26,702	3.5	445	296	32	12	2.7	162	282	194	96	3.5	425	283	41	15	3.0	211	345	115	44

LEGEND: E—Excellent; G—Good; F—Fair; P—Poor; WA—Weighted Average based on a weighting of 4 for Excellent, 3 for Good, 2 for Fair, and 1 for Poor.

Line Use Optimization with Multiplexing Equipment

Problem:

Many communications terminals operate at relatively low speeds, for example 110 or 300 bps. Since a typical voice grade switched telephone line can comfortably handle 2400 bps, it follows that multiple low speed terminals could be transmitting over a single line simultaneously, provided the data streams of each terminal could be clearly identified. The device which provides this capability is called a multiplexer.

There are different types of multiplexers with varying capabilities and price ranges. This report explains these variations and also presents user ratings of several manufacturers' products obtained from a recent Datapro user survey.

Solution:

The traditional value of multiplexing has been savings on communications facility charges; the equipment concentrates data that otherwise would be routed over many separate lines into one composite stream that can be delivered over one link. Recent advances in technology, particularly microprocessors, have added an additional advantage to multiplexing: error correction.

Multiplexers can be separated into two major categories: Frequency Division Multiplexers (FDM) and Time Division Multiplexers (TDM). In addition, three classes of TDM's are analyzed. The principal differences between the FDM and TDM technologies and the advantages and disadvantages of each are presented in this report.

FDM Versus TDM

The multiplexing techniques utilized by FDM and TDM hardware are drastically different. FDM is generally used on AT&T 3002 voice grade lines; it subdivides the available bandwidth into frequency slots. The amount of bandwidth (frequency slot) that is assigned to each channel is dependent upon the speed at which that channel is to operate. Naturally, the bandwidth required increases in proportion to the operating speed. FDM can be viewed as a series of parallel modems, each operating at a different frequency and connected in parallel to the line. The basic function of moving data consists of translating a serial input bit stream into discrete audio tones for transmission over analog telephone lines. The "1's" and "0's" on the digital input are converted to Mark and Space tones at the analog output. At the distant end, the reverse process takes place; one end "talks" and the other end "listens". Most of the FDM hardware available today is equipped with plug-in filters and frequency determining networks. This allows the transmitter and receiver of any particular channel to operate at different frequencies, thus allowing full duplex (simultaneous two-way transmission) over a 2-wire facility. No other hardware is required.

The advantages of the FDM technique include: minimum common equipment (no modem or "sweep" electronics required since each channel operates inde-

Line Use Optimization with Multiplexing Equipment

pendently), and the ability to multipoint the hardware for polling applications at minimum cost. The principal disadvantage of FDM is that it is limited in speed and in the number of channels it can handle by the bandwidth available.

Control signal passage using FDM also gets tricky. Because of the bandwidth limitation, an automatic restriction is imposed on the number of controls that can be passed in-band (within the frequency allocation of the channel). The most simple method of passing a single control through an FDM channel is that of carrier on/off. This is satisfactory for some applications but not for most and it is subject to error due to frequency spill-over and noise.

More sophisticated approaches to control signal passage are available in some models. These include "warbling" the center frequency of the channel or passing all controls through a separate channel dedicated to that purpose. The warbling approach can be dangerous on noisy lines. The separate channel concept can be expensive because coder/decoder circuits are required, usually out-board of the channel card, to separate and direct the control functions to the proper data channel.

In contrast, TDM's are strictly digital devices. They accept multiple digital inputs and convert them to one composite digital output. If transmission is to be over conventional analog facilities, the TDM must direct its output through a modem to the line. With the advent of Bell's Dataphone Digital Service (DDS), a direct digital interface is possible and no modem is required. However, DDS is not universally available. DDS is presently scheduled to be operational in 96 cities. But even in those metropolitan areas, only certain telephone exchanges are served; consequently, a modem may be required just to get to the access point of the DDS network.

The input data to a TDM is loaded into a channel buffer and is swept out by common electronics. Each channel has its turn, or time slot, in the overall composite stream. Each sweep of the common electronics is traditionally called a frame. These frames are developed in different ways depending upon the equipment type, which can be identified as bit interleaved, traditional character interleaved, and statistical. These terms are reasonably self explanatory; the hardware will be described later in this report.

The principal advantages of the TDM are: the ability to handle high input speeds, a large number of inputs, and (in statistical multiplexers) error detection/correction. The TDM equipment can also pass, transparently, a number of control signals, which are necessary when the input channel is dial-in; e.g., timesharing applications.

A synopsis of the characteristics of FDM and TDM is presented in the following:

	FDM	TDM
Speeds accepted	Up to 600 bps	Virtually unlimited
Maximum number channels	Up to 32@50 bps	248 possible
Modem required	No	In most cases
Complexity common equipment	Minimum	Complex
Common equipment back-up	No	Yes
Dial back-up	Not normally	Possible
Full duplex 2-wire	Yes	No

From a cost stand point, both FDM and TDM average out at between $300-$400 per channel end, with the nod going to FDM on small systems and to TDM on larger systems. Obviously, the application also plays a large roll in the selection and may limit the buyer to one technology or the other. The speed limitation on FDM, or the number of channels to be serviced, or the number of controls that must be passed, figure into it.

TDM Analysis

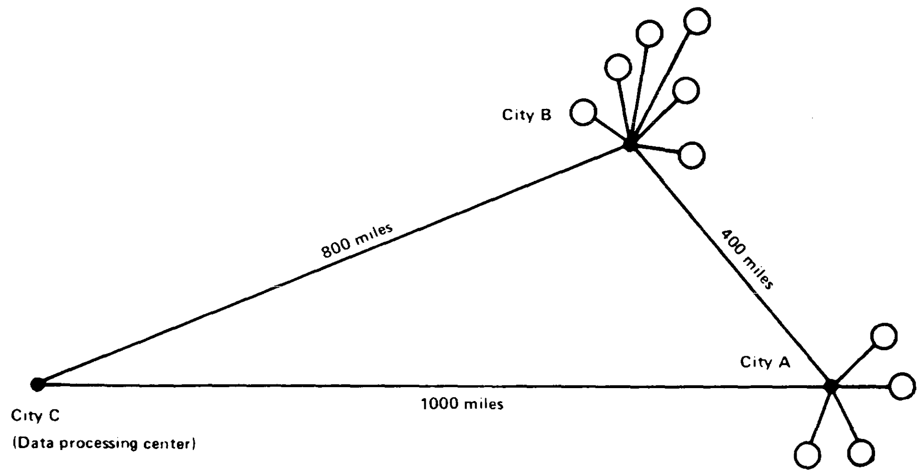
The equipment presently available using this technology can be classified into three groups: bit interleaved, traditional character interleaved, and statistical multiplexers. Each group has strong and weak points, which will be presented in the following paragraphs along with the differences in mode of operation. As a brief overview, the following table provides a comparison of some of the capabilities/limitations:

	Bit Int.	Char. Int.	Statistical
Apparent efficiency	90-98%	100-115%	200-400% (typ.)
Channel capacity	96	162	248
Synchronization time	Slow	Avg.	Fast
Strength of Synchronization	Weak	Good	Strong
Integrity of control signals	Weak	Strong	Strong
Autospeed detection	No*	Opt.	Opt.
Echoplex applications	Strong	Weak	Weak
Polling applications	Good	Strong	Weak

*Not required; transparent to rated speed

The observations presented in this table are based upon our knowledge of the hardware, its multiplexing technique, and the applications and user responses that we have encountered through surveys, interviews, and experience. They are group classifications;

Line Use Optimization with Multiplexing Equipment



Terminal-hours per day		DDD Costs, assuming calls average 10 minutes each		Leased Line Costs		Difference in Line Costs between DDD and Leased Charges	
From City A	From City B	Calls per month	Toll Charges per month	Two-lines (AC & BC)	Multipoint line (ABC)	Two-lines	Multipoint lines
4	6	1260	$4,568.40	$1,406.40	$1,010.40	$ 3,162.00	$ 3,558.00
10	20	3780	13,708.80	1,406.40	1,010.40	12,302.40	12,698.40
20	30	6300	22,932.00	1,406.40	1,010.40	21,525.60	21,921.60
40	60	12,600	45,864.00	1,406.40	1,010.40	44,457.60	44,853.60
80	100	22,680	82,756.80	1,406.40	1,010.40	81,350.40	81,746.40

This example illustrates the potential for savings through the use of multiplexing techniques over leased lines in place of Direct Distance Dialing. The indicated differences are for communications facility costs only and do not reflect the required multiplexing equipment, local lines between the terminals and the multiplexer, and modems. See the text for a complete explanation. The line costs are based on rates for daytime telephone usage and AT&T Series 2000 and 3000 Multi-schedule Private Line Service and are representative. It is assumed that cities A, B, and C are MPL category A cities.

naturally, any vendor or equipment model could be better or worse than the rating. The apparent efficiency and channel quantity numbers were obtained from the vendor survey conducted in August/ September 1979.

Bit Interleaved Multiplexers

As the name implies, this machine loads one bit at a time from each channel input to create a composite stream. Whenever a channel buffer has a bit in it, that bit representation becomes part of the composite; consequently, start and stop bits in asynchronous communications also get loaded, which accounts for the relatively low efficiency rating for the device. Most bit machines operate using a fixed frame so that when a channel is idle, a mark hold bit is entered into that channel's time slot in the frame—another reason for reduced efficiency. Although the synchronization "word" or byte may be selected very carefully to avoid its occurrence in a random data stream, finding it and verifying it usually takes longer than with other types of multiplexers. All time division multiplexers obtain synchronization in the same way. They "bit slide" the incoming data stream into a register where it is compared against the assigned sync. word. Upon finding the correct bit pattern, the multiplexer knows when it should occur again because it knows the length of the frame. After recognizing the proper pattern at the appointed time on the next frame, it is reasonably sure that synchronization has been achieved. Some models let it go at that, others wait for further confirmation. Because the device is bit interleaved, the composite stream received from the distant end is totally random, which is why it is difficult to achieve synchronization and to re-gain it when lost due to line outages or pertubations. For the same reason, passing control signals is a problem; not only because they decrease efficiency by occupying a portion of the frame, but also because of the difficulty and/or response time required to properly interpret the byte and to react.

But bit interleaved multiplexers obviously have advantages also. They are ideally suited to mix synchronous streams and, because a minimum delay is imposed, they are a natural selection for echoplex applications. They are also quite cost competitive.

Traditional Character Interleaved Multiplexers

These machines form the composite stream by reading a character from each channel buffer per scan or frame. They are fixed frame devices and the scan rate

Line Use Optimization with Multiplexing Equipment

is set to be slightly higher than the fastest input rate. When a channel buffer is not completely full, the time slot in the frame assigned to it refreshes the control information. In asynchronous applications, the multiplexer strips off the start and stop bits of incoming characters and thereby achieves an apparent efficiency of over 100 percent. The start and stop bits are automatically inserted at the receiving end, which can cause a problem.

Many systems use a long space disconnect option, particularly in time sharing. Because the receiving channel automatically inserts the stop bit, which is a mark, the long space (break signal) can not be passed transparently. This is countered by adding an indicator bit to each character to show if that character is data or control. One of the bit positions, or a certain combination of bits in the byte, indicates the long space situation and the system thus operates transparently. Character interleaved multiplexers have good synchronization time because the sync. word can be selected such that occurance in random data is remote. This is true because the frame time slots are filled a character at a time and all of the bit configurations for all of the codes are known; therefore, the sync. word can be assigned as a bit pattern that does not occur in normal asynchronous or even synchronous traffic. One application, that of clear text transmission for program loading etc., nullifies this conclusion because any bit pattern is possible—but synchronization is still very strong.

The character interleaved multiplexer was originally designed to accommodate asynchronous traffic, but most machines how have synchronous option cards. These devices are very good in polling applications but because of the delay inherent to character interleaving (2-4 character times, round trip) they are not usually acceptable for echoplex applications. Most of them pass three to five control signals in each direction so they are also widely used in dial-up applications. The price of these systems is reasonable, particularly at high quantity channel levels.

Statistical Multiplexers

These devices are relative newcomers but have gained wide acceptance for several reasons, among them efficiency and error control. The equipment is designed for interactive applications where there is idle or "dead" time between characters and/or transmissions. Also, transmission is usually half duplex. The sum of all of these dead times is taken advantage of by dynamically allocating trunk throughput only to active channels. In this way, many more devices can be serviced than with a traditional character interleaved, fixed frame device. The frame length is variable and is developed according to input activity. The statistical multiplexer contains a buffer where the input data is stored until the frame is developed and transmitted. A copy of the frame is retained in the buffer however, and permits retransmission in case an error is detected.

Statistical multiplexers use some form of bit oriented protocol on the high speed link that allows for error checking and achieves very strong synchronization. Most of them use a form of High-level Data Link Control (HDLC), which calls for a bit pattern of 01111110 that is called a flag and is used as the sync. word. To prevent this pattern from occurring in random data, a technique termed "zero bit insertion" is employed. Whenever the multiplexer receives an input of five consecutive 1's, it automatically inserts a 0 in the bit stream, thus precluding the possibility of six consecutive 1's that might otherwise be mistaken for a flag. Immediately following the synchronizing flag, most multiplexers insert supervisory information including the length of that particular frame (because frame length is variable). Channel data follows, and at the conclusion of the frame, the data is analyzed for bit structure and a Cyclic Redundancy Check is performed (this is actually happening continuously). The results of the CRC are coded into a 16-bit pattern termed a Frame Sequence Check (FSC), which is appended to the frame. The last item to add to the frame is a trailing flag. At the distant end, the FSC is compared to the receiving multiplexer's analysis of the incoming data. If the two concur, an ACK is returned to the sending multiplexer, which allows it to clear the acknowledged frame from memory to make room for more input. If the two FSC's do not check, ARQ or automatic request for retransmission occurs and the sending multiplexer retransmits the frame, which will be retained in memory until an acknowledgement is received.

The zero bit insertion technique and frame construction result in extremely strong synchronization performance. Of course, the zero bit insertion must be reversed at the receiving end such that anytime the pattern 111110 is received, the zero bit is striped. These devices are even used in applications where the efficiency improvement cannot be realized (such as continuous input data on all channels) just for the error correcting benefits.

Where true statistical multiplexing is implemented, flow control procedures are utilized. When the multiplexer buffer fills beyond a certain threshold (usually due to line problems that necessitate many retransmissions) the machine stops the attached terminals from inputting data either by X-ON/X-OFF procedures, which is a character set sent in the bit stream, or by lowering Clear To Send. Flow is reenabled when the buffer declines to normal fill.

Line Use Optimization with Multiplexing Equipment

USERS' RATINGS OF MULTIPLEXERS

Multiplexer Vendor and Model	Type	Number of Responses	Number of Units	Users Responses and Ratings																								
				Overall Performance					Hardware Reliability					Maintenance Service					Ease of Installation					Ease of Expansion				
				WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
Codex 910	Char.	3	8	3.0	1	1	1	0	3.3	1	2	0	0	3.0	1	1	1	0	3.3	1	2	0	0	2.0	1	0	0	2
Codex 920	Char.	6	28	3.0	1	4	1	0	3.2	1	5	0	0	2.8	0	5	1	0	3.0	1	3	1	0	3.3	3	2	1	0
Codex 6030	Stat.	5	29	3.5	2	2	0	0	3.3	1	2	0	0	3.2	1	3	0	0	3.3	1	2	0	0	3.3	1	2	0	0
Codex 6040	Stat.	5	20	3.6	3	2	0	0	3.4	2	3	0	0	3.4	2	3	0	0	3.6	3	2	0	0	3.4	3	1	1	0
Codex, others	—	6	20	3.8	5	1	0	0	3.7	4	2	0	0	3.3	2	4	0	0	3.2	2	2	1	0	3.0	2	2	0	1
Codex, totals	—	25	105	3.4	12	10	2	0	3.3	9	14	0	0	3.2	6	16	2	0	3.3	8	11	2	0	3.1	10	7	2	3
Digital Comm. Asc., all models	Stat.	3	9	3.0	0	3	0	0	2.0	0	0	3	0	2.0	0	0	3	0	4.0	3	0	0	0	4.0	3	0	0	0
Infotron 180	Char.	4	20	4.0	4	0	0	0	3.8	3	1	0	0	2.0	1	0	0	2	3.3	1	3	0	0	3.5	2	2	0	0
Infotron 780	Stat.	6	24	3.7	5	0	1	0	3.7	5	0	1	0	3.5	5	0	0	1	3.2	3	2	0	1	3.4	2	3	0	0
Infotron, others	–	3	12	3.0	1	1	1	0	3.0	1	1	1	0	2.3	1	0	1	1	2.3	1	0	1	1	3.5	1	1	0	0
Infotron, totals	—	13	56	3.6	10	1	2	0	3.5	9	2	2	0	2.8	7	0	1	4	3.2	5	5	1	2	3.5	5	6	0	0
Micom 800	Stat.	6	26	3.7	4	2	0	0	3.5	3	3	0	0	3.0	1	3	1	0	3.7	4	2	0	0	3.2	3	2	0	1
Timeplex T Series	Char.	7	93	3.4	4	2	1	0	3.1	3	2	2	0	1.9	0	3	0	3	3.1	4	1	1	1	3.0	3	2	1	1
All others	—	6	45	3.7	4	2	0	0	3.7	4	2	0	0	3.2	2	3	1	0	3.4	2	3	0	0	3.8	4	1	0	0
Grand Totals	-	60	334	3.5	34	20	5	0	3.4	28	23	7	0	2.9	16	25	8	7	3.3	26	22	4	3	3.3	28	18	3	5

LEGEND: E—Excellent; G—Good; F—Fair; P—Poor; WA—Weighted Average based on a weighting of 4 for Excellent, 3 for Good, 2 for Fair, and 1 for Poor for each response.

A statistical multiplexer is not generally suited for echoplex applications because, in addition to being character oriented, additional delay is incurred in developing the frame. Some machines have a near-end echoplex capability where the near end multiplexer echoes back the data that the terminal sent in, as opposed to traveling to the central site and being echoed back by the front end. This arrangement is usually acceptable considering the strong error protection techniques used.

Another area where trouble could be encountered is in a polled environment. Most terminals do not acknowledge a poll from the CPU with their address, they simply ACK. Because of the storage in the multiplexer due to frame development, it is possible for the CPU to get a response out of queue and not be able to identify the source. Most statistical multiplexers are microprocessor based and are competitively priced for most channel capacities.

Potential Dollar Savings

The accompanying diagram shows the requirements for a hypothetical system involving but three points. Assume that a central data processing facility is located in City C, together with multiple locations clustered around Cities B and A that need independent, rapid-turnaround communications with the processing facility in City C. Further assume that the locations clustered about Cities B and A are within local-call range of their respective cities.

Of the various possible approaches for configuring the system, we will consider three:

- Direct calls (DDD) from each location to the processing facility in City C, with the calls averaging 10 minutes in length.
- Leased voice-grade lines between Cities B and C and between Cities A and C. Each cluster of locations will be multiplexed over the respective leased line. For the sake of discussion, we will assume that line BC and line AC require no conditioning.
- A leased voice-grade line extending from A to B and then to C. No conditioning will be assumed.

The nature of the computations will be to see the differences in cost among the three arrangements for the communications facilities only. *The differences will show the amount of money that can be spent on multiplexing equipment to achieve the same total costs.* If multiplexing equipment can be acquired for less money, then that difference represents savings. The parameter used as a variable is terminal-hours, or the number of terminals multiplied by their average usage per day. This figure gives the cost of using DDD for communications. Use of the multiplexing equipment and leased lines is a fixed monthly cost and is independent of usage.

From the chart in the illustration, you can see that the potential savings are great indeed, even for

Line Use Optimization with Multiplexing Equipment

moderate terminal usage. Naturally, the calculations have been simplified to some extent. In this particular case, the greatest savings would probably be accomplished by routing one line from point A through B to C. In this arrangement, three "ends" would be required, one at each city. The multiplexing equipment would have to have multidrop capability, which would increase the cost of the individual stations, but probably not as much as the requirement for four "ends" for the arrangement using two leased lines into City C. The geographical locations would, of course, play an important role in determining the best configuration for a particular system. This method can be used to advantage to quickly gain perspective as to which of several arrangements offers the most potential.

One additional benefit of and reason for multiplexing is shown by this exercise. The entire cost differential between using DDD and leased lines can be saved by ignoring multiplexing and running a party line among all stations and the processing center at City C. But there are two serious drawbacks to this approach. One is that only one station at a time can communicate. This leads directly to the second limitation. At most, there are only 24 terminal hours available per day for the one line. Multiplexing, then, provides simultaneous data paths and greatly increases the capacity of a leased line in terms of the number of data exchanges that can be accommodated from low-speed terminals.

USER EXPERIENCE

A Reader Survey Form, included in the March 1979 supplement of both DATAPRO 70 and DATAPRO REPORTS ON DATA COMMUNICATIONS, requested that our subscribers share their experience with us concerning data communications multiplexers. We received a total of 48 usable replies by the editorial cutoff date in September. Some users had more than one type of multiplexer, so the replies represent a total of 62 product ratings. The total population reported on is 299 multiplexers.

The accompanying table presents a summary of the product ratings assigned by the responding users. The categories are self explanatory. As has been the case in previous surveys, the users appear generally satisfied with the overall performance, hardware reliability, ease of installation, and ease of expansion, but are somewhat less generous in their evaluation of the maintenance services of the vendors.

Several questions were included in the Reader Survey Form to determine usage patterns, including low-speed side line types and speeds and high-speed side line types and speed. All but 6 of the respondents indicated the type of facility used to connect the terminal to the multiplexer. About 40 percent of the respondents used dial-up and almost 80 percent used leased lines (some used both). In addition, an assortment of private facilities were used.

A query concerning the operating input speeds yielded the following results:

Low-Speed Inputs, bps	Number of Responses
Up to 75	1
110	4
134.5	5
300	37
600	1
1200	38
2400	23
4800	6
7200	1
9600	1

It is obvious from these totals that the trend toward higher speeds is a fact with 300 and 1200 bps operation almost equal. It is rather surprising that so little 110 bps operations was reported, because there are so many of these machines still in use. Based upon the number of 2400 bps inputs, it would seem that the statistical multiplexer has made appreciable inroads, at least with our subscriber base; most of those reporting 2400 bps applications were using statistical multiplexers. A few "pioneers" are operating with inputs of 4800 bps and above. This area would appear to be a natural for the statistical multiplexer, depending upon application.

The totals presented in the preceding table indicate the total usage of those responding to this question; i.e., responses listing multiple speeds were counted for each speed used.

For time division equipment, the subscribers were asked to identify the high-speed composite rate at which the machines were operating. The results are as follows:

High-Speed Rate, bps	Number of Responses
1200	1
2400	10
4800	22
7200	1
9600	26
19.2K	3
64K	1

The facilities that were utilized included standard voice grade lines, private wire, and some satellite channels. No optical links were identified, but we expect this in the future. The responses can be broken down as 85

Line Use Optimization with Multiplexing Equipment

percent analog and 15 percent digital (DDS). We were unable to identify the type of data set used on the private wire systems.

Longevity is another topic. The length of time that each multiplexer had been in service was answered as follows:

Time in Service	Number of Responses
Less than one year	13
One to two years	29
Two to four years	12
Over four years	5

We also asked a question concerning options being used, with the following results:

Option	Number of Responses
Autospeed	14
Echoplex	2
Outward auto-dial	1
Redundant common electronics	11
Dial back-up	10
Statistics reporting	21

The number of responses did not permit us to make a detailed comparison of options in use versus the size of communications network in use, but it appeared, as one would expect, that the use of these options tended to be in larger networks with many channel ends and terminal sites. □

Communications Processors—Special Kinds of Computer-Helping Computers

Problem:

Some of the harsher analysts of man's intellectual capabilities claim that we cannot really handle more than seven variables at the same time and that the practical limit is probably closer to three. Perhaps. It seems to be self-evident that humans tend to solve narrowly defined problems in a very linear way—computer flowcharts are a paper monument to this characteristic, and the permutational complexities of chess are what turn a patently simple game into a challenging intellectual exercise. But this failing is probably why we invented computers in the first place, just as we invented other tools to compensate for other failings.

The variables in a communications system can very quickly exceed the magic limit of seven, and so we have almost from the very beginning relegated the details of communications management to the computer. Early systems and most of today's simpler systems borrow a piece of time from the same computer that handles the payroll (e.g.) to manage the communications chores. The borrowed time is nonproductive in the sense that it subtracts from the computer's problem-solving power, but the time is well used in the context of the total system, so we tolerate it and lump it under the general heading of Unavoidable Overhead . . . . but only up to a point. Beyond that point, the nonproductive overhead time becomes too costly to tolerate, and we have to look for some way to separate the communications management chores from the computer's more important problem-solving duties. "That point" is hard to define because many factors contribute to it—the number of devices, the traffic density, the traffic speed mix, the processing mix (batch vs. interactive), and so forth, but it shows up clearly as a sharp drop-off in the processing throughput rate and/or in the terminal response time.

The solution to the problem is almost trivial—get another computer and hook it up at the "front end" of the main computer to handle all the communications management chores.

Communications Processors—Special Kinds of Computer-Helping Computers

Solution:

Separation of the data communications control function from the remainder of the processing functions is made possible by a communications processor.

When used as a front-end, a communications processor performs only the communications function for a host data processor. This arrangement has demonstrated dramatic throughput capabilities and advantageous price/performance. When used as a stand-alone system, a communications processor performs both the communications and the data processing functions. The data communications function software and hardware retain individual identities to facilitate modification. The management decision to have a distributed network or a master/slave (centralized) network is as dependent on the company's organizational structure as it is on the technical system's parameters.

Data communications has come a long way from the time when the communications, software was intertwined with the applications software and the communications hardware was integrally wired into the central processor. Two major developments and a responsible endeavor towards standardization have led to a dramatic increase in the use of communications processors with ever increasing capacity and capability.

The first major development was recognizing that the data communications functions must be segregated from other data processing functions. This resulted in modular communications software packages and communications interfaces that permitted alteration of the communications environment without major surgery to the hardware and the software. It also permitted the organization of communications processing functions, relative to other processing functions, along assembly-line principles. The assembly-line technique segments a job into discrete elements for exclusive execution by specialized persons or equipment; the assembly-line total output significantly exceeds the output of the same persons or equipment with each performing the total job. The development of specialized components to perform essential line handling functions resulted in the front-end processor, which freed the host processor of this time consuming task. A front-end/host configuration is able to handle a significantly greater data volume than an equivalently powered single processor that performs both the line handling and the data processing function.

The second major development was low-cost microprocessors that economically permitted creation of a distributed network of many small systems performing specific specialized data processing functions. This constitutes the employment of the assembly-line approach to data processing as well as to data communications control. The pros and cons of distributed data processing, as opposed to the master/slave approach with a few processors performing all data processing functions centrally, has split users into many factions. Most users are designing systems with elements of both distributed and master/slave network arrangements and balance the modular flexibility of distributed processing against a single integrated central data base that assures corporate procedural conformity and integrity. In essence, network design, as a subset of systems design, must contend with the continual struggle within an organization between the proponents of centralization and decentralization.

The cost of transmission facilities has not paralleled the rapidly declining cost of data processing, which is due to the introduction of microprocessors. This fact is frequently cited as justification for the trend towards decentralized processing. Decentralized, or distributed data processing has given rise to a new type of data processing module; the small processor, or minicomputer which performs both data and communications processing. IBM's System 8100 and Sperry Univac's V77 family of minicomputers are two examples of processors which can serve either as independent processors, or as distributed systems which offer significant communications control capabilities for network compatibility and integration.

Support for the centralized processing approach has come from some terminal suppliers who are marketing terminals with greater and greater data editing capabilities (under the misnomer of intelligent terminals). When a large enough central computer has been installed, such devices are economically attractive, compared to a distributed terminal capable of performing major data processing tasks.

VENDORS AND STANDARDS

The effort on the part of most vendors towards standardization is a little heralded revolution within data communications. This ongoing effort, along with hardware architectural improvements, is reducing the investment, inventory, and software support necessary to support a variety of different terminal and line disciplines, which are different for few justifiable reasons. Standardization, in addition to reducing costs to existing users, will continually increase the user base that can economically justify the use of electronic communications in their operations.

Communications Processors—Special Kinds of Computer-Helping Computers

The ground rules of network architectures announced by many of the large mainframe and minicomputer manufacturers have codified their communications standards. IBM's Systems Network Architecture, DEC's DECnet, Sperry Univac's Distributed Communications Architecture, and Honeywell's Distributed Systems Environment are examples of such architectures. Bit-oriented protocols are rapidly being adopted that improve the performance and error checking/recovery capabilities of data transmissions. Minor variations of the international HDLC or IBM's SDLC bit-oriented protocols are now supported by many suppliers of communications equipment.

Communications processor hardware and software architecture are continually being changed. Transistors are being replaced by multi-layered, electronically coupled, chip circuits. Throughput capabilities are enhanced by using multiple microprocessors within the communications processor to perform specialized functions. Altering the microcode or stored logic (either directly by the user or indirectly by such features as IBM's Extended Facilities) has added a new dimension to throughput improvement techniques. Multiport memory access has facilitated warm-start back-up systems. Virtual operating systems are taken for granted and full-capability data base management systems are being given serious consideration by installations previously reluctant to accept the associated CPU overhead.

While the communications processor manufacturers have been improving their hardware and software, the common and specialized carriers have not been idle. AT&T, in Chicago, is testing fibre optic cable in place of existing line cable and is currently slogging through the regulatory mire with its proposed ACS packet switching facility. IBM and Xerox, two major noncommunications vendors, are also competing directly with AT&T through their proposed SBS (IBM, et al) and XTEN (Xerox) offerings. ITT has committed itself to join Telenet, Tymnet, and Graphnet in offering a packet switching service that is characterized by charges for data transmission that are independent of distance. Certified equipment can be attached to phone lines without AT&T's, DAA protective device. Voice-grade lines can be used to transmit at up to 9600 bits per second. The changes in the costs of communications lines and in their capabilities varies with individual situations and is sufficiently complex so as to almost defy summation. There is very little doubt that the entire field of communications will be totally revolutionized during the next five years.

BASIC DEFINITIONS AND APPLICATIONS

A communications processor, in the context of this report, is simply a digital computer that has been specifically programmed to perform one or more control and or processing functions in a data communications network. As a self-contained system, it may or may not include the following components, depending on its specific application: communications lines multiplexor, line adapters, central computer system interface, and on-line peripheral devices. It always includes a specific set of user-modifiable software.

Communications processors are not a new system design concept. During the industry's second generation, in the early 1960's, such processors were offered by several of the major main-frame suppliers, including Control Data's 8090/8050, General Electric's DATANET-30, and IBM's 7740. Also, Collins Radio Company (now Collins Communications Switching, Rockwell International) delivered its first Collins Data Central programmable communications system way back in 1963. In almost all such early uses, the systems were used primarily in message switching applications, acting simply as message routers and dispatchers in a data communications network.

The principal differences today lie in the diversity of application areas, the relatively low cost of the units, and, by consequence, the trend toward widespread usage. Listed below are some of the principal uses of programmable communications processors in current data processing systems. It is important to note that many such units can be used in a variety of application areas, with specific sets of software and interface units for each application. The current popular types of applications include:

- **Message switching.** The message switching processor receives messages from remote terminals, analyzes them to determine their proper destination, performs any code conversions that may be necessary, and transmits them to other remote terminals. The sending and, or receiving remote terminals may themselves be computer systems. Most message switching systems are of the store-and-forward type, in which the programmable processor stores the messages it receives on on-line auxiliary storage units, such as disks, drums, or magnetic tape. The length of time the messages are stored before transmission to other terminals or computers can range from a few seconds to an entire day or more, depending on the specific application needs and traffic volumes. The programmable processor performs little if any processing on the messages; it acts principally as a traffic director.

- **Front-end processing.** The programmable communications processor replaces a hard-wired communications controller as the interface between the central data processing system and the data communications network. The front-end processor

Communications Processors—Special Kinds of Computer-Helping Computers

not only receives and transmits all data passing through the network, but also, and significantly, can be programmed to pre- and post-process this data in a variety of ways in order to relieve the system's central processing unit from time-consuming overhead activities related to message formatting and control. Front-end processors can perform their functions in support of a wide variety of data processing applications. Additionally, the more sophisticated communications processors can be employed with software which permits them to be automatically reconfigured from a front end mode of operation, to that of a remote communications processor. This feature permits a single front end to automatically switch to a backup host in the event of a primary host failure, and also to perform communications processing for both local and remote hosts simultaneously.

- **Line concentration.** Programmable communications processors sometimes fill the relatively simple role of communications line concentrators. Here the processor generally terminates a number of low-speed transmission lines and interfaces them to one or two higher-speed lines for more efficient and economical data transmission. Little, if any, processing of the transmitted data is performed. The programmable aspect of the processors is probably less used in this application than in any of the other currently popular uses. Hard-wired concentrators are generally equally effective, suffering by comparison only in their lack of flexiblity. Microprocessors may change the significance and flexibility of this class of applications equipment.

- **Dedicated processing.** Many of the programmable communication processors now have enough storage capacity and processing power to enable them to serve as the sole or principal computers in dedicated application systems of various types. In inquiry/response systems, for example, the processor receives inquiry messages from remote and or locally connected terminals, processes the messages to determine the specific information required, retrieves the information from on-line random-access storage units, and sends it back to the inquiring terminals. In systems of this type, application-oriented processing is as important as message receipt and transmission.

WHY PROGRAMMABLE COMMUNICATIONS PROCESSORS?

Programmable communications processors are enjoying increased popularity in various parts of data communications systems because they are demonstrably more effective on a price performance basis than their predecessor hard-wired controllers. General advantages that contribute to this price performance edge include the following:

1. **Price.** Through the economies afforded by integrated circuitry, today's general-purpose minicomputers can often be purchased for less money than specialized hard-wired controllers. Even when the cost of specific software routines is added to the cost of the minicomputer to adapt it for specific data communications functions, the net price of the mini-based controller will often be substantially less than the hard-wired equivalent.

2. **Flexibility.** Since the programmable processors can be modified at any time and at comparatively low cost to user or vendor, they are eminently well suited for key roles in data communication systems, which are typically characterized by bewildering variety and constant change. Advances in communication line facilities are being made by the common carriers and by the independent companies, making available new, faster, and lower-cost transmission services. Remote terminals have long since surpassed the simple typewriter-style devices and now abound in a variety of shapes, styles, formats, functions, and transmission characteristics. Other dynamics at work include the tendency of most data communication networks to expand as workloads grow and applications increase. Programmable communications processors can readily serve such a continually changing environment by permitting efficient systems growth and guarding against obsolescence.

3. **Distribution of labor.** Since these processors can be programmed to perform varying amounts of productive processing, often in conjunction with their own on-line peripheral devices, they can share portions of the overall processing load with other processors in the system. Peak loads can be more effectively handled and critical bottlenecks more likely avoided.

4. **Fail-soft capability.** In data communications systems that include at least one other computer, programmable communications processors can provide some form of continued system operation when one or more of the other computers fail. The degree and effectiveness of this fail-soft capability depend not only on the capabilities of the programmable processor but also, perhaps more importantly, on the skill displayed by the system architect in his provisions for redundant components and fall-back procedures.

5. **Independent processing.** When programmable communications processors are not involved in their principal data communications tasks, they can often be used as stand-alone data processing systems provided, of course, that their configuration includes some input/output devices. Simple media conversion tasks, such as card-to-tape and tape-to-print, can be valuable by-products from these otherwise com-

munications-oriented processors, and localized time-sharing can yield added benefits.

FRONT-END PROCESSORS

The most important use of programmable communications processors today is front-end processing, in which the processor replaces a central computer system's hard-wired communications line controller, handles all message control activities, and performs enough preprocessing of the transmitted data to relieve the central processing unit of the communications housekeeping activities that otherwise threaten to choke its productivity. Front-end processors can be rated as most important because their usage is potentially the most widespread and because their benefits to the overall system are most valuable.

The concept of front-end processing essentially involves off-loading or removing the data communications control function from the central processing unit and setting it up as an external, largely self-contained system. This decentralized approach to the distribution of processing labor permits both the communications and central processors to perform their primary functions in parallel and with little interference. Data is passed between the processors only when necessary and with as high a degree of efficiency as is possible in circuit design.

A typical front-end processor might control a hundred or more communications lines of varying speeds and types attached to a large number of diverse remote terminals. The front-end processor would ideally assume all terminal, line, buffering, and message control functions, permitting the central processing unit and the user application programs to treat the communications network as just another high-speed, on-line peripheral device.

The concept of front-end processing is not new. General Electric and IBM offered users of their second-generation computer systems the availability of programmable front-end processors (the GE DATANET-30 and the IBM 7740 Communication Control System). But the concept has begun receiving widespread attention only in recent years, as processor and on-line storage costs have dropped and data communications software know-how has advanced. Enthusiastic promotion of front-end processing by minicomputer manufacturers and independent systems houses has in turn drawn the major main-frame suppliers into the market. This formidable marketing assemblage ensures that a typical centralized data processing installation using or contemplating a data communications subsystem will be strongly urged, perhaps for the first time, to consider installation of a front-end processor.

Front-End Components

The essential components of every front-end processing system are the following:

1. **Processor.** The processor element can be a stored-program digital computer of almost any size. It must have its own main memory, but it may or may not use on-line peripheral devices. The processors should have excellent interrupt handling and strong bit manipulation capabilities.

2. **Central processor interface.** The front-end processor component must include the proper hardware interface to permit it to connect directly to a standard input/output channel of the central processing unit (or host computer). The interface must permit the host computer to communicate with the front-end processor as if it were a standard peripheral device control unit, requiring little if any operating system software modification.

3. **Communications multiplexer.** This component provides a logically independent data channel into the front-end processor's main memory for every transmission line being served. The multiplexor serves as the front-end processor's functional interface to the data transmission lines. Control of incoming and outgoing data is coordinated between the multiplexor and the processor via interrupts.

4. **Line interface units.** These components are hard-wired devices that link the multiplexor with the modems that terminate each communications line. Like the modems, the line interface units are specifically tailored to serve the speed transmission characteristics of the lines they terminate. The lines are, in turn, generally selected according to the transmission requirements of the remote terminal devices.

5. **Software.** The front-end processing hardware components become an integrated, functioning system only through the inclusion of software—some generalized, and some high specialized. The software programs should include terminal control, line control, message control, and central system interface procedures. Depending on the supplier, the user may be asked to write some portion of this software.

Front-End Functions

Because a front-end processor is essentially a programmable computer, it can be programmed to perform an almost limitless variety of functions. But in its role as external controller of a centralized data communications network, the specific functions generally programmed are those that relate to data and message control. The following functions are the most important ones offered with the more comprehensive

Communications Processors—Special Kinds of Computer-Helping Computers

front-end processing systems. Some systems will not provide all these functions, as all are not required in specific installations.

1. **Line Control.** This involves the periodic polling of terminals to determine readiness to transmit and receive data. Automatic call answering, acknowledgement, and dial-up can also be handled.

2. **Character and message assembly.** Bits are assembled (and disassembled) into parallel characters, and control characters are recognized to permit the assembly and disassembly of entire messages. The host computer should receive only completed messages for processing. Data can be handled at varying line speeds and in synchronous or asynchronous formats; start-stop bits and synchronizing characters are handled automatically.

3. **Data conversion.** The data transmission codes (such as Baudot, ASCII, etc.) are converted into a common structure that is equivalent to the native data code (such as EBCDIC) of the host computer.

4. **Data and message editing.** This is a general function that can include application-oriented reformatting, removal of spaces and zeros (and other kinds of data compression), and other data restructuring to permit more efficient data transmission on the one hand and more efficient processing by the host computer on the other.

5. **Error control.** Using both hardware and software techniques, the front-end processor can detect and correct data transmission errors before they reach the host computer. As a result, the host computer can rightly assume that all messages it receives contain pre-validated data ready for processing.

6. **Message buffering and queuing.** The front-end processor can buffer several messages in its main memory before passing them to the host computer, with the intention of interrupting the host as infrequently as possible. Also, if the host computer cannot process incoming messages as fast as they arrive, the front-end processor can queue the messages in its own auxiliary storage units, and can transfer the messages to the host computer when processing time becomes available. Queue management can be arranged in several different ways, including a system of priorities.

7. **Message switching.** When the front-end processor serves more than one host computer, it will analyze message headers and addresses and send each incoming message to the proper central computer. This situation can occur when several separated, dedicated central computers share a data communications network.

8. **Message answering.** Certain messages, such as simple inquiries, can be completely processed by the front-end processor without any contact with the central data processing system. Since many front-end processors permit attachment of on-line auxiliary storage units, these processors can store and access their own private data bases. Some systems also permit the front-end processors to directly access the auxiliary storage subsystems and data files of the host computer.

9. **Message recording.** Vital inbound messages can be passed on to the host computer while being simultaneously recorded in the front-end processor's auxiliary storage. Such message recording can assist in system restart operations in case the central system should malfunction and lose either its messages or the results of processing the messages. Also, it may be advisable in some systems to store a journal record of every message received during each processing period.

10. **Statistics recording.** The front-end processor can keep a running record of all data communications traffic, including statistics such as total number of messages processed, number of messages delivered to each destination, number of line errors, average length of time in queue, number of busy signals, etc. These statistics can be dumped on demand or in the form of reports at the end of each processing cycle.

Other application-oriented functions can be progammed by the front-end supplier, by the user, or by some combination of the two. It must be remembered, however, that the front-end processor, like the host computer, has only a finite amount of processing power. The more functions that are added to it, the more likely it is to run out of power, especially in active, growing communications networks. A front-end processor pushed beyond its capacity will result in lost messages and, ultimately, in system failure.

Advantages and Disadvantages of Front-End Processing

The possible front-end processing functions noted above can be translated into specific advantages over the more conventional hard-wired communications controllers. These advantages are not unlike the general advantages of programmable communication processors stated earlier in this report. The advantages include:

- **Flexibility**—handling many line speeds and transmission characteristics in uniform or interchangeable circuitry.

- **Adaptability**—supporting a wide variety of remote terminals from the main-frame and independent

Communications Processors—Special Kinds of Computer-Helping Computers

suppliers, regardless of their transmission speeds, line control conventions, synchronization techniques, and data codes.

- **Expandability**—permitting relatively easy growth of the data communications network, principally by adding line interface units and modifying the control programs.

- **Performance**—handling more and higher-speed data communications lines than hard-wired counterparts, with less host intervention or overhead.

- **CPU relief**—controlling the entire data communications subsystem will relieve the system's central processing unit on two counts: processing time and main memory space. Central control of data communications networks can consume 40 to 50 percent of the available processing time in typical situations. And the resident software control routines can easily consume in excess of 50K bytes and frequently use up to 300K bytes or more of main memory space, depending on the functions performed. Efficient utilization of front-end processors can provide almost full relief in both processing time and memory space overheads. (If the host processor is not overburdened, the need for a programmable unit may be harder to justify.)

- **Economy**—a specific front-end processor system can deliver many years of service depending on its flexibility, adaptability, and expandability. Important economies will also derive from freeing the host computer to turn out 40 to 50 percent more productive work during each processing period. As a corollary, the host central processing unit will itself tend to have a much longer useful life span before requiring an upgrade to a larger, more expensive model. Also, one front-end processor can often serve as many lines as two or more hard-wired controllers, and one such procesor can also serve mutliple host computers. There is even the possibility of replacing a large host computer with a smaller model after the communications load has been shifted to a front-end processor.

- **Fail-soft**—adding a front-end processor to a central data processing system can, if so designed, enable portions of the communications network to keep on operating—although in degraded mode—when the central processing unit malfunctions.

- **Reliability**—utilizing a monitor unit, a system operator can interrogate a front-end processor at any time for information on the operational status of the data communications network. With these diagnostics, component failures can be readily identified and corrected.

- **High-level user interface**—permitting the user programs to address the data communications network as a standard peripheral device. The complexities of the network can remain transparent to the user.

- **Independent processing**—permitting the front-end processor to answer inquiries from its own data banks in on-line mode, and also to perform simple, independent tasks in off-line mode, such as card-to-tape, tape-to-print, etc. In off-line mode, the processor can also be adapted to serve specialized I/O devices, such as plotters and OCR devices, that the central system may not be able to handle.

Now for Some of the Disadvantages

Front-end processors deserve careful investigation because of their many apparent advantages over hard-wired communications controllers. Such investigations should include as many probing questions as possible, because there are potentially serious pitfalls to be avoided.

One potential problem is the question of overloading the front-end processor, with the resultant loss of data. Sophisticated data and message control programs will consume large quantities of the front-end processor's computing and memory facilities, just as they do in a centrally-based communications system. Since many front-end processors are based on minicomputers, the possibility of overloading is all the more real. A tendency toward overloading can easily negate any apparent advantages of expandability and growth potential.

Another serious question is that of software. The body of software required for terminal control, line control, and message control activities, not to mention application-oriented pre-processing, is unquestionably complex. It is also vital to the operation of these systems. The prospective user must determine whether or not the supplier is capable of supplying this software, at what level of completeness, with what assurance of bugfree stability, with what chances of interfacing smoothly with the central system software, and with how much installation assistance.

Obviously, if the software doesn't work properly, the system is of little value. From another point of view, a system whose software works but performs very few and very basic functions may still offer little more than a typical hard-wired controller.

Another consideration is that the hardware software combination that makes up a front-end processor may require far more time and effort to install and make operational than a hard-wired controller, especially when the supplier of the front-end equipment is dif-

Communications Processors—Special Kinds of Computer-Helping Computers

ferent from that of the host computer system. Apart from the traditional problems (real or imagined) of divided vendor responsibility, there exists the very real problem of integrating two completely different sets of hardware and software.

A currently operational data communications installation considering replacing its hard-wired communications controller(s) with a front-end processor must carefully evaluate the problems of conversion. Beyond the usual problems of data integrity and the logistics of arranging the conversion process, the user may also be faced with the prospect of modifying either his central system control software or his body of application programs that use the communications network.

Evaluating a front-end processing system on a cost value basis is extremely complex and can be almost meaningless when performed in the abstract. Costs will vary with the size and diversity of the network being controlled, with the size and processing power of the front-end processor, with the number of control and preprocessing functions incorporated (software is expensive, whether hidden in a "bundled" system price or not), and with the number of on-line peripheral devices. Keeping costs to an absolute minimum will probably result in a system that is capable of little more than the hard-wired controller it is replacing. In this case, the cost differential is easily measured, but it will not likely be significant in either direction.

Adding functions that will permit use of "foreign" terminals, relieve the central processor of intolerable overheads, and allow independent and back-up processing may increase the costs as it increases the value. In order to evaluate the reasonableness of the cost of the front-end processor and the potential cost savings throughout the system, an effort must be made to associate specific dollar figures with the expected values to be derived from re-orienting a host-controlled data communications system to an externally controlled one. In summary, it should be clear that costs and values of front-end processing can be assessed only in terms of specific situations and specific systems.

BUYING GUIDANCE

The front-end processing products have not matured to the point where their descriptive terminology is in any way standardized or consistent. As a result, the prospective buyer must make every effort to determine exactly what he will be getting and what he will not. The sales brochures and technical manuals are often not sufficiently informative (and sometimes downright misleading). We have explained some of the most important features under the next heading.

First of all, there are at present two distinctly different kinds of front-end processors. The first and more basic variety is designed to simply replace the functions and services of the central system's hard-wired controller. It is meant to be a plug-compatible replacement, requiring few, if any, changes to the central system's communications control software or the user's application programs. It does not necessarily relieve the central system of any software control overheads, but simply provides a more flexible interface to the communications network for accommodation of additional and varied lines and terminals in the future.

The most prevalent examples of this type of front-end processor are the many available units designed to replace or emulate the IBM 2701 Data Adapter Unit and the IBM 2702 and 2703 Transmission Control Units. These front-end processors function with the IBM System 360 or System 370 computer systems through the standard IBM BTAM, QTAM, and TCAM communications control software.

The second and more powerful variety of front-end processor is designed to replace not only the functions and services of the hard-wired controller, but also most or all of the data communications control functions normally performed by the central system's processing unit and resident software. This variety of front-end processor, by freeing the central processing unit for productive work, provides valuable advantages not only in data communications flexibility but also in systems throughput.

It is possible that a user may want to install the basic kind of front-end processor initially and then gradually add functions to it to relieve the central processing unit's communications overheads. However, the user must make sure that his selected front-end processor has enough processing and memory capacity to permit the gradual build-up of substantial message control routines, and that the various responsibilities of both the vendor and the user are clearly assigned.

Another buyer's tip is to look for the word turnkey. Turnkey installation of front-end processors usually means that the supplier takes on full responsibility for hardware, software, and interfaces required to essentially plug in his product. From a user's point of view, this approach is highly desirable, since it can save him money, time, and aggravation. But the user must still determine what product with what promised functions is being offered on the turnkey basis. It may still be a somewhat limited front-end product.

A low list price can be totally misleading, since it may include only a minicomputer and an associated communications multiplexor. The cost and effort of establishing the proper interfaces and writing the all-important software can be dropped squarely on the buyer, who may have been trapped by an attractive low-price bid.

Since software development is such a critical question, the buyer should determine early in the proceedings exactly what software is provided with the basic front-end system and at the basic price. If certain software is lacking, such as specific remote terminal handlers, or message queuing routines, then implementation and integration responsibilities should be clearly fixed, and with firm price quotations.

The buyer will also ask the competing bidders for clear statements of service and support after installation of the front-end processor. Since data communications subsystems can be complex and demanding in any environment, it must be considered an extremely valuable system feature if the prospective supplier of the front-end processor offers to assume full operating and service responsiblity for the externally controlled communications network that is directed by his product.

When considering a front-end processor from a source other than the supplier of the central computer equipment, the buyer should insist on receiving concrete performance data, drawn from installed systems, to substantiate the supplier's claims. The buyer should beware if the supplier refuses to back up his claims with actual case studies. As further evidence of proven performance, the buyer should personally contact as many previous users as possible, probing not only for their degree of satisfaction, but also for the extent to which the installed systems reflect his own intended system design and functional objectives. However, even in highly specialized reference accounts, meaningful information can be derived regarding the supplier's competence and willingness to help, and the basic reliability of the hardware software package.

When the proposed supplier is a major main-frame manufacturer, the buyer will also want evidence of proven performance. This evidence should apply to the overall performance of the total, integrated data processing system, and not just the front-end subsystem. However, when the main-frame supplier offers a choice of a front-end processor or a hard-wired controller (as several now do), then the buyer will again want specific, tangible performance data to justify selection of front-end processing. Of course, the main-frame supplier can forcibly persuade adoption of the front-end concept, even without offering convincing performance data, by simply indicating that the newer product will receive all future support and that the former one will be essentially dropped from the product line.

HOW TO ASSESS COMMUNICATIONS PROCESSORS FEATURES AND CHARACTERISTICS

With one exception, all the non-economic characteristics of communications processors reduce themselves to one consideration: the throughput capabilities of the equipment relative to the specific systems requirements. The exception is where the physical attachment limitations are exceeded before the processing capabilities are fully used.

For example, the number of high speed communications lines that are physically attachable to a processor usually exceeds the throughput capabilities. For that reason, most vendors specify a smaller value for the number of lines attachable at the higher speeds than the equipment can physically accommodate. The numbers more accurately describe the outer limits of a processor's throughput limitations rather than its physical limitations. Vendors generally want you to realize that the line mix and the resource mix could radically alter the number of lines that could be supported, physical port availability notwithstanding.

Network Arrangements Supported

Most processors operating as a front-end are restricted to supporting the host computer systems of specific mainframe manufacturers. However, some vendors include in their product lines front ends that can be customized. Also, many older mainframes that have been fully written-off from an accounting standpoint can be offered at low enough prices to justify tailoring and dedicating the overqualified equipment to function as a front-end.

From a network arrangement standpoint, the number of direct connections a front-end can support to one host and the number of hosts a front-end can support become important considerations especially for fall-back considerations. Usually, a small number represents a special, direct connection. A high number indicates that the connection is via a regular communications line port and does not mean that the vendor is suggesting that so many connections to one or more host is designed capability.

When only one station can be polled on one line, the system, as standard, supports only point-to-point terminal arrangements. When the communications processor functions as a remote concentrator, the number of host concentrator connections is also a consideration from a network standpoint. Again, the number of connections permitted is primarily an indication of whether a special interface or a regular communications line interface is used.

Communications Processors—Special Kinds of Computer-Helping Computers

Processor/Controller	Number of User Responses	Number of Units Installed	Average No. of Lines per Unit	Average No. of Terminals per Line	User Ratings*														
					Overall Satisfaction					Ease of Installation					Throughput				
					WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
Hard-Wired Controllers																			
Burroughs X350 Series	7	12	13.3	3.8	3.7	5	2	0	0	3.7	5	2	0	0	3.4	3	4	0	0
IBM 2701	21	22	1.3	4.3	3.4	9	8	2	0	3.2	8	7	3	1	3.2	6	10	3	0
IBM 2702/3	3	3	29.3	2.8	3.3	1	2	0	0	3.0	0	2	0	0	3.5	1	1	0	0
IBM ICA	4	6	4.0	6.6	3.3	1	3	0	0	3.0	2	1	0	1	2.5	0	2	2	0
Memorex 1270	40	65	28.0	2.1	3.6	25	12	2	0	3.3	18	15	4	1	3.5	19	18	0	0
Sperry Univac CTMC	4	3	21.3	2.1	3.5	2	2	0	0	3.3	2	1	1	0	3.5	2	2	0	0
Others	12	18	13.3	2.7	3.3	5	6	1	0	3.3	5	5	2	0	3.2	4	6	2	0
TOTALS	91	129	17.5	2.4	3.5	48	35	5	0	3.3	40	33	10	3	3.3	35	43	7	0
Front-End Processors																			
Burroughs DCP	3	3	101.3	6.5	3.7	2	1	0	0	3.7	2	1	0	0	4.0	3	0	0	0
Burroughs, Others	3	6	7.0	9.7	3.3	1	2	0	0	3.0	1	1	1	0	3.0	0	3	0	0
Comten 3650	15	25	42.9	2.1	3.6	9	4	1	0	3.1	7	4	1	2	3.6	9	5	0	0
Comten 3670	15	27	115.2	3.1	3.5	8	7	0	0	3.3	5	10	0	0	3.6	9	4	1	0
CDC 17/1700	4	4	36.3	9.1	3.3	1	3	0	0	3.3	1	3	0	0	3.3	1	3	0	0
CDC 2550	2	2	83.0	—	3.0	0	2	0	0	2.5	0	1	1	0	3.5	1	1	0	0
DEC PDP-11	3	39	34.3	3.9	2.7	0	2	1	0	2.3	0	1	2	0	3.0	0	2	0	0
Harris 4705	3	5	9.0	6.2	2.3	0	2	0	1	2.3	0	1	2	0	2.7	0	2	1	0
Honeywell DN 355	3	3	35.3	1.8	4.0	3	0	0	0	3.3	1	2	0	0	4.0	3	0	0	0
Honeywell DN 30	4	5	8.0	3.6	2.8	0	3	1	0	2.5	0	2	2	0	1.8	0	0	3	1
Honeywell DN 2000	5	6	7.6	2.3	2.6	0	3	2	0	2.4	0	2	3	0	2.6	1	2	1	1
Honeywell 66XX	4	5	27.5	1.9	3.8	3	1	0	0	3.3	1	3	0	0	3.5	2	2	0	0
IBM 3704	71	78	7.4	4.9	3.4	32	31	3	0	3.1	21	37	8	3	3.2	18	43	6	0
IBM 3705	81	110	42.9	2.8	3.5	39	34	3	0	3.2	31	32	9	5	3.2	25	42	7	1
Periphonics T Comm 7	4	5	25.0	6.4	2.8	0	3	1	0	2.5	0	2	2	0	2.8	0	3	1	0
Others	15	22	21.2	2.7	2.9	5	4	2	3	2.5	3	4	6	2	2.9	7	2	4	2
TOTALS	235	345	32.9	3.0	3.4	103	102	14	4	3.1	73	106	37	12	3.2	79	114	24	5

*User ratings report the number of users responding Excellent (E), Good (G), Fair (F), and Poor (P) for each category. The weighted averages (WA) were calculated by weighting the four ratings on a 4, 3, 2, 1 basis.

Users' ratings of communications processors/controllers

Since the prime purpose in burdening communications lines around the world with data is to either retreive information or to add to the store of information, the nature of the data base system supported should not be overlooked. Actually, it represents the "end" for which one selects a "(communications processor) means." Of course, a buyer may be already committed to a file maintenance or data base system and not be interested in this type of support.

Properly depicting communications line capacity is the most difficult and the most controversial vendor specification. As a reasonable alternative. Datapro evaluates the number of half-duplex lines that can be physically attached to the processor presuming all lines were operating within a given speed range. The ranges chosen were: up to 1800 bps, 2000 to 9600 bps, and over 9600 bps. The number of low speed lines usually represents the physical and throughput limitation for asynchronous lines. Generally, the medium and high speed lines represent the outer limits of the throughput capabilities.

The terminal protocols supported by the processors are important features. Even though the protocols supported mostly depend on the marketing philosophy of the vendors, a large number of vendors are beginning to support the standardized bit-oriented protocols.

Processor Characteristics

If the processor is microprogrammable by the user, one can expect the capability for increasing throughput by properly microcoding frequently-used, time-critical functions. If not properly done, the capability could adversely effect the installation. Main memory cycle time, main memory word size, and main memory storage capacity offer a very general feel for throughput speed possibilities. However, sophisticated internal architecture may enable the processor to be many times faster than another processor with the same cycle time and word size. That is another reason why we emphasize detailed analysis after initial selection.

The manner of data transfer between memory and communications lines, memory and mass storage, and memory and othe supported peripherals becomes critical as volume requirements rise and/or response

Communications Processors—Special Kinds of Computer-Helping Computers

Processor/Controller	User Ratings*																								
	Hardware Reliability					Promptness of Maintenance					Quality of Maintenance					Software					Technical Support				
	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
Hard-Wired Controllers																									
Burroughs X350 Series	3.9	6	1	0	0	3.6	4	3	0	0	3.6	4	3	0	0	3.8	5	1	0	0	2.3	1	3	1	1
IBM 2701	3.7	13	6	0	0	3.5	10	8	1	0	3.3	9	8	1	1	2.7	2	9	2	2	2.8	3	11	4	1
IBM 2702/3	3.7	2	1	0	0	3.0	1	0	1	0	3.0	1	0	1	0	2.5	0	1	1	0	2.5	0	1	1	0
IBM ICA	3.8	3	1	0	0	3.3	1	3	0	0	3.0	1	2	1	0	2.0	0	0	2	0	2.3	0	1	2	0
Memorex 1270	3.7	27	10	0	0	3.2	15	14	5	2	3.1	13	14	8	1	3.0	1	3	1	0	2.9	7	16	7	1
Sperry Univac CTMC	3.3	2	1	1	0	3.5	2	2	0	0	3.5	2	2	0	0	3.0	0	1	0	0	1.7	0	1	0	2
Others	3.6	8	3	1	0	3.5	7	4	1	0	3.3	20	15	4	0	3.1	3	5	2	0	2.8	2	5	4	0
TOTALS	3.7	61	23	2	0	3.3	40	34	8	2	3.3	50	44	15	2	3.0	11	20	8	2	2.8	13	38	19	5
Front-End Processors																									
Burroughs DCP	3.7	2	1	0	0	2.3	0	1	2	0	2.7	0	2	1	0	4.0	3	0	0	0	2.7	1	1	0	1
Burroughs, Others	3.3	1	2	0	0	2.7	1	0	2	0	2.7	1	0	2	0	3.0	0	3	0	0	2.7	1	1	0	1
Comten 3650	3.7	11	2	1	0	3.4	7	6	1	0	3.4	8	4	2	0	3.4	7	6	1	0	3.2	7	4	2	1
Comten 3670	3.5	9	5	1	0	3.3	6	8	1	0	3.3	6	8	1	0	3.1	5	7	3	0	2.9	4	5	4	1
CDC 17/1700	3.5	2	2	0	0	3.5	2	2	0	0	3.3	2	1	1	0	2.3	0	2	0	1	2.7	0	2	1	0
CDC 2550	2.5	0	1	1	0	4.0	2	0	0	0	3.5	1	1	0	0	3.0	1	0	1	0	3.0	0	1	0	0
DEC PDP-11	2.3	0	2	0	1	2.0	0	1	0	1	2.0	0	1	0	1	2.0	0	0	1	0	2.0	0	0	1	0
Harris 4705	2.7	1	1	0	1	2.3	0	2	0	1	2.0	0	1	1	1	2.0	0	1	1	1	2.0	0	0	3	0
Honeywell DN 355	3.7	2	1	0	0	3.7	2	1	0	0	3.7	2	1	0	0	3.3	1	2	0	0	3.0	1	1	1	0
Honeywell DN 30	2.5	1	1	1	1	3.0	0	1	0	0	2.3	0	1	2	0	1.5	0	0	1	1	1.5	0	0	1	1
Honeywell DN 2000	3.2	2	2	1	0	2.8	0	4	1	0	2.8	0	4	1	0	2.2	0	2	2	1	2.2	0	2	2	1
Honeywell 66XX	3.8	3	1	0	0	3.5	2	2	0	0	3.0	1	2	1	0	3.0	1	2	1	0	2.5	1	1	1	1
IBM 3704	3.6	44	22	2	1	3.3	28	35	7	0	3.3	27	36	6	1	2.8	13	35	19	3	2.8	14	34	18	4
IBM 3705	3.7	56	19	1	0	3.5	44	28	4	0	3.4	37	34	4	1	3.0	19	35	14	4	2.9	22	32	18	4
Periphonics T Comm 7	2.8	1	1	2	0	2.5	1	0	3	0	2.5	0	2	2	0	2.0	0	1	2	1	2.1	0	3	2	2
Others	2.6	5	3	3	4	2.7	2	7	6	0	2.8	4	6	3	2	2.3	4	3	2	6	2.2	3	3	3	6
TOTALS	3.5	140	66	13	8	3.3	97	98	27	2	3.2	89	104	27	6	2.9	54	99	48	18	2.8	54	90	57	23

*User ratings report the number of users responding Excellent (E), Good (G), Fair (F), and Poor (P) for each category. The weighted averages (WA) were calculated by weighting the four ratings ona 4, 3, 2, 1 basis.

Users' ratings of communications processors/controllers (continued)

times are reduced. For high speed, high volume transmissions, Direct Memory Access transfers instead of character interrupt transfers become mandatory for reasonable throughput rates.

USERS' REACTIONS AND RATINGS

In the November 1978 supplement to both DATAPRO 70 and DATAPRO REPORTS ON DATA COMMUNICATIONS, we published a Reader Survey Form on Communications Controllers/Processors. Although the subject of this report is communications processors, considerable feedback was received on popular hard-wired controllers, and summaries of these products are also included to provide a frame of reference.

By our editorial cut-off date of January 15, 1979, we had received a total of 292 usable replies representing 430 controllers and processors. The number of processor responses, and the number of processors represented was almost exactly the same number as the previous survey conducted in 1977. However, the number of hard-wired controller responses decreased by almost a third.

The 350 communications processors represented were controlling some 17,000 lines connecting about 65,000 remote terminals. It is interesting to note that compared to the 1977 survey, the average number of communications lines controlled by each processor increased from 32.9 to 48.4, and the average number of terminals per line increased slightly from 3.0 to 3.8. While a one-for-one comparison between responses to this and the previous survey could not be accomplished, the figures indicate that many of the users' networks have grown considerably since 1977.

The content of these responses is tabulated in the accompanying table. The table is arranged into the two basic equipment groups; hard-wired controllers and front-end processors.

In an attempt to identify how communications processors were being used, we asked the users to

Communications Processors—Special Kinds of Computer-Helping Computers

check one or more usages in a list of five: front-end, 270X emulation, remote concentrator, message switching, and other. The purpose was to determine the level of sophistication among users in the use of communications processors. The results are summarized below, but be sure to read the notes following the presentation.

Processor usage	Percent of Responses
Front-end	37%
270X emulation	57
Remote concentrator	6
Message switching	3
Other	3

Because the percentages total over 100 percent, it is obvious that some users reported more than one usage. In a few cases, it appeared that multiple units were being used in different fashions. For example, a user of two processors might have indicated that one was employed as a front end, while the other performed as a remote concentrator.

Two distinct patterns of usage were indicated by the users of the IBM 3704 and 3705. Some 88 percent of the 3704 users indicated that usage was exclusively 270X emulation, and only 10 percent said that their 3704's were operating in a partitioned mode; performing both network control and 270X emulation functions. One response indicated that a 3704 was operating exclusively as a network-controlling front end, but the operating software indicated that it was, in fact, performing only 270X emulation.

In comparison, users of 3705's indicated a much higher usage as front ends; a dramatic increase over the 1977 survey. Some 40 percent of the 3705 respondees indicated that their units were performing either as network-controlling front ends, or both as front ends and 270X emulators. This represents an 800 percent increase over the 1977 survey, wherein only four percent indicated that their 3705's were performing as front ends. Of the balance of 3705 responses, 56 percent indicated that usage was only in 270X emulation; two percent were being used as remote concentrators; and the remaining two percent gave an unclear indication of whether usage was as a front end, 270X emulator, or both. The percentage of Comten users indicating front end usage was 37 percent, just slightly below that of the 3705 responses. All other responses, taken as a group, which accounted for 24 percent of the total responses, reported a total of 52 percent front end usage.

While it is apparent that many users are still not making use of the full power of front-end processors, it would appear that this percentage is dwindling. It could not be determined from this survey what impact, if any, the increased marketing of distributed systems such as the IBM 8100 or System/38 has on the network control or communications processing, but Datapro believes that the increased integration of such systems can only expedite increased network and communications sophistication and efficiency.

The survey provided for identifying areas of major difficulty. Based on the overall number of responses (292), the users' assessments are summarized below:

Major difficulty	Percent of Responses Reporting
Communications processor software	22%*
Host system software	13
Throughput	4
Communications lines	18
Modems	11
Terminals	10
System Expansion (installing more lines)	14

*Based on 235 processor responses.

In only two areas were there significant deviations between controller users and processor users; throughput and expansion. The hard-wired controller users cited throughput as a major difficulty more than twice as often as the processor users, and 19 percent of the controller users indicated system expansion was a major problem, compared to 12 percent of the processor users. In postulating a possible explanation, we should first note that the controller users stated that the average number of lines per controller was 29.2, which is almost double the 17.5 average calculated from the 1977 survey. While noting again that a one-for-one comparison between responses was not made, it might be concluded that many controller users have approached or reached the maximum capacity of their controllers, hence accounting for the increased comments concerning throughput and expansion.□

Private Telephone Systems—PBX and PABX

Problem:

The telephone segment of a communications system all too frequently tends to get lost in the general excitement and volatility of data communications equipment and techniques, but an internal, private telephone system that may or may not interface with the public utility is probably the most sensitive part of a communications systems because it is used by everyone in the company and thus must be a practically foolproof system. Since the 1968 Carterfone decision, you now have almost complete freedom to build or buy non-AT&T equipment and to hook the equipment into AT&T's telephone system (check carefully before you do this though; there are some hook-up restrictions).

This report is presented as a refresher for the data communications manager who wants to re-examine the basic technology and nomenclature of PBX and PABX equipment. It also offers a succinct and useful summary of the interconnect options now available to the general user of AT&T facilities.

Solution:

When the term "private telephone system" is used in this report, it refers to telephone switching equipment dedicated to a specific user. The telephone companies refer to a telephone switching center as a telephone "exchange." A public telephone exchange is available to a great number of residential and business users. The telephones we have in our homes access the public telephone exchange. But when a business has a number of telephones, it is desirable to remove them from the public exchange and provide that business with its own private telephone exchange. This consists of telephone switching equipment located at the business site and dedicated to that business' use.

The original terminology for such a set-up was a "Private Branch Exchange" or PBX. The term PABX originally meant a private automatic branch exchange. Such systems permitted the users to automatically dial other users on the same exchange without requiring the involvement of the switchboard attendant or operator. PABX then came to imply a different kind of distinction. The term PBX was generally used to refer to a private telephone system provided by the local telephone company; and PABX usually implied a customer-owned private telephone system or an interconnection telephone system not provided by the telephone company. Now the two terms are used interchangeably.

In the broadcast sense, private telephone systems regardless of their source, can be divided into two classifications:

- Key Systems and
- PBX/PABX Systems.

A Key System does not have any switching capability. All associated telephone sets must "pick-up" all external telephone lines that access the local public tele-

Private Telephone Systems—PBX and PABX

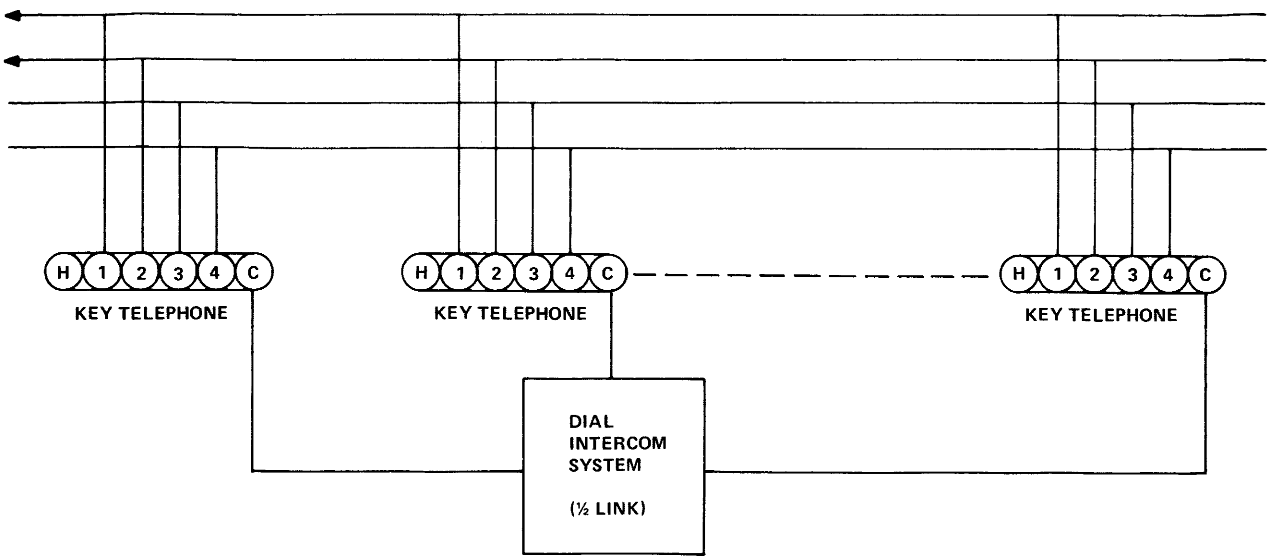


Figure 1. The simplest telephone system permitting facility sharing is a group of key telephones, which permit access to a number of lines by several users

phone exchange. Most key systems have a means by which the associated telephone sets can signal and, possibly, converse, with each other. The PBX/PABX Systems are small telephone switching centers that allow the users to dial each other or dial out to the local public telephone exchange. Each of these classifications is analyzed in the following paragraphs. Typically, the key systems are useful in business environments with only a few telephone sets. The PBX/PABX systems are generally necessary for those applications that have a large number of telephone sets.

When under 25 telephone sets are required, virtually all applications can be handled by a key system. But with more than 50 telephone sets, only a PBX/PABX type of system will meet typical telephone usage objectives. Between 25 and 50 telephone sets, either approach may be warranted depending upon the environment's operational characteristics and demands.

KEY TELEPHONE SYSTEMS

As shown in Figure 1, each telephone set associated with a key system must access the outside telephone lines. Since more than one outside telephone line is usually associated with the key system, button telephones or key telephones must be used. All incoming and outgoing calls must be placed over these shared outside telephone lines.

The usual practice is to designate one of the key telephones as the primary answering point. All incoming calls are answered at that point. The receptionist must then locate the desired called party and instruct them to access that outside telephone line, which has the incoming call. In order to permit the incoming call to be announced and the desired called party instructed, a means of intrasystem communications among the associated telephone sets must be provided.

The required intercom capability can be provided for one or two simultaneous conversations. The system would thus be referred to as either a one link or two link key system, respectively. Usually, access to the intercom circuits requires one of the line button positions on the key telephone. To call another associated key telephone, the intercom button is depressed and a code assigned to the key telephone to be called, is dialed. This will signal the called key telephone and, when the other party depresses his intercom button, their internal conversation may take place.

One major difference among key systems is the signal heard by the intercom caller if the dialed key telephone is busy or off-hook. Some key systems will return a normal busy signal, while others will continue to sound a ringing signal at the calling telephone that does not ring at the called telephone.

The number of key telephones that can be physically associated with a key system is virtually unlimited. Assuming the proper signal amplification, you could have a key system with four outside telephone lines and one hundred key telephone sets. Operationally, such a configuration would be impractical; yet assuming that the signal could be properly amplified, the company could boast that its key system operates with up to one hundred telephone sets.

Private Telephone Systems—PBX and PABX

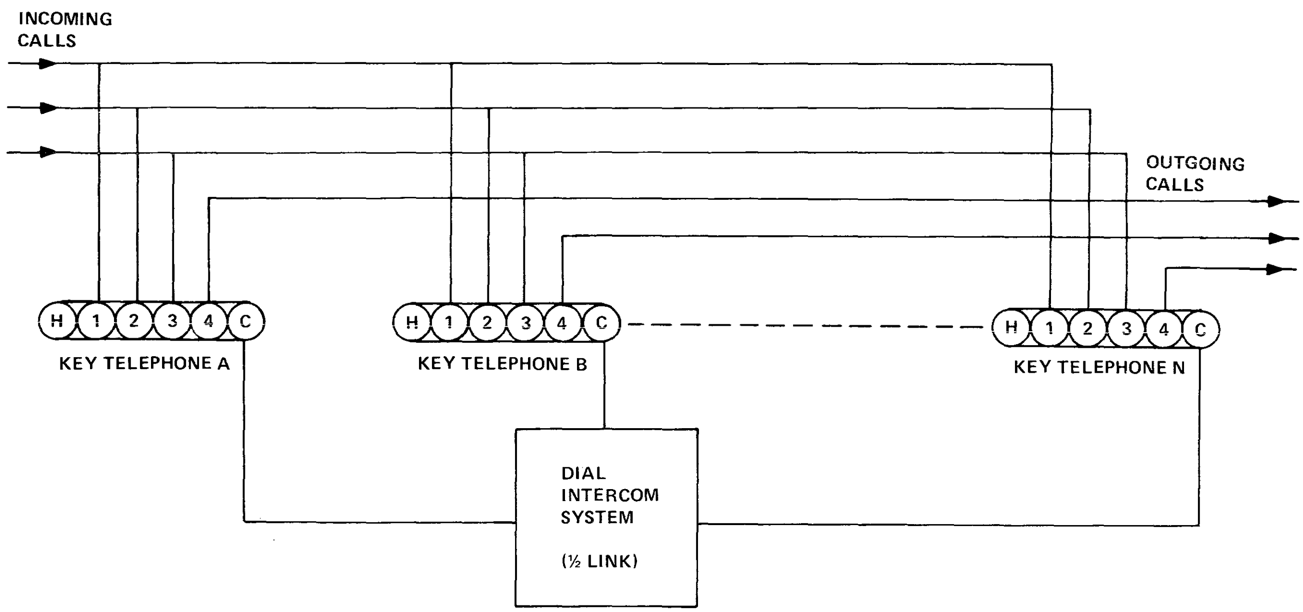


Figure 2. Imaginative grouping of lines can extend the usefulness of a key telephone system. In the expanded arrangement above, each telephone set can answer calls on any of three incoming lines or can place outgoing calls independently. Implied above, although not shown directly, these can be multiple sets of incoming trunks and a group of telephones can share an outgoing trunk or set of trunks. Such groupings must be planned around the actual mix of incoming and outgoing calls within each department as well as the total company

The critical factor in determining the practical capacity of a key system is the number of key telephone set codes that can be dialed on the intercom circuits. If only one digit intercom dialing is provided, the maximum number of key telephones is limited to ten addressable key telephones. If two digit codes are available, the capacity is, therefore, much greater. It is important that the number of unique intercom codes available be identified.

Usually the outside telephone lines are used for both incoming and outgoing calls. The published telephone directory number is assigned to the first line in the group. The feature of Trunk Hunting is always provided so that if the first line is busy, the telephone equipment at the public telephone exchange will search the other lines for one that is idle and, therefore, available to receive the incoming calls. As the number of calls terminating and originating with a key system begins to increase, a conflict develops for the available outside telephone lines. If all the outside telephone lines are occupied with outgoing calls, an incoming call will be blocked by a busy signal. Conversely, if all available lines are being used for incoming calls, personnel wishing to place an outside call must wait and hence reduce their effectiveness.

There are two possible solutions to this contention dilemma. The first is to increase the capacity of all the key telephone sets and install more outside telephone lines. This can be a costly solution. If the key system is rented from the local telephone company, the monthly rental for a ten button key telephone set may be double that required for a six button key telephone set. In addition, expansion to twelve, eighteen, and thirty button key telephone sets carries a proportional monthly license. And if you own the key system with the six button key telephone sets, you must now buy a new key system with greater line capacity and more key telephone sets.

The second method of resolving this growth congestion is to segregate the outside telephone lines for incoming and outgoing outside calls. As shown in Figure 2, a number of outside telephone lines are shared by all key telephone sets for incoming calls only. But each key telephone set has its own private outside telephone line for placing outgoing calls. This configuration is particularly effective for organizations that have a majority of their total calls originating from within. As a potential management benefit, your local telephone company will usually bill the toll from each of the private outside telephone lines for outgoing calls separately, or at least provide subtotals on a common telephone bill. This creates excellent visibility of those who properly use their telephone system and those who tend to talk excessively.

PBX/PABX TELEPHONE SYSTEMS

This classification of private telephone systems includes those devices that function as small, on-premises telephone switching centers. As shown in Figure 3, the three basic elements of a PBX/PABX telephone system are the following:

- Switching Equipment,
- Attendant Console/Switchboard, and

Private Telephone Systems—PBX and PABX

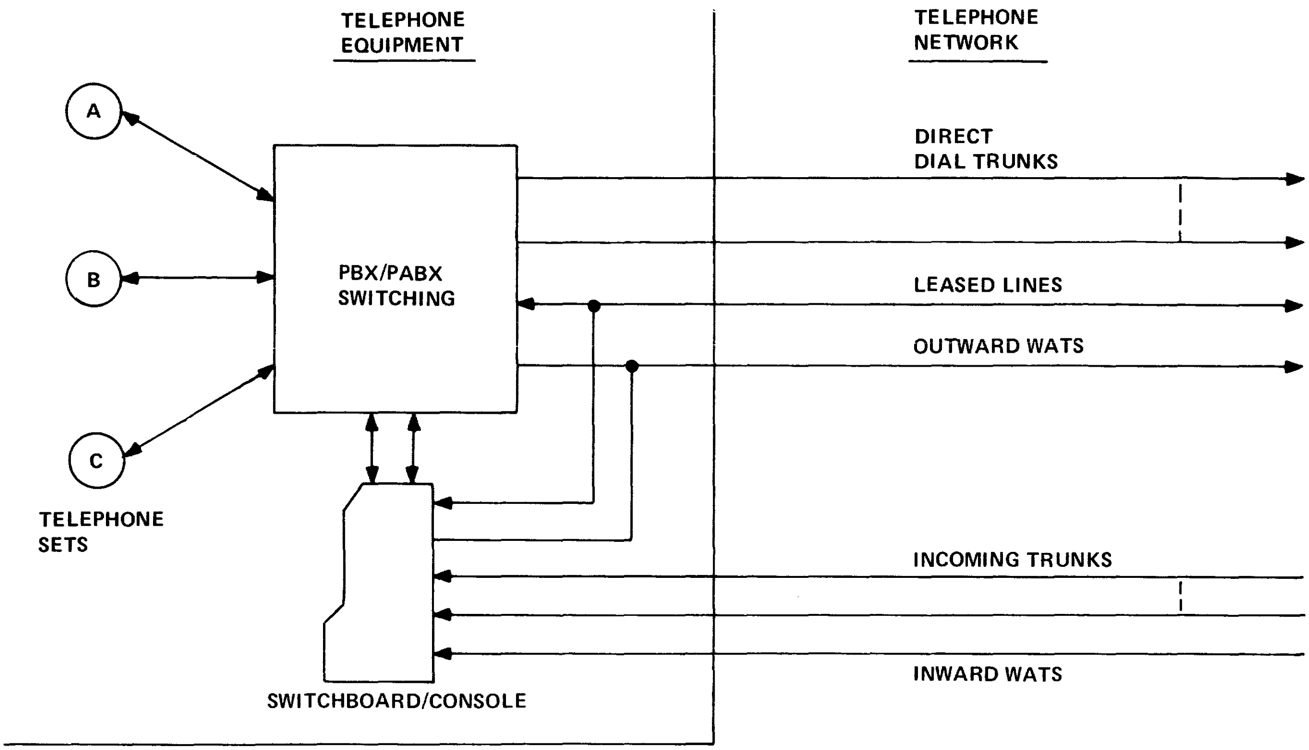


Figure 3. Basic components of a PBX/PABX private telephone system. For added flexibility, the telephone sets (A, B, and C) can be key sets with some buttons controlling access to lines not going through the PBX switching apparatus

- Station Equipment (telephone sets).

The Switching Equipment design establishes the capabilities and features that are available with a particular PBX/PABX system. The Attendant Console or Switchboard controls the answering of incoming calls and controls a number of features and capabilities for all the system users. The difference between a console and switchboard is usually that the console is all push-button operation while the switchboard uses the traditional jacks and cords to establish connections.

The Station Equipment comprises the actual telephone sets used by PBX/PABX system users. The terminology of "station" is generally used to refer to a telephone set and that telephone set user is generally referred to as the "station user."

Other configuration concepts associated with the PBX/PABX systems refer to "primary" and "secondary" lines. A "primary line" is the first appearance of a specific extension number on a station. The "secondary line" permits other stations to pick up or also have access to that same extension number.

Each of these PBX/PABX system elements are analyzed below.

Switching Equipment

Since the Switching Equipment is the heart of the PBX/PABX system, it is also the most critical. The availability or non-availability of features is dictated by the manufactured design of the switching equipment. Also, the capacity and expansion limits of a particular PBX/PABX system are directly controlled by its design.

There are only two types of switching equipment design in use today. In this context the types are not concerned with the physical components being used, but primarily with the configurations of those components. The two types are Step-by-Step (Sequential), and Common Control. These switching design types will be discussed later in this report.

But with respect to component technology, we can identify the following electro-mechanical component possibilities for both the switching which connects the telephone circuits and the control which directs which telephone circuits are to be connected:

Group A—Switching:

- Strowager—electromechanical,
- Crossbar—electromechanical,

Private Telephone Systems—PBX and PABX

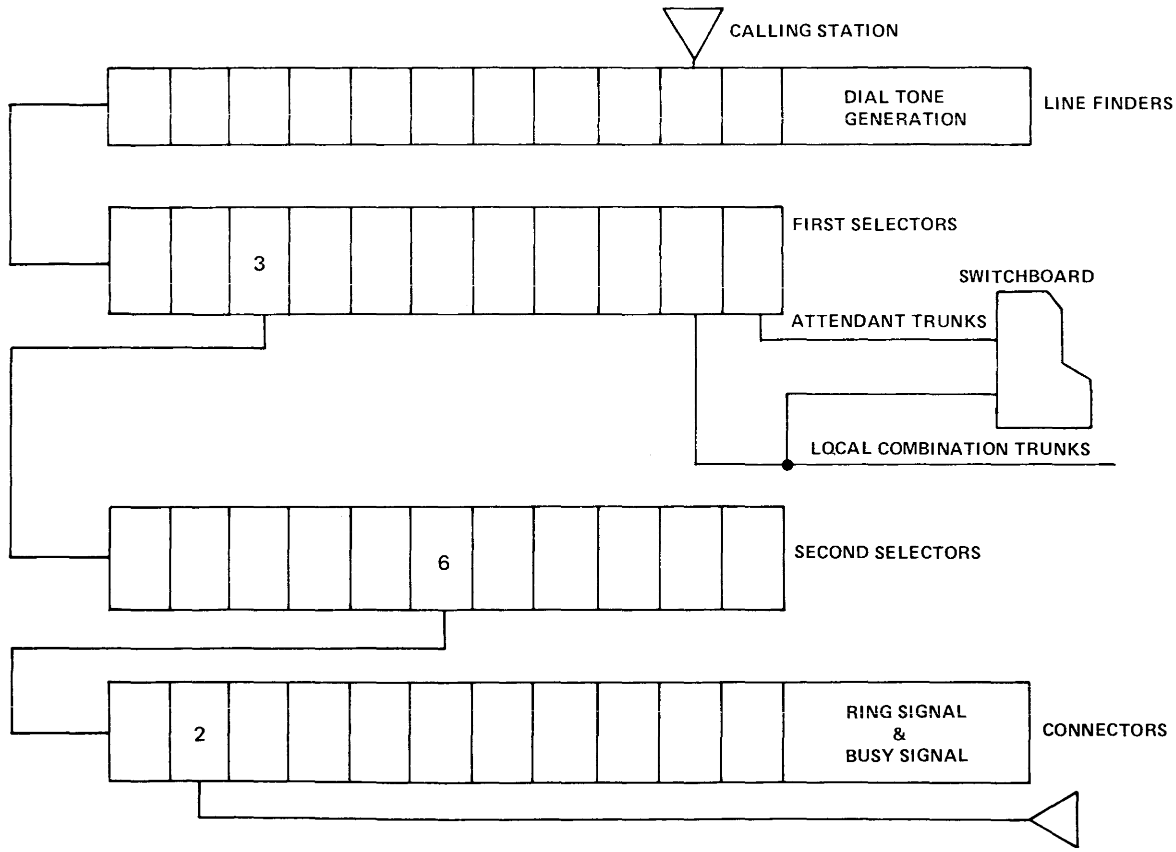


Figure 4. Typical Step-by-Step PBX system diagram

- Reed Relay—electromechanical,
- Semiconductor—electronic, and
- Time Division Multiplexed—electronic.

Group B—Control:

- Strowager—electromechanical,
- Multicontact Relay—electromechanical,
- Transistorized—electronic, and
- Computerized—electronic programmable.

By taking any item of Group A and putting it with any item of Group B, you will be able to define a total PBX/PABX system that probably exists. This potential variety of component designs creates much of the confusion associated with selection of PBX/PABX systems. The actual capabilities and capacities of the system are, however, not primarily dictated by the physical componentry, but by the design of the system, i.e., either Step-by-Step or Common Control.

The oldest design configuration is the sequential or Step-by-Step system. That builds up a telephone connection as each digit is dialed. The different levels or selectors are interconnected by a limited number of electrical paths.

With this type of system it is important that the calling loads and utilization be carefully analyzed, both with respect to internal calls and to calls to the outside public network.

These systems provide virtually unlimited station and trunk capacity. To increase capacity, all that's required is to add more line finder groups for call origination and connectors for call terminations. The simplified diagram shown in Figure 4 shows the connection path for a calling station to be connected to extension 362. If the calling station wished to reach the switchboard attendant, dialing "0" would connect the calling station to the last position in the first selectors, which are the attendant trunks. To place a call to the outside public telephone network, dialing "9" would access an available outside telephone line and the calling station would hear dial tone from the public telephone exchange.

Private Telephone Systems—PBX and PABX

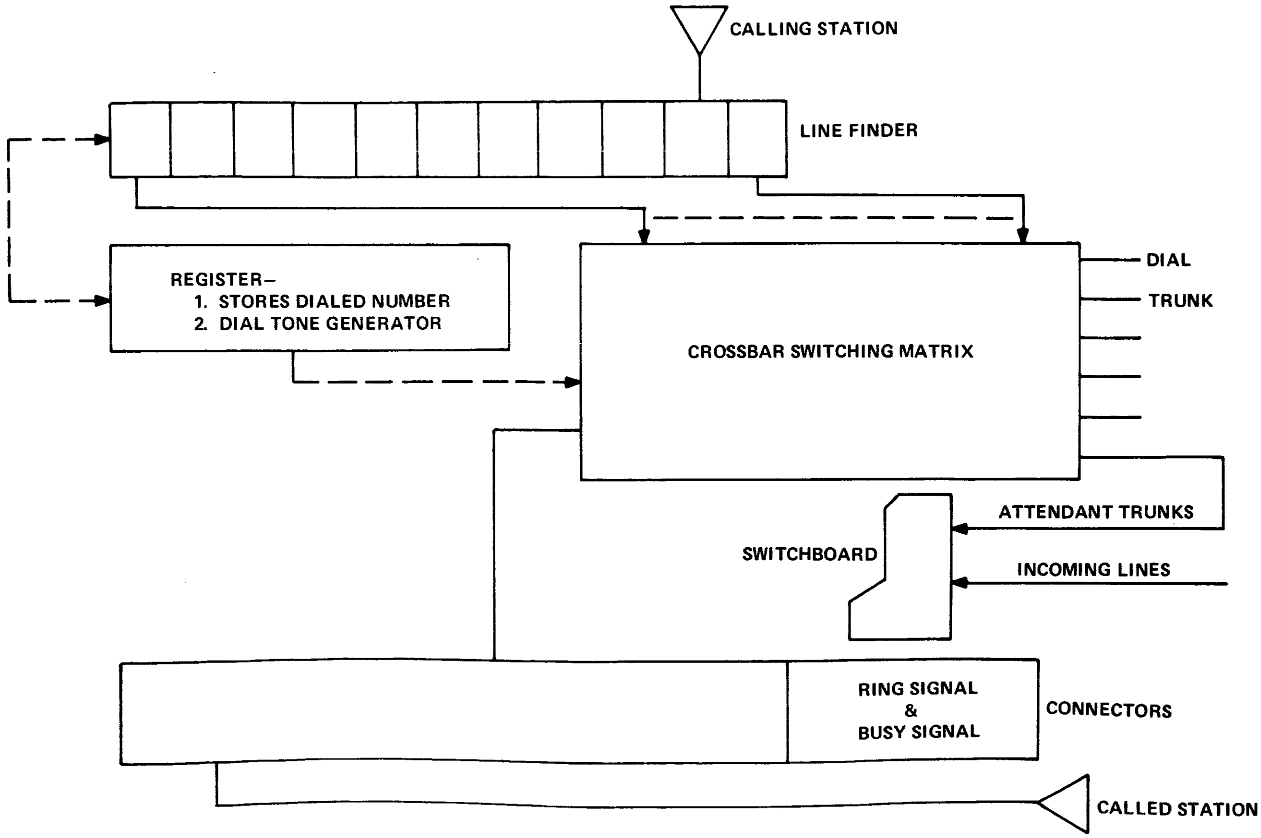


Figure 5. Common Control crossbar PBX system diagram

One of the major disadvantages of these traditional "workhorses" of the telephone industry is the inability to capture the dialed number before activating the switching mechanisms. If the dialed number could be captured, the connection line would be quicker and, therefore, more efficient use made of the shared switching mechanism. For example, an attempted call to a busy station could possibly be checked before utilizing the capacity of the switching mechanism. Other features such as station class of services, abbreviated dialing, etc., could also be provided.

A major disadvantage of the Step-by-Step arrangement is also derived from its greatest advantage. This advantage is its inherent reliability due to redundant switching paths and the virtual absence of any significant common equipment. This reliability is basically due to the multi-tiered configuration of the switching equipment. Unfortunately, however, this design demands significant floor space and impressive power requirements. The Step-by-Step PBX/PABX has been characterized as telephone switching by brute force. Today there are still over 80,000 public telephone exchanges in the United States that use the Step-by-Step switching configuration. In addition, there are well over 150,000 PBX/PABX systems in operation that are of a Step-by-Step design.

The other major design type, the Common Control PBX/PABX system, uses a configuration similar to that shown in Figure 5. The major difference between the two types of systems is that the Common Control systems first store the entire dialed number before the switching mechanism is activated. This method of operation overcomes a number of the disadvantages and limitations of the Step-by-Step design. Different and more simplified switching devices can be used when all digit recognition and decoding functions are assigned to centralized electromechanical or electronic devices.

All Step-by-Step systems use a version of the Strowager switch which was originally designed in 1893. This switch had to recognize and decode a dialed digit and also close the necessary connections for the telephone circuit.

By contrast, the most familiar switching matrix used in a Common Control system is the crossbar switch. This device will typically connect two or four conductors for the telephone connection. The X-Y configuration as shown in Figure 6 permits the direct connection of the proper station to the proper trunk. The trunk as shown in this diagram may be another station associated with the system or the attendant's position.

Private Telephone Systems—PBX and PABX

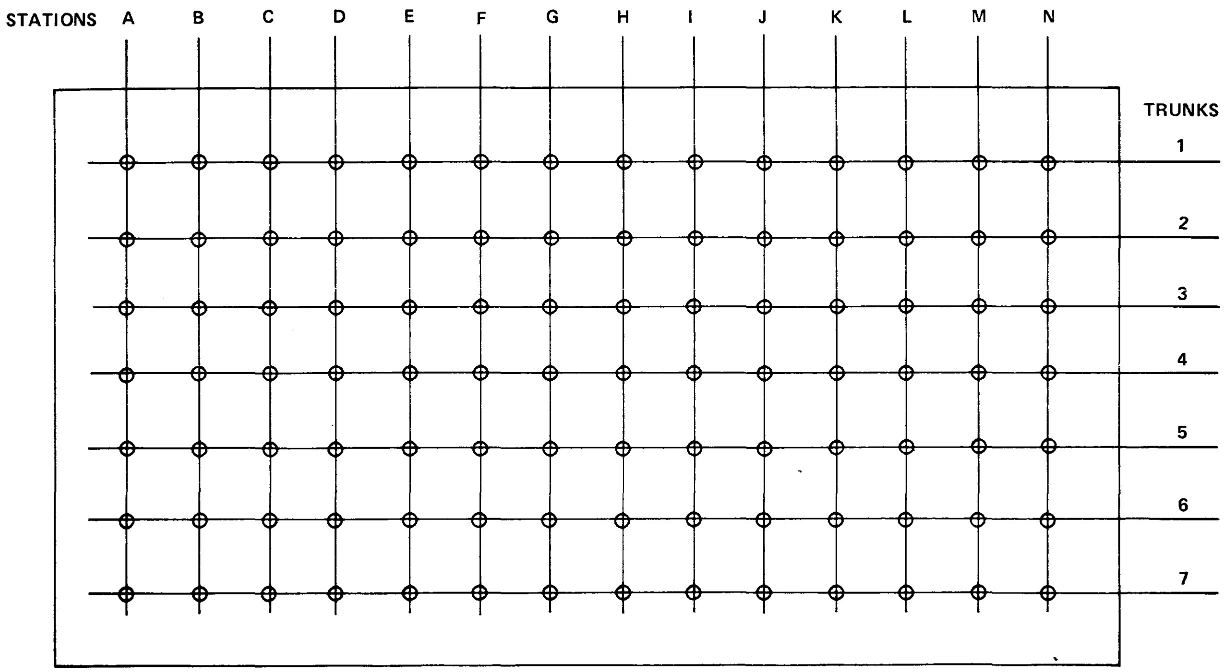


Figure 6. Crossbar switching matrix diagram

Other switching devices include the reed relay and various electronic semiconductor switching methods. There also exist a number of PBX/PABX systems which do not physically switch electrical connections, but which time-divide each telephone connection. The voice conversation is actually sliced into thin time pieces. A few of these pieces are mixed in a known sequence with pieces of other telephone conversations. All the intermixed time slices are transmitted over a common connection to the outside telephone lines. At that point the proper slices are distributed to appropriate output.

Attendant Console/Switchboard

Somewhat analogous to the communications command post, the attendant console or switchboard is the control point of the PBX/PABX system. In the early PBX systems before the concept of "automatic" was applied, the switchboard was the PBX. The operator, or as presently called, the attendant; represented a complex, common control system requiring manual dexterity to handle the switching mechanism and demanding above-average employee fringe benefits. The operator stored the desired number prior to activating the switching mechanism and was even able to properly place a call when no number was identified but only a desired name or address presented. These manual switchboards required operator intervention for intra-system station calling, for the processing of incoming calls, and for originating outside calls.

With the introduction of automated telephone systems, one of the major objectives was to reduce the number of tasks required of the attendant. In today's thinking, the attendant's primary activity is to properly represent the business' image to incoming callers and to efficiently route those calls to the desired called party. A second important activity that might be justified might require the attendant to direct, or at least monitor, the utilization of high-cost outside trunks, such as WATS lines. These objectives can usually be realized by two methods:

- Remove unnecessary attendant tasks that can be performed automatically or by the station users.
- Provide the attendant with features that permit the primary tasks to be accomplished in a more efficient manner.

An attendant switchboard permits and requires the direct connection and maintenance of that connection for all calls completed by the switchboard. An incoming call answered by the attendant must be physically connected to the desired station. When the call has ended, the switchboard attendant must physically disconnect that connection. A switchboard is a continuous and direct connection device.

A console, by contrast, primarily functions as a monitoring vehicle. It becomes involved in the processing of a call only during the transition periods. Beyond that point the console is no longer directly involved in that call. An incoming call is normally answered at the console. Through the use of a busy lamp field, the attendant can visually determine if the desired called station is available. Assuming that the called station is idle, the console attendant merely push-button switches the incoming call to the called station. At this point the connection is made in the associated switching equipment. When the call terminates, the call automatically disconnects without any console intervention.

These differences between an attendant console and switchboard constitute their basic functional differences. The availability of the console and its features, however, also provides the opportunity for other functional efficiencies that, because of the demands required by switchboard operating, could not otherwise be practically considered.

Station Equipment (Telephone Sets)

Generally, the station equipment or the user telephone sets are transparent to the private telephone system. The same type of station equipment is used on both key systems and PBX/PABX systems. Basic station equipment can include the following devices:

- Single Line Telephone Set
- Two Line Telephone Set
- Six Button Telephone Set
- Ten Button Telephone Set
- Twelve Button Telephone Set
- Eighteen Button Telephone Set
- Thirty Button Telephone Set

In addition, the station user may also add peripheral devices such as speaker telephones, automatic answering equipment, etc., to his basic station equipment.

Certain manufacturers may provide station equipment that is uniquely designed for the associated private telephone system. This is more commonly done with key systems rather than with PBX/PABX systems.

INTERCONNECT

Since the 1968 Carterphone decision, interconnect has been continually in the news. The outcome of the 1968 decision allowed modems and other devices to be directly connected to private (leased) lines. The FCC deferred to the telephone companies and permitted imposition of a connecting arrangement device between any equipment and the public telephone network to "protect" the network. At that time, devices connected to the public telephone network were not considered part of the telephone service, and regulation was governed by local regulatory agencies, not the FCC.

Beginning in 1972, the FCC started proceedings to investigate the direct (i.e., without any connecting arrangement device) connection of devices to the public telephone network. The outcome of this study was a ruling on November 7, 1975 that instituted a registration/certification program for any device other than main telephones, PBXs, key telephone systems, and coin telephones to be connected directly to the public telephone network. The effective date was to be May 1, 1976. In May, the FCC added the other equipment to the list, with an effective date of August 1, 1976. Also in May, the U.S. Circuit Court of Appeals in Richmond, Virginia, implemented a stay in the execution of the FCC ruling. In June, that order was vacated, so the registration program is on again for modems and ancillary devices, including independently made connecting arrangements, but the stay for PBXs and telephone systems is still in effect.

Permission to connect independently manufactured DAA's, or connecting arrangements, if certified, permits the connection of non-certified equipment to the public telephone network without resort to a telephone company device.

Discussion of the legality and morality of interconnection is great coffee-time conversation, but the "whether and when" is of far greater significance in planning your telephone communications arrangements. Undoubtedly, the direct connection of certified PBX and key telephone equipment from independent manufacturers will be implemented. It may have happened by the time you read this article. Meanwhile, it is a worthwhile talking point with the vendors of independent telephone systems. Make sure that they clearly address the problem in any proposal. Pinning them down on just how their system will be connected to the outside world is a worthwhile inclusion in your investigations.

CENTREX, DIMENSION, AND CARNATION

Any discussion of private telephone systems would be incomplete without mention of what the Bell System and IBM are doing. IBM? Yes. The computer giant is making several moves in the non-data processing communications industry.

As an outgrowth of the Bell System's development of stored program (computer) controlled switching gear (called ESS or Electronic Switching System), and before independents were allowed into the PBX

Private Telephone Systems—PBX and PABX

arena, Bell introduced the Centrex system. Available with the switching center located on telephone company or customer premises, Centrex was highly advanced for its day. It was also expensive. Some of the features incorporated included direct dialing of extensions from locations outside the company's switchboard, call forwarding (rerouting calls automatically from one extension to another), hunting (automatic seeking among a specified group of extensions for one that was not busy), pick up of the night line from any extension, call hold from all extensions, direct call transfer from one extension to another, conference call set up from any extension, and abbreviated dialing (a dialed 3-digit code was automatically extended to 7 or 10 digits). Debugging early systems was fun if frustrating. In addition, personnel had to learn new telephone usage habits. No longer did a brief call to the operator suffice to get a number, or set up a conference call, or transfer a call, etc.

The Bell Dimension system was introduced in January 1975 as a lower capacity, lower cost system with reduced features. The original Dimension 400 (up to 400 stations) was supplemented in June 1976 with the Dimension 100 (up to 100 stations), Dimension 2000 (lots and lots of stations), and Dimension Custom Telephone System (if you can describe a feature, it can probably be implemented). Also introduced for the whole Dimension line was call monitoring (recording of usage on a per station, per message basis), and alternate facility routing (automatic routing of outgoing calls over WATS, FX, or other facilities).

IBM entered the PBX market in Europe awhile ago with a system with the internal project name of Carnation. To the external world, it was called the 3750 Switching System. Built around a minicomputer (what else?), the 3750 includes special facilities for accommodating digital data transmission. Not only does it handle conventional terminals, but it also encourages the use of the telephone instrument itself for data entry and inquiry, with replies generated by a voice response unit (not included in the 3750 system). Plant security and monitoring devices can also be accommodated.

Many of the features pioneered by Bell's Centrex and present in Dimension are available with PBX equipment from independent vendors. Each particular system offers combinations of cost and features that make it ideal for a particular communications environment.

The very latest PABX equipment is probably best represented by the microprocessor-based systems produced by a relatively small firm, Mitel, Inc., which offers a 40-line 8-trunk system for about $11,000. Larger companies with less mobility are rapidly catching up, but the immediate future probably belongs to the smaller companies who can quickly incorporate the fruits of LSI/VLSI technology into their equipment without having to worry about disrupting a large user base.□

The Equipment and Techniques of Digital Telephony

Problem:

The general movement toward transmission facilities with more usable bandwidth is prompting a parallel movement toward standardized all-digital encoding techniques for every possible type of information input to a communications network. Voice inputs are one of the last holdouts because quite a bit of information can still be transferred cheaply on a very limited band of about 3000 or 4000 Hertz. To convert a standard voice channel to an equivalent digital channel, the voice channel needs to be sampled at a rate of roughly 8000 Hertz. Each sample typically yields an 8-bit code, so a 3000-4000 Hertz voice channel requires a 64Kbps digital channel to transfer the same information. A developing communications standard is the 32-channel 2.048 Mbps digital "highway", which could handle over 500 standard (nondigital) voice channels. From a sheer capacity point of view, standard voice channels obviously use a given bandwidth far more effectively than equivalent digital channels, by a ratio of roughly 16 to 1, so what is the advantage of digital telephony? The main advantage is vastly improved switching control. Standard voice traffic uses human-compatible but grossly wasteful line switching techniques; digital traffic uses far more efficient message and packet switching techniques.

The introduction of inexpensive microprocessors and LSI/VLSI circuitry into communications switching equipment is rapidly wiping out the 16:1 volume advantage of nondigital voice transmission techniques and is tipping the scales sharply in favor of the switching advantages to be gained by digitizing voice inputs. This report is offered to clarify those advantages and to give you a perspective on the current status of digital telephony equipment and techniques for your current and future communications expansion plans.

Solution:

The telephone industry in the United States and abroad is in the process of adopting digital technology for interoffice transmission systems, local and tandem (interoffice trunk switching) offices, private automatic branch exchanges (PABX's) and in a few instances, local subscriber loops and telephones. The impetus behind the recent acceleration in the evolution of practical fully digital telephone switching was, for the most part, the availability of microcircuit technology at a price competitive with other hardware alternatives, and the proven reliability of supply and performance of devices based on large scale integration (LSI).

The development of the microcomputer as a small, powerful, and versatile tool for use by digital switching and transmission design engineers led to the develop-

"Microcomputer Applications in Telephony" by Donald K. Melvin from *Proceedings of the IEEE,* Vol. 66, No. 2, February 1978. Reprinted by permission.

ment of many new digital systems in the mid-1970's. Microcomputer-controlled PABX's have since been announced by several major manufacturers of telephone switching systems. Independent (non-Bell) manufacturers have already introduced digital PABX's and digital toll offices (class-4) and digital local or end offices (class-5). Bell System has already cut over to service its first toll tandem (trunk switching) Number 4 Electronic Switching System (ESS) offices and is also producing its Dimension[1] PBX. Most of these switching systems are microcomputer controlled to some degree, and use digital transmission techniques (see Reference 1). Prior to the availability of microcomputers, most new electronic telephone switching systems were designed around space-division switching (usually using reed-relays) and custom computer controlled architectures. Time-division multiplex (TDM) switching systems with digital control in which voice channels from lines and trunks (interoffice circuits) are sampled and converted into either pulse-amplitude modulation (PAM), pulse-code modulation (PCM), or delta modulation (ΔM) signals for switching purposes appeared on the market on only a limited basis until the inexpensive microcomputer became available in mid-1972 (see References 2 and 3 for early TDM systems).

MICROCOMPUTERS IN TELEPHONE SWITCHING SYSTEMS

Even after the early microcomputers such as the Intel 4004 and the 8008 became available (Reference 4), it was several years before their usefulness in telephone systems became fully appreciated. Telephone system designers had been accustomed to designing common control switching systems with wired logic and circuit redundancy for reliability, and for the most part were not familiar with the programming techniques and interface specifications of microcomputers. Also, their power and flexibility of application were not fully appreciated. Programmed control has been used for many years in telephone switching systems, however. As more powerful models of the microprocessor became available, such as Intel's 8080A, 8085, and 8048, and similar products from other semiconductor manufacturers, design aids were also developed to allow simplified programming techniques and more convenient interface access.

A number of telephone switching system architectures had been proposed as early as a decade ago, but they were not economically feasible until low-priced memory, control, and logic functions became available through LSI technology. The distributed switching system (one system that utilizes remote digital line concentrators to provide central office or private branch exchange types of service at some distance from the central exchange) is now evolving around new LSI technology. These remote line groups tie into the central exchange with digital trunks for voice path switching, control, routing, billing, access to long distance circuits, etc.

The trends now indicate that multiprocessor, distributed control approaches in modular switching system architectures are appropriate applications of microcomputers where full use of their capabilities can be realized, and system objectives can be met.

Microcomputer control is now finding its way into telephone systems such as PABX's, key telephone (multiline) systems, line concentrators, central office switching systems, and a variety of related test equipment and service monitoring systems.

Typical microcomputer applications include common control or central processing units for PABX's, line group control for central offices, remote line concentrators, remote line groups (nonconcentrating), and distributed multiprocessor applications in central office systems. Also, they are being applied in electronic telephones, electronic key telephone systems, and other business telephone systems.

Since microcomputers are now becoming used in such a wide variety of switching, concentrating, and control applications in telephony, a basic small PABX has been chosen as an example of a typical microcomputer system application. The PABX described here incorporates many of the elements of switching and transmission control as well as line scanning, signaling, and supervision and is thus representative of the kinds of tasks that can be performed by a microcomputer in a practical system.

A MICROCOMPUTER CONTROLLED DIGITAL PABX

Figure 1 shows the functional block diagram of a typical PABX that uses PCM techniques for transmission within the system and digital switching techniques. This digital PABX is a 100-port microcomputer-based telephone switching system that uses T1[2] compatible transmission highways (one for each direction of speech transmission), with a serial, 24 time-slot, 8-bit per time-slot digital transmission format. One hundred lines can be comfortably handled by one microcomputer. More are possible if the system service features are limited.

The Codec (ideally a single-chip coder-decoder) is used for A/D and D/A conversions in each line and trunk circuit. The PABX may use a stored program central processing unit (CPU) for common control functions, including assignment of time-slots on the digital

[1]A Trademark of American Telephone and Telegraph Company.
[2]Note: T1 is AT&T's designation code for the North American digital transmission line standard.

The Equipment and Techniques of Digital Telephony

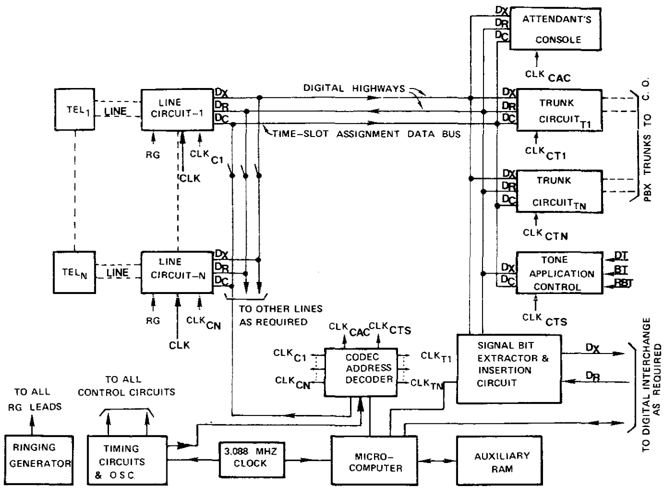


Figure 1. Block diagram of PCM-PABX

transmission highways, and all-call processing functions such as dial pulse registration, ringing control, and trunk signaling.

All analog signals received over the connecting lines and trunks are converted into digitized equivalent signals to facilitate time-division-multiplexed switching of the digitized voice and signaling samples between various time-slots on the transmission highways. The time-slots are computer assigned as required by the traffic to interconnect calling and called lines and trunks.

The time bases for the common control CPU and the digital transmission highway may be different and isolated from each other. The computer may operate from its own internal time base, while the transmission highways operate at a 1.544-MHz bit rate. A 3.088 MHz clock is divided by two to produce the 1.544 Mbps T1-compatible transmission bit-rate, and two clock phases are available for transmission-related control.

The serial highways are multipled to all line and trunk circuits as well as to an attendant's console control and a tone application control. These highways are the transmit data highway (D_X) the receive data highway (D_R), and the computer data highway (D_C). Highways D_X and D_R are the 1.544 Mbps transmission highways, and they need not operate in phase with each other. In a 100-line PABX with just single transmit and receive highways, they may be operated synchronously. Highway Dc carries computer generated data for time-slot assignment to the codecs and thus operates at a rate determined by the computer. The codec may operate with either on-chip or off-chip generated time-slot clocking pulses. Both transmission highways (D_X and D_R) carry bit-streams having a 125 μs frame period (for 8 kHz sampling of each voice channel). Each frame is composed of 24 8-bit time-slots, plus one frame synchronization bit, totalling 193 bits per frame.

Each 8-bit time-slot carries digitally encoded amplitude samples for one direction of transmission of one conversation. The telephone line or trunk, etc. assigned to a given transmission highway time-slot may change from moment to moment under computer control. Each port does not have a fixed time-slot assignment on the transmission highways but has a variable assignment; thus the highway time-slots can represent active conversation paths or conversation "links".

LSI DEVICES FOR THE TRANSMISSION PATH

With the availability of LSI devices such as microprocessors and their peripheral devices for use in the distributed or common control portions for a digital switching system, the next requirement for an economically feasible digital switching system was the availability of microcircuit devices for use in the voice transmission path.

In a digital PABX or Class-5 central office, a digital transmission path through the system may typically include the following basic circuit functions, in various combinations:

1. line interface circuit;
2. hybrid (2/4 wire conversion);
3. transmit and receive channel filters;
4. codec (digital coder and decoder);
5. digital transmission highways;
6. digital interchange;
7. analog trunk interface circuit;
8. digital trunk interface circuit.

Most of these functions are potential new product areas for LSI technology, however, considerable architecture changes over present electromechanical and nondigital electronic architectures are required to optimize the impact of this new technology. Thus, along with the new digital hardware comes the potential for computer-like system designs, which result in new feature offerings to the user, improved transmission on long distance calls, shorter call set-up times, and faster trouble diagnosis times.

TELEPHONE LINE INTERFACE CIRCUIT

Figure 2 shows a block diagram of a telephone line interface circuit for a digital switching system such

The Equipment and Techniques of Digital Telephony

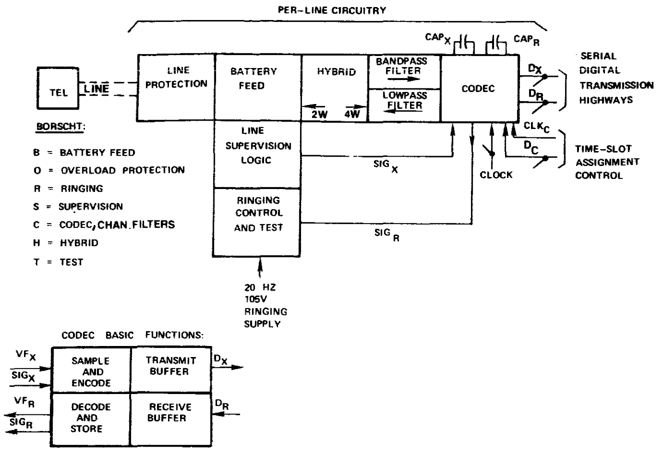


Figure 2. Block diagram of line circuit for digital switching system

as a PABX or central office. This is a single-party line circuit that serves a conventional telephone line equipped with conventional telephones such as Western Electric Company's type 500 or similar telephones with high voltage electromechanical ringers and either rotary or TOUCH-TONE[3] type dials. These telephones also have automatic varistor compensation for adjusting sidetone (your own voice heard in your receiver) and transmitter current as a function of line current, and must be accommodated by any new battery-feed design.

The telephone industry has coined the word "BORSCHT", which here refers to all the basic functions required in a digital system line interface circuit. These features are:

B battery-feed
O overload protection
R ringing
S supervision
C codec and channel filters
H hybrid
T test

The line protection circuitry protects the line circuit from overload damage by high-voltage transients on the telephone line due to lightning, electrostatic coupling from other sources, from power line induced or conducted currents, etc.

The battery-feed circuit provides a balanced dc talk battery supply to the line to provide line current for telephone sets that use a carbon transmitter and require a nominal 20-80 mA line current. This battery-feed circuit must provide a low dc impedance for battery coupling to the line (nominally 400 Ω) and an ac terminating impedance that matches the characteristic impedance of the telephone cable, nominally either 600 or 900 Ω.

The line supervision logic circuitry monitors the line loop current, and when current is detected above a certain threshold (approximately 15ma) it produces a logic level "1" for signaling the CPU that this line is engaged. This circuit also repeats rotary dial loop pulses (when required) to the CPU in the same manner as for off-hook supervision.

Ringing control circuitry provides a means for applying interrupted high-voltage ringing signals to the telephone line to ring a called telephone. This circuitry applies ringing from a common ringing generator (nominally 20 Hz, 100 V) by means of a relay or other switching circuit that removes the normal battery-feed circuit while applying the ringing voltage to the line. Also, this circuit provides the circuitry for ring-trip, i.e. disconnecting and locking out the ringing voltage once the call has been answered.

Test circuitry consists, usually, of a relay that transfers the line to a test circuit and extends the line to a test desk or other test facility. A test relay is usually used to perform the transfer under control of the CPU.

The line hybrid circuitry is used at the transition between two-wire (two-way) or four-wire (separate transmit and receive) operation. Digital transmission is unidirectional in nature, so separate transmit and receive paths are necessary for bidirectional communication. The line hybrid (formerly accomplished by a multiwinding transformer) must prevent or substantially reduce the coupling between the one-way receive and one-way transmit paths (called return loss) to prevent undesirable echos or sidetone effect in the connecting telephones. These hybrids are used on the analog side of the codec.

Hybrids usually have four sets of terminals: four-wire transmit, four-wire receive, two-wire line, and balancing network. The hybrid is usually balanced for maximum echo cancellation by closely matching the balancing network impedance to the 2-wire line impedance. Adaptive echo cancelling techniques may also be used to enhance the performance of the hybrid to reduce the critical nature of network balance, and to give improved return loss over the entire channel bandwidth for all line lengths to be accommodated.

When a codec is used to encode and decode digital signals, channel filters are required to limit the bandwidth of the voice channels. In the transmit direction, the bandpass is nominally 180 to 3400 Hz, according to AT&T's D3 channel bank compatibility specification and a comparable CCITT[4] specification. Also,

[3]TOUCH-TONE is a registered mark of the American Telephone & Telegraph Company.

[4]CCITT is the International Telegraph and Telephone Consultative Committee of the International Telecommunications Union (ITU), Geneva, Switzerland.

The Equipment and Techniques of Digital Telephony

the receive channel filter is a low-pass filter that passes frequencies below 3400 Hz. The channel filters must be compatible with the codec analog interfaces.

TRUNK INTERFACE CIRCUIT

Figure 3 shows a block diagram of a trunk circuit that interfaces a PCM-PABX with a conventional 2-way PBX trunk from a central office. In this circuit, the codec, filters, and hybrid shown are the same as used in the line circuits. The trunk signaling logic provides ground-start circuitry (to prevent call collisions), trunk-loop closure, loop pulsing, and ringing detection circuitry. Both inward and outward supervisory signals may be transmitted to the PABX common control microcomputer via the codec signaling leads, or via special signaling buses multipled to each trunk circuit board. One-way trunks are similarly implemented.

Solid-state implementation of most of this circuitry can be achieved using high voltage devices, and varistor and diode protection of the circuitry against ringing voltages, switching transients and other potentially hazardous voltages and currents. A protection specification being adopted by some manufacturers states that line and trunk protection circuits must protect against transients of 1000 V at 1 A for 1 ms.

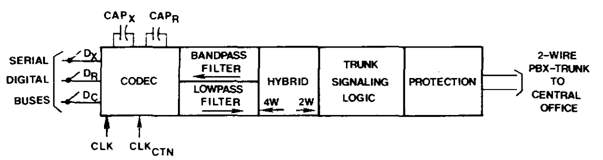


Figure 3. Two-wire PBX trunk circuit for digital telephone switching system

THE CODEC

The codec is the key component in the digital switching system line or trunk interface or in a channel bank channel unit. The basic function of the codec is to provide a means for encoding transmit direction voice signals into a 64 Kbps bit-stream for decoding a similar received bit-stream into receive direction voice frequencies, and, in the North American version, to provide an end-to-end per-channel signaling means. For North American μ-law[5] codecs, one signaling channel in each direction of transmission must be provided by the codec for each voice circuit. Signaling information is transmitted in the eight bit once in every sixth transmission frame by borrowing the eighth bit from the least significant position of the eight-bit code used for PCM channel transmission. Actually, when the signaling channel is provided, the voice channel uses 7-5/6 bit encoding of the voice signal instead of the full 8-bit coding.

[5]The μ-law is a mathematical law that defines the compressing and expanding (companding) characteristics of the codec. Companding reduces the quantizing noise for low-level signals by using finer quantizing steps at low levels than at high transmission levels. A similar A-law is used as a European standard companding characteristic.

When 24 similar codecs are connected to the same transmission highways, the highway bit-rate may be T1 compatible, i.e., operate at 1.544 Mbps. The codec may handle bit-rates up to twice this rate and handle up to 48 codecs on one pair of transmission highways. This makes it compatible with existing 24, 30, 32, or 48 channel system architectures or with any other desired configuration of codec-per-line groupings from one to 48. The 24 and 48 channel systems usually use μ-law codecs (e.g., Intel type 2910), and the 30 and 32 channel systems usually use the A-law codecs (e.g., Intel type 2911). (See Reference 8).

Companding methods used on the digital signals can match the international transmission standards of the μ-law or A-law as specified by the CCITT (see Reference 5). The μ-law companding characteristic is used primarily in North America, and the A-law primarily in Europe.

The application of the codec in a telephone switching system is enhanced if it is capable of dynamically changing its internally stored channel time-slot assignments under control of a microcomputer that serves that line group. Time-slot assignments need not remain fixed as they are in carrier system channel bank applications but can be changed for each call, or for tone assignments, conversation, or conference modes of operation. This allows switching system architectures that provide as many highway time-slots as maximum simultaneous call requirements dictate, not one time-slot per line or codec since time-slots need to be assigned only as required.

DIGITAL TRANSMISSION HIGHWAYS WITHIN PABX

The terms transmit and receive indicate transmission direction relative to a local telephone. Figures 4 and 5 show the time-slot composition of highways D_X and D_R respectively. On D_X, the odd time-slots, TS-1 through TS-19 may be calling party transmit time-slots, and the even time-slots, TS-2 through TS-20 may be called party transmit time-slots. Similarly, on high-way D_R, the odd time-slots TS-1 through TS-19 may be called party receive time-slots, and the even time-slots, TS-2 through TS-20 may be calling party receive time-slots. Digitized voice samples on highway D_X are coupled back to highway D_R unchanged, and in the same time-slots. Signaling data bits (the 8th bit in each time-slot in every sixth frame) are modified by the signal bit extractor and insertion circuit.

The Equipment and Techniques of Digital Telephony

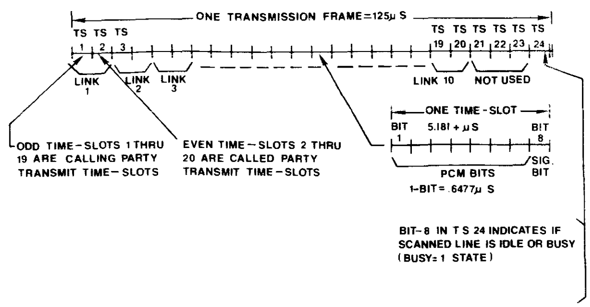


Figure 4. Transmit highway (Dx) transmission frame

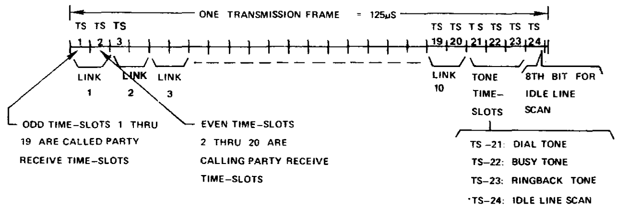


Figure 5. Receive highway (DR) transmission frame

Signaling bits on the transmit highway Dx relate to line and trunk supervision and rotary dial pulse transmission, while signaling bits on the receive highway DR relate to ringing control for the lines and signaling and supervision for the trunks and attendants console.

Since two time-slots are required on each transmission highway for each conversation, the related time-slots involved in the same conversation are sometimes referred to as a link, thus the system described here is a 10-link system. A practical PABX might have 90 lines and 10 trunks, 40 lines and 8 trunks, 20 lines and 6 trunks, or many other combinations of lines and trunks, limited only by the link capacity.

Other system architectures could use all time-slots for links with none dedicated to tone assignment or scanning. A 32-channel CCITT standard digital transmission format can be used for additional trunking capacity in PABX and remote concentrator applications.

The 2.048 Mbps, CCITT standard digital transmission format usually requires an A-law companding codec and could be used for all line, trunk, tone application and attendant's console interface circuits instead of the North American T1 standard μ-law codec described here. With the T1 format, each transmission time-slot contains eight bits, each of which has a duration of 0.6477μs., thus the period of one time-slot is 5.181μs.

CALL-PROGRESS TONES

Figure 6 shows one method of applying three of the call-progress tones (i.e., dial tone, busy tone, and ringback tone) to the system transmission highways. The tones can be generated in a digital form and applied to the receive highway DR by gates, or they can be generated by analog circuits and then compressed and digitized by using the transmit side only of a codec. With this latter method, test jacks can allow the analog tones to be monitored or they may be used for other analog applications without the need for digital signal decoders. The tones shown are those recommended for use where they may be heard over international circuits and are the same as those tones used in many new systems currently being produced in the U.S.

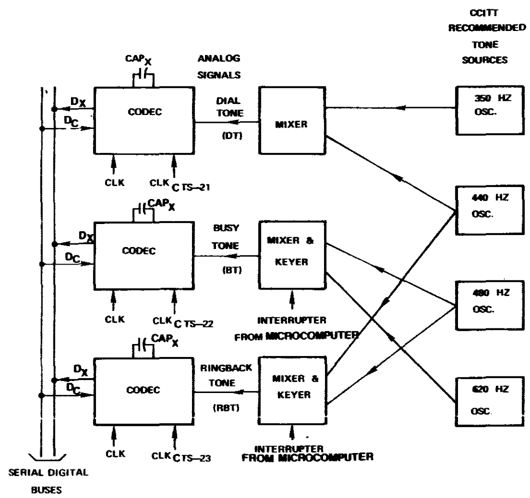


Figure 6. Supervisory tone application circuit

With the T1 format, there are 24 time-slots per frame, thus allowing 24 simultaneous conversations per highway. In the hypothetical system shown in Figure 1, three of the 24 time-slots (see Figure 5) are permanently assigned to call progress tones, namely:

Time-Slot	Tone	Abbreviation
21	Dial Tone	DT
22	Busy Tone	BT
23	Ringback Tone	RBT

Also, in order to simplify the per line circuitry, the idle line scanning function is performed via a special computer-assisted technique that uses time-slot 24 on the receive highway DR.

With 20 time-slots left for encoded voice transmission, the highways will serve up to twenty active telephones, or 20 percent of the ports in a 100-port system. If the number of ports is reduced, the grade of service can be improved. If the number of ports is reduced to 20, the system becomes nonblocking, i.e., all calls can be processed without abnormal delay. This would also be true for a 20-line concentrator with T1 trunks to a central office and no analog trunks.

The Equipment and Techniques of Digital Telephony

For higher grade of service capabilities, all 24 time-slots can be used for links, or the CCITT standard 32 channel (2.048 Mbps) transmission standards may be adopted, or the codecs may be connected for 48 channel operations, i.e., at a 3.088 or 3.152 Mbps bit-rate. In these cases, dedicated time-slots for call progress tones or idle line scanning, as described here, need not be used. Scanning can be handled independently of the codec, and tones applied from special tone ports.

PABX ATTENDANT'S CONSOLE

A block diagram of an attendant's console for a digital PABX is shown in Figure 7. Here, there are two sets of interfaces with the central control:

1. Serial data highways—
   —D_X transmit data highway
   —D_R receive data highway
   —D_C computer data highway

2. CPU I/O ports between the 8085 central processor and the buttons, lamps and other console circuits.

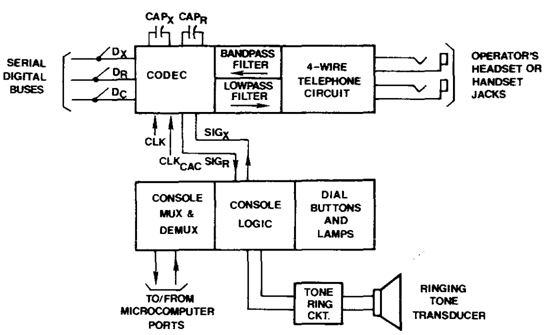


Figure 7. Attendant's console block diagram for a digital PABX

The microcomputer CPU assigns transmit and receive time-slots to the console codec via highway D_C. A two-way voice path is established between the attendant's telephone and the local telephone or trunk via the assigned time-slots on transmission highways D_X and D_R.

The console tone ringer may be keyed on by signals received via lead SIG_R of the codec, as a function of lamp and button signals where attendant action is required, or as determined by signals received from CPU ports and the button states.

Lamp signals from the CPU are demultiplexed to drive lamp drivers and lamps, and button states are multiplexed to transmit their states to the CPU over attendant console data buses. Battery-feed for the attendant's telephone circuit, indications of head-set-plugged-in, and listening key operated are transmitted over codec signaling channel SIG_X and/or the attendant's console CPU data leads.

MICROCOMPUTER SYSTEM

(Note: Intel part-numbers are shown, but many of the devices described here are also available from other semiconductor manufacturers.)

Figure 8 shows the functional block diagram of a typical central processing unit. The Intel 8085 has been shown here as an example. When used in an application such as a small conventional PABX, one microprocessor can typically handle all common control functions for several hundred lines.

Peripherally interfacing with the microprocessor are a number of memory, I/O and control devices:

1. Programmable READ ONLY Memory (PROM): This semi-permanent memory is for storing common control functions and class-of-service data for lines, trunks, and other ports. (Approximately 4K bytes of PROM are required for systems of up to about 256 ports, depending on system features.)

2. Random Access Memory (RAM): This memory is used for the temporary storage of call processing data, such as:

   a. present line status;

   b. previous line status;

   c. time-slot status/assignment;

   d. call sequence state;

   e. dial pulse count;

   f. dialled digit counter;

   g. dialled number;

   h. trunk status;

   i. source/destination address;

   j. destination/source address;

   k. special feature status and sequence;

   l. stack data for 8085 microprocessor.

Figures 9(a) and (b) show the trends in reduced component count and instruction cycle times as a function of product evolution. Figures 9(c), (d), and (e) show the CPU interfaces, and Figure 10 shows a microcomputer common control for a small PABX.

The Equipment and Techniques of Digital Telephony

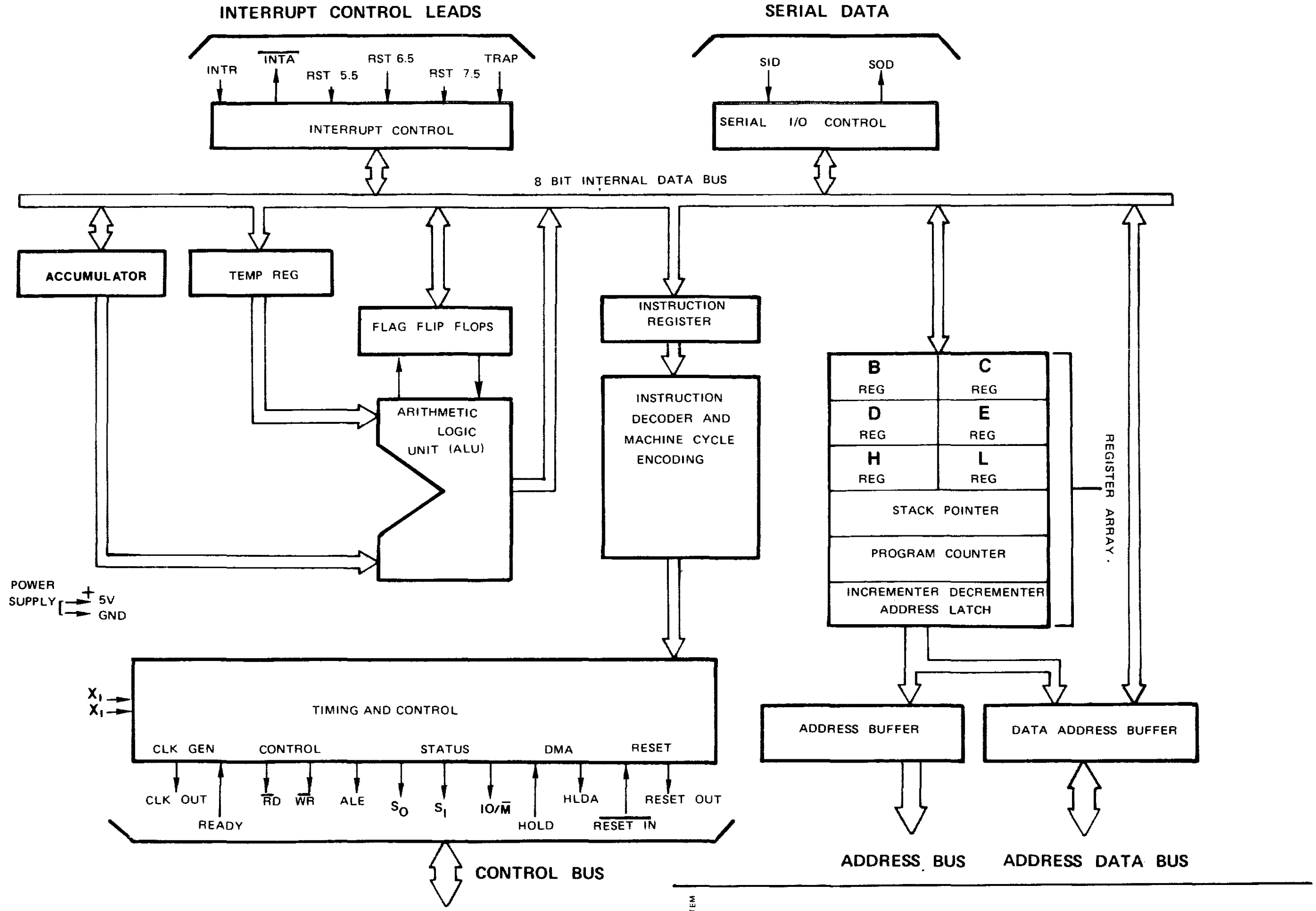


Figure 8. 8085 microprocessor

3. Programmable Peripheral Interface (8255A5): This chip provides the interface between external controlled and controlling devices and signals, and the computer parallel bidirectional data buses. These I/O ports can be used to accept and produce line and trunk signaling data, for transmission to and from all line and trunk circuits, etc.

4. I/O Driver/Terminator Interface: This is conventional TTL I/O circuitry for driving output circuits and terminating input circuits at the computer system interface.

5. Programmable Communications Interface (USART)[6] (8251A): This device provides an interface between the computer buses and Teletype interfaces or RS-232C data line or terminal interfaces. This permits direct Teletype or data line or terminal interfacing with the system central processing unit. Thus system behavior can be monitored or changed readily via local or remote terminal devices.

The USART accepts data characters from the CPU (8085) in parallel format, and then converts them into a continuous serial data stream for transmission to a data line, data terminal, or Teletype. The USART can simultaneously receive serial data streams and convert them to parallel data characters for the CPU.

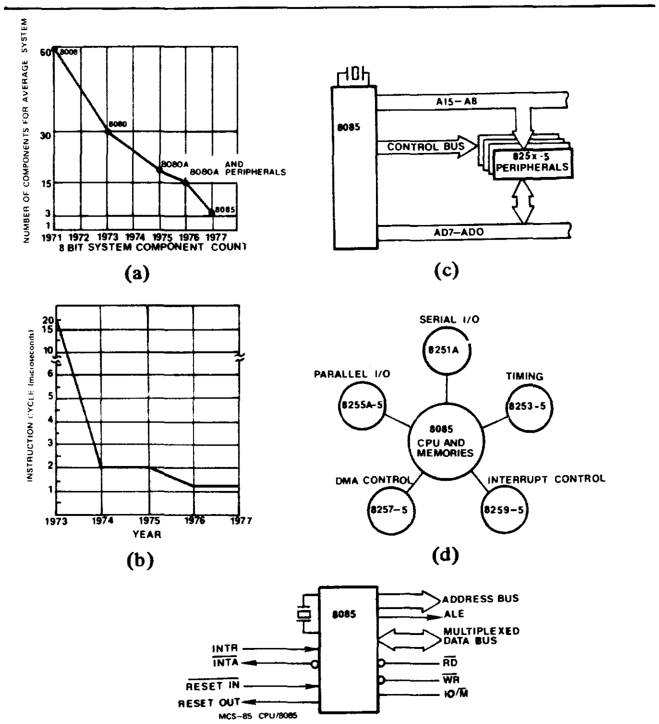


Figure 9. 8-bit microcomputer system evolution. (a) 8-bit system component count 1971-1977. (b) 8-bit system instruction cycle performance 1973-1977. (c) 8085 address and control buses. (d) MCS-85 TM PERIPHERALS. (e) MCS-85 TM CPU/8085 (basic functions).

The Equipment and Techniques of Digital Telephony

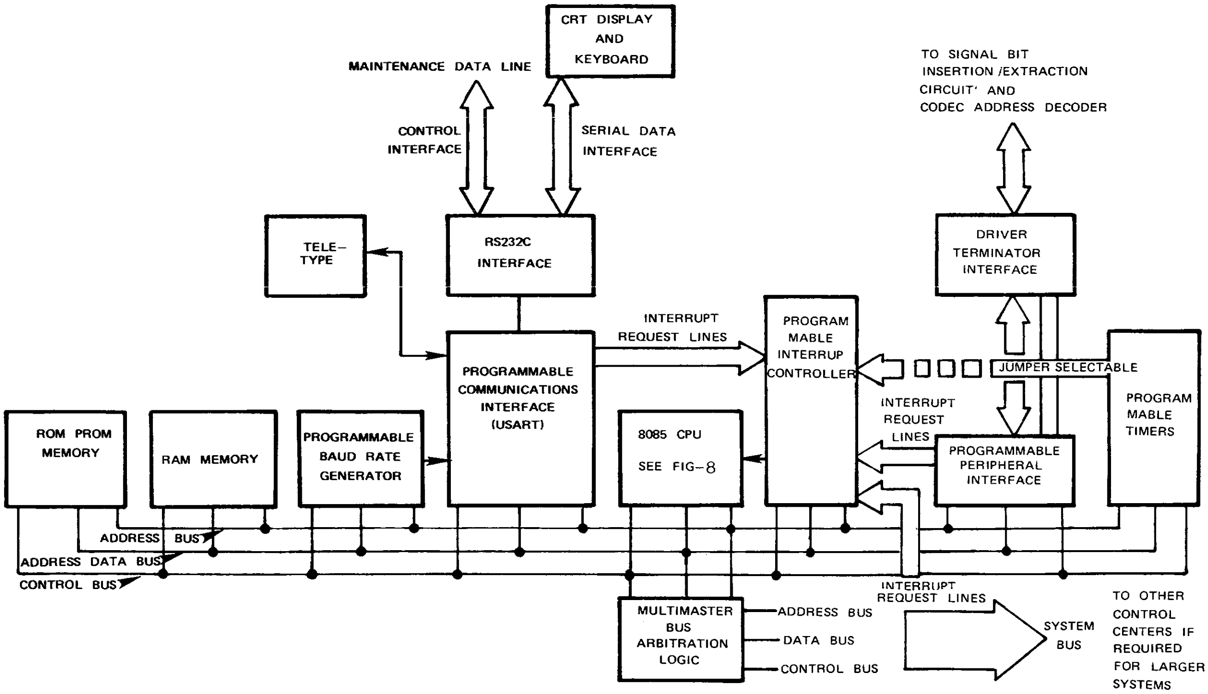


Figure 10. Microcomputer common control system for small PABX

6. RS232 Interface: This interface consists of inverting drivers which drive standard RS232 data lines, a CRT display, or a data coupler from signals produced by the 8251 USART,[6] and vice versa.

7. Multi-Master Bus Arbitration Logic (8218): This chip allows the computer system to share a common system bus with other master devices such as other CPU's, or DMA devices, thus sharing common memory and I/O resources. This device arbitrates all system requests for use of the system bus synchronously with respect to the bus clock.

8. Programmable Interval Timers (8253-5): These timers can be used for a variety of timing functions but are well suited to generating microcomputer interrupt signals to cause the computer to momentarily stop its processing action and to scan line and trunk port signaling leads. As an example, a 10ms interrupt signal can be produced by the timers to cause the CPU to scan a different idle port signaling lead every 10 ms. This interrupt function could also be accomplished under software control, but use of the interval timers allows precise control without danger of accidental modifications to the interrupt signal with software changes.

The 10 ms interrupt rate would produce an average dial-tone delay upon line seizure (off-hook condition) of 0.5 s in a 100-port PABX or concentrator. The worst case dial-tone delay would be one second, and that would occur in about 1 percent of the calls.

[6] Universal synchronous/asynchronous receiver/transmitter chip.

9. In other 8085 peripherals, a number of the functions mentioned above are grouped so that a three-chip set can provide essentially all the computer functions, as follows—

—8085 CPU;
—8155 2048-bit static MOS RAM with I/O ports and timer;
—8355 16 384-bit ROM with I/O.

All the CPU and peripheral chips described above comprise the common control system for a PBX (see Figure 10). These components are available separately for custom computer designs, or as a completely assembled and tested single board computer, such as the SBC 80/20 or other standard SBC versions. The SBC 80/20 is a single PC board which measures approximately 6¾″ x 12″. Additional memory RAM cards and a special mounting frame assembly are also available as production items (see References 6 and 7). The SBC 80/20 with its programmable peripherals is suggested primarily for ease of system development.

DISTRIBUTED SWITCHING SYSTEMS

Figure 11 shows a block diagram of a modular digital switching system. Various sized line groups are shown, where each line group represents one D_X-D_R highway pair and associated line circuits. When a line group is colocated with the central digital interchange the D_X and D_R highways may be in phase

The Equipment and Techniques of Digital Telephony

with the PABX system clock, so that the digital interchange can function synchronously with the line group transmission highways.

When the line groups are remotely located with respect to a central digital interchange as in a large PABX, or in a distributed central office architecture, it becomes necessary to adjust the phase of digital bit-streams received over the Dx highways from the line groups so that they are put into phase with the digital interchange timing base. A digital transmission line controller (shown in Figure 11 as DTLC) provides phase adjustment and dc line feed for span-time repeaters (if required) to extend the lines over several miles from the central controller.

New telephone central office switching system designs are generally either of a centralized architecture or a distributed architecture with remote concentrators that serve as modular line groups. With digital transmission rapidly coming into practice, telephone lines can now terminate in codec equipped line circuits located in a remote concentrator. Voice transmission signals are coded and decoded by the codec in each line circuit such that several telephone lines may share common digital transmission lines between the concentrator and the central switching equipment. Each line group may be microcomputer controlled, as could be the central digital interchange. In small systems, or in PABX's, two transmission line pairs (one pair for each direction of transmission) would be appropriate for each line group. In larger telephone system applications, multiple highways in each direction of transmission are frequently required for system reliability and improved traffic handling capacity.

In a distributed switching system, call completion capability between stations served by the same remote concentrator is frequently required when the line group is large, say over 24 lines, and/or is a great distance from the serving central office, say, 10 miles or so. With this approach, the remote concentrator takes on some of the internal switching complexity of a PABX. Outgoing dial signals may be monitored, and if a call is for another party served by the same concentrator, the transmission path to the central office is dropped out, and the calling path is completed locally within the remote concentrator. Alternately, dialing may be detected centrally, and instructions for a local-to-local connection can be transmitted to the remote line unit.

Remote line units may sometimes serve 24 or 48 lines using T1 or T1C transmission lines operating at 1.544 or 3.152 Mbps, respectively, and no concentration at the remote location. If switching concentration is used, then up to approximately 100 or 200 lines or more can be served by a single T1 or T1C span line, respectively, between the central office and remote concentrator. The number of lines is actually a function of the grade of service, which is also a function of frequency of service blockage and service delays allowable in a given application.

Using either μ-law or A-law codecs, up to 30 or 32 channels can be provided over a 2.048 Mbps digital span-line, depending on the signaling method between the concentrator and the serving central office. This provides a better grade of service for a given number of lines, or more lines can be served with the same grade of service, than when a 24 channel T1 compatible approach is used. The 2.048 Mbps bit-rate is therefore beginning to appear in North America on digital lines serving remote concentrators as satellites of a digital central office, even though the 2.048 Mbps bit-rate was heretofore primarily a European standard.

SUMMARY OF TRENDS IN TELEPHONE SWITCHING SYSTEM DESIGNS

Industry trends indicated by many sources, including trade journal articles, seminars, and private discussions, show the rapid adoption of digital techniques for voice and data transmission and switching in systems now in design or being introduced into field applications. The North American T1, 24-channel digital transmission system is becoming a standard, not only for interoffice telephone trunks (per AT&T's D3 channel bank specification) but is also appearing within switching systems on the internal transmission highways. The European CCITT standard 32-channel digital transmission system is also being adopted for PCM switching system internal digital highways in Europe and also North America because of its greater traffic handling capacity. Also, 48-channel highways are now appearing in some switching systems.

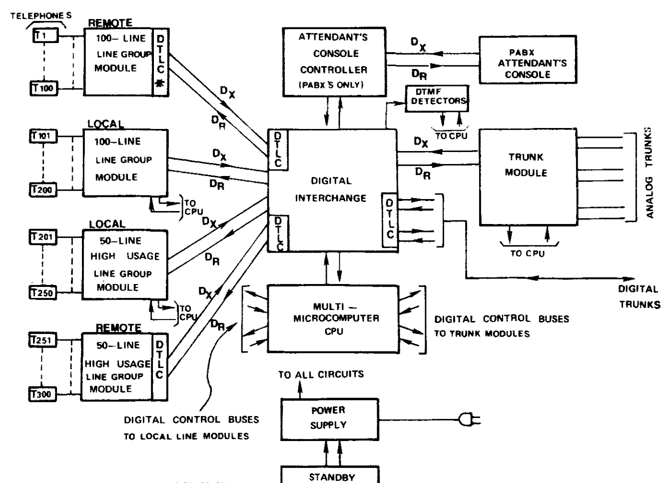


Figure 11. Block diagram of a modular digital switching system

The Equipment and Techniques of Digital Telephony

Single serial transmit and receive highways per line group appear to be primarily used in small PABX's or small line groups while multi-highway approaches are used for reliability and improved traffic handling efficiency plus diversity of span-line (digital transmission line) routing in the case of distributed central office approaches.

LSI versions of a codec and channel filters in each line interface circuit, trunk circuit, attendant's console control, etc. will lead to standardization of components, companding characteristic compatibility, and uniformity of encoding and decoding processes. It also simplifies line circuit modularization and helps reduce the price of the codec-filter set as a function of total market for a given device.

The plan for digitizing voice circuits as close to the customers telephone as is economically feasible is now being implemented. Having a codec at the first line interface with switching or transmission facilities hastens the evolutionary removal of subsequent A/D and D/A conversions and their associated hybrid requirements at junctions between switching systems and carrier transmission systems. This can result in the integration of switching system multiplexed highways with channel bank multiplexed span-lines such that conventional channel bank requirements can be substantially reduced, and interoffice trunks can originate and terminate in span-line compatible digital formats. This approach will eliminate analog trunk requirements for the most part in the long run.

A long range goal of the telephone industry in the United States is digital transmission, telephone-to-telephone, for domestic calls. The introduction of LSI devices such as the codec, PCM channel filters, microcomputers and their various peripherals all speed the evolution of the "all-digital" system goal.

DIGITAL CARRIER CHANNEL BANKS

Many digital carrier channel banks and digital switching systems have the requirement for codecs, filters, and 2/4 wire hybrids in common. In channel banks, the time-slot assignment for a given codec is permanently fixed. In some digital switching systems however, it is desirable to change the time-slot assignments for a given codec from one call to another, and sometimes during a given call set-up process.

Codecs that are designed for use in either a direct (fixed) time-slot mode, or in a programmable (variable) time-slot mode of operation may be conveniently used in both channel bank and switching system applications.

Furthermore, the transmit and receive sections of codecs must be able to operate asynchronously with respect to each other, a requirement for many channel bank applications. A power down mode is required for codecs and their companion PCM channel filters, so that they can be operated in a low-current drain mode during idle circuit conditions.

Channel banks of from two to 48 channels can be produced using per channel serial highway codecs. Either one or two signaling channels per direction can be provided by each codec. Signaling for only one signaling channel occupies the eighth bit of a transsmission time-slot in every sixth transmission frame. Two signaling channels can be provided by alternate sixth frame use of the eighth bit of a transmission time-slot. Thus, two signaling channels (A and B) can be provided for special signaling applications while still meeting the standard D3 channel bank transmission specifications.

For CCITT 30-channel carrier systems, the eighth bit in each time-slot is not used for signaling purposes. Instead, two time-slots per frame are used to carry multiplexed signaling and synchronization data for all 30 voice channels. This is a form of common channel signaling, where signaling for all voice channels is carried by wide-band data channels between central offices.

MICROCOMPUTERS IN HIGH FEATURE ELECTRONIC TELEPHONES

Prior to the anticipated widespread introduction of fully digital telephone instruments that connect with their local switching office by a digital line, a number of new and unusual telephone features can be implemented by including microcomputer control within the telephone set itself. That is, a high feature telephone can be implemented today with readily available off-the-shelf devices and yet connect to a conventional telephone line.

Figure 12 shows the block diagram of a hypothetical electronic telephone instrument (ETI). This product is based on an inexpensive CPU such as an 8084. In the example shown, the telephone could provide the following features:

1. handset operation;

2. Touch-Tone dial or touch-to-rotary dial pulse conversion;

3. tone ringing with variable amplitude and frequency controls;

4. digital display of:
   a) time-of-day
   b) lapsed time of call
   c) number being called
   d) last number dialled
   e) number recalled from repertory memory
   f) identification of other ETI party;

The Equipment and Techniques of Digital Telephony

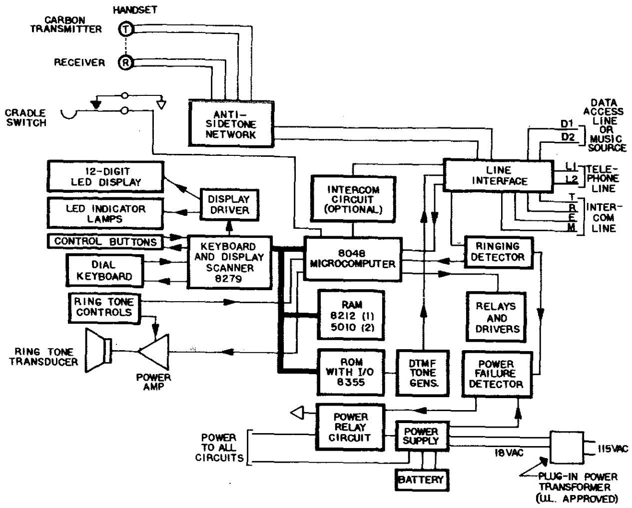


Figure 12. Electronic telephone with intercom and clock (based on 8048 microcomputer)

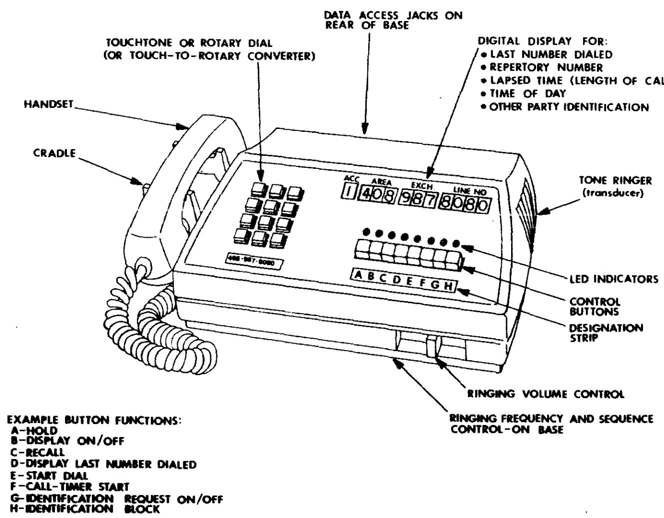


Figure 13. Electronic telephone

5. time of day set by dial button pad;

6. internal line hold and lamp flashing circuitry;

7. repertory dialing with storage and number access via dial pad;

8. redial last number dialled, using Redial button;

9. start-dial control (start button pressed after dial-tone heard);

10. programmed pause after access code (automatic dialling pauses until second dial-tone heard);

11. emergency number programmable for single button automatic dialing;

12. LED lamps for indicating what functions are active at a given time;

13. intercom access by dial pad button;

14. programmable and erasable repertory memory for one or two digital repertory number access codes;

15. Optional data coupler or music-on-hold features.

This instrument can be self-contained and powered by a small external power transformer. An internal long-life battery can provide memory power during commercial power failures and can be automatically recharged when commercial power is on.

An 8279 programmable keyboard/display interface chip can be used to interface between the CPU and the lamps, digital display, control buttons, and dial pad buttons.

A conventional transformer type antisidetone transmission network, or an electronic equivalent can be used for the voice circuitry. Ringing tones can be CPU generated or oscillator generated, with a multi-frequency tone that may be user changed by a control. This same control can also adjust the tone keying rate to provide a wide variety of distinctive ringing tones. A convenient volume control permits user control of the ringing signal loudness.

Lamp flashing, ringing signal tone generation, digital display controls and repertory dialing functions can all be provided by standard readily available LSI devices, such as those shown in Figure 12. Automatic display of other party number requires additional signal tone generators and detectors not shown. This feature would function only between two similarly equipped instruments.

The general appearance of a self-contained high feature telephone, such as has been described above, is shown in Figure 13.

DIGITAL TELEPHONES OF THE FUTURE

Digital encoding and decoding of voice frequency (VF) signals within the telephone instrument are a long-range goal of telephone system design engineers. The ultimate elimination of 2-wire to 4-wire line conversions (and vice versa) at trunk interfaces, and the possibility of having the equivalent of end-to-end four-wire (i.e., separate transmit and receive) transmission paths, could help eliminate several bothersome aspects of telephone transmission engineering, namely 4-wire terminating sets, with their related network balancing problems. The present antisidetone circuits which cancel local talker voice signals from his telephone receiver circuit, could also be eliminated in an equivalent 4-wire telephone. Future optical fibre transmission on customer lines may also help bring about the equivalent 4-wire digital telephone set.

The Equipment and Techniques of Digital Telephony

Both end-to-end digital transmission and common channel interoffice signaling (CCIS) systems can bring improved performance for the telephones of the future. With high-speed common channel interoffice signaling capabilities on a national basis, such new features as displays of the number of a calling telephone at a called station, or storage of numbers of callers to your telephone in your absence, or selective answering of desired calls only, and many other new features may soon become technically feasible, but much remains to be done before the all-digital network becomes a reality.

Once the high-voltage ringing schemes that have been in use since the early days of telephony can be phased out, the chances of using inexpensive, small, LSI components in all-electronic, or all-digital telephones can be hastened. In the meantime, much can be done to apply LSI to the most highly repetitive elements of the existing telephone network, such as telephone instruments, line and trunk interface circuits, digital line controllers, digital interchanges, span-line repeaters, etc.

CONCLUSIONS

Once the analog voice signals are digitized in a telephone system, the processing and control functions can be performed by digital hardware as used in common by manufacturers of computers, data systems, telephone switching systems, transmission systems, and many other types of digital systems. This mass application of commonly used components will help improve system economies and thus benefit both the equipment operators and the end users by providing improved services at reasonable costs.

We have reviewed but a few of the many possible applications of microcomputers, memory devices, and other computer peripherals including codecs, and channel filters to the telephone field. The transition to an all-digital mode of operation in telephony is rapidly increasing as the various key elements needed for digital switching and transmission systems become available.

ACKNOWLEDGMENT

Computer programming for laboratory verification of the PABX techniques described herein was conducted by W.H. Li and M.A. Townsend, and for the electronic telephone study by W.H. Li. Also, periodic discussions with Dr. M.E. Hoff during the study of the systems described in this report were helpful and appreciated.

REFERENCES

[1]S.G. Pitroda, "Switching systems—Section 5," *Telecommun.* vol. 11, no. 13, pp. 75-86, July 1977.
[2]D.K. Melvin, "A fully electronic automatic telephone exchange using time-division-multiplex technique," *General Telephone Tech. J.*, vol. 6, no. 4, Apr. 1959.
[3]W.A. Depp and M.A. Townsend, "Electronic-PBX telephone switching system (ESS-101)," *IEEE Trans. Commun.*, vol. COM-83, pp. 329-331, July 1964.
[4]Intel Corporation, "Data Catalog, 1977," sections 8, 9, and 10.
[5]International Telecommunications Union (C.C.I.T.T.) Fifth Plenary Assembly, Green Book, Volume III-2, "Line Transmission," Published in Geneva, Switzerland, 1973.
[6]G. Adams, T. Morgan, and J. Zarrella, Intel Corp. "A microcomputer tailored for multiprocessor control applications" *Control Eng.*, AR-28, Sept. 1976.
[7]Intel Corporation, "OEM Computers—SBC 80/20 single board computer hardware reference manual," published in 1976.
[8]J. Huggins, M. Hoff, and B. Waren, "A single chip NMOS PCM voice codec," in *Int. Solid-State Circuits Conf. Dig.*, Feb. 1978.

RELATED READINGS

A.R. Meier, "Class 5 digital switching—Where does it stand today? *Telephony*, p. 25, July 11, 1977.
M.J. Kelly, "Study of a rural digital end office," *GTE-Automatic Electric J.*, p. 294, May 1977.
R.C. Garavalia, "New digital electronic PABX family," *GTE-Automatic Electric J.*, p. 303, May, 1977.
R. Gundlach, "Large-scale integration is ready to answer the call of telecommunications," *Electronics*, p. 93, Apr. 28, 1977.
D.S. Rendall and S.L. Soni, "A digital world: The way of the future," *Telephony*, p. 46, July 19, 1976.
K. Gueldenfennig, "Conceptual systems approach for a digital PABX," a paper published by Stromberg-Carlson Corporation, Rochester, NY, June 1977.
S.L. Russell, "System Century® DBX real-time processing," A paper published by Stromberg-Carlson Corporation, Rochester, NY, June, 1977.
H.E. Vaughan, "An introduction to No. 4 ESS," presented at Int. Switching Symp., June, 1972.
J.D. Beirle, "Special facilities in Metaconta 10C large toll application," presented at Int. Switching Symp., June 1972.
C. Rozmaryn, J. Trelut, and L. Viellevoye, "The E11 switching system" presented at Int. Switching Symp., Sept. 1974.
E.B. Daly and P.A. Zelinski, "No. 1 EAX description and performance statistics," presented at Int. Switching Symp., Sept. 1974.
B.H. Shanta, G.B. Meemamsi, C.K. Sane, and V. Kumar, "SPC-1 switching system-general description," presented at Int. Switching Symp., Oct. 1974.
M.J. Kelly and A.W. Kobylar, "Operation of a time division multiplexed PCM switching martix," *Automatic Electric Tech. J.*, vol. 13, pp. 308-314, Apr. 1973.
R.E. Pickett and N.J. Skaperda "A PCM switching system for all uses," *Telephony*, p. 36, July 19, 1976.
J.C. McDonald and J.R. Baichtal, "A new integrated digital switching system" presented at Nat. Telecommunications Conf., Nov. 1976.

The Equipment and Techniques of Digital Telephony

M. Hashimoto, "Development of a computer controlled digital EPABX," presented at Int. Switching Sym., Oct. 1976.

S.G. Pitroda, W.A. Fechalos, and C.J. Stehman, "The microprocessor controlled 580 digital switching system," presented at Int. Switching Sym., Kyoto, Japan, Oct. 1976.

L.N. Bauman, E.J. Braun, D.G. Hill, J.C. Moran, and D.C. Trimble, "The Dimension PBX System and hardware," presented at Nat. Telecommunications Conf., 1975.

H.E. Theloosen, "Time division switching, PAM and PCM in an EPABX," *Telephony,* Nov. 4, 1974.□

Current Transmission Facility Characteristics

Problem:

There are a few immutable constants in human affairs. Death and taxes certainly qualify, and war is a strong candidate. Now we propose two more for consideration: computers and communications. Each constant is based on something we can't avoid, like a biological law or a quirk of human nature, to make it stick. For computers and communications, the "can't avoids" are becoming the dominant institutions of IBM and AT&T.

Before 1968 (the Carterfone Decision), it was impossible to consider a large-scale communications system (the military excepted) without accounting for AT&T equipment or facilities. After 1968, it became possible to consider non-AT&T equipment and facilities, but it is still impossible to completely exclude AT&T from any communications planning scenario. In spite of SBS, XTEN, MCI, and an alphabetic potpourri of dozens of other would-be communications competitors, the hard realities of the communications business lie in what AT&T has to offer and how much they want for it.

This report is a survey of the basic characteristics of the currently predominant transmission facilities, which are almost exclusively provided by AT&T, and of several of the just-starting or upcoming facilities. Inevitably, and almost inescapably, the most significant upcoming facility is AT&T's Advanced Communications Service (ACS). Communications alternatives and costs are explored more thoroughly in other reports throughout this segment.

Solution:

There are two basic forms of transmission in use today—analog and digital. Analog is dominant. Digital transmission is a thing of the mid-70's. All data communication bit-streams must be converted from their computer or terminal forms (i.e., currents, voltages) into a form suitable for sending over common carrier transmission facilities. In almost all cases the conversion has been performed by the ever-present data set (modem), which converts the bit-streams into tone signals. The exception has been the low-speed teletypewriters, which operate without modem conversion on a local loop from the serving central office. At that office, however, the chances are that the signals are converted to tones for further transmission.

The reason for converting the bit-streams to tones is that the common carrier facilities were designed to

Current Transmission Facility Characteristics

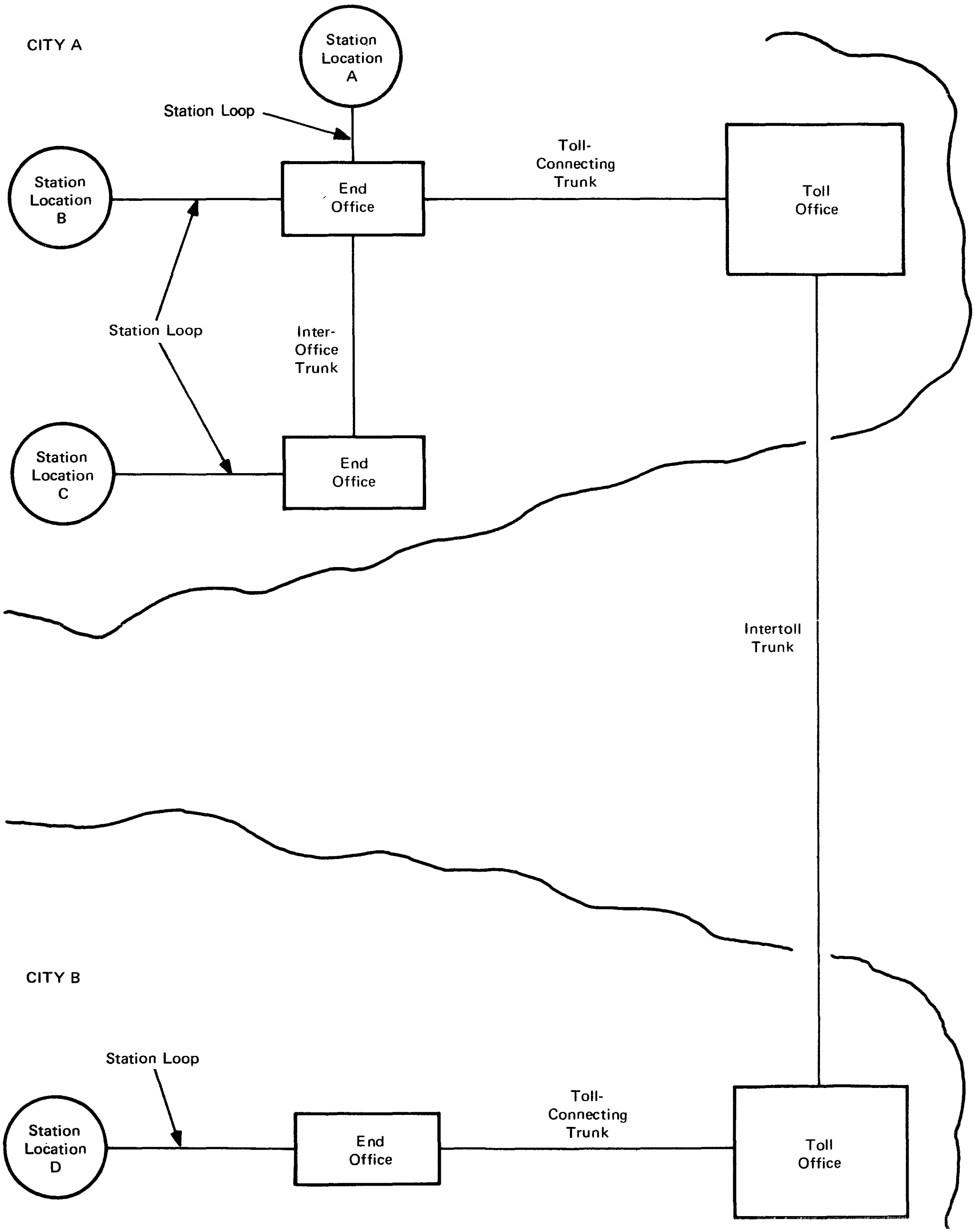


Figure 1: Typical routing method for connections on the telephone network

Current Transmission Facility Characteristics

carry tones—tones of the human voice, but nevertheless tones. Because the facilities were so designed and implemented, the on-off nature (two value or binary) of bit streams could not be carried without first being converted into tones. By way of example, a widely used technique is to transmit all binary ones as a tone of one frequency, and all binary zeroes as a tone of another frequency. At the destination the receiving portion of the modem is able to differentiate between the tones and reconstruct them into a serial bit-stream. Such tone transmission systems are referred to as analog because they were designed to handle signal intensities and frequencies that are electrical analogs of the intensities and frequencies of the human voice.

In the second half of the 60's, the Bell System began to install a new kind of transmission system. Referred to by Bell as T-Carrier, this system converts voice analog signals into digital codes, transmits the digital signal, and converts from digital code back to an analog signal at the distant end. While designed and planned for all communication needs, digital transmission is able to carry data transmission bit-streams directly, that is, without first converting them to tones. While the modem is eliminated, there is still an interface box between the computer or terminal and the service. This unit serves to perform certain voltage level and logical polarity conversions between the business machine and the network service and to isolate the two sides of the link from each other for protective purposes. Digital transmission has emerged in one of AT&T's service offerings, Dataphone Digital Service, and is also the basis of SPC's Datadial I and Datadial II services. (SPC obtained these two services from Datran after it went bankrupt.) AT&T's Dataphone Switched Digital Service (DSDS) is also based on digital transmission principles.

The principal advantage of digital transmission is that it is less sensitive to error-inducing electrical noise. Noise is latent in all electronic systems and can also be induced in a circuit from a multitude of sources, including atmospheric events and high power levels in adjacent circuits or equipment.

In analog systems, the circuit signal must be periodically amplified because it weakens with distance. The amplification unfortunately also amplifies any previously induced noise (and may add noise), making it harder for a receiver to differentiate between information and noise. Analog signals can also be distorted, which interferes with retrieval of information. Noise adds energy to a circuit; distortion rearranges the energy distribution. Digital signals are periodically reshaped or regenerated enroute by devices known as repeaters. Passing through successive repeaters can cause a form of distortion known as phase jitter (causes signal elements to be of uneven length). However, digital transmission still yields overall error rate performance superior to that of analog.

In the following paragraphs, the basic characteristics of the most commonly used facilities for data transmission are presented as follows:

- The public telephone network or Bell's DDD.
- Value Added Networks or Packet Switching.
- Bell's Advanced Communications Service (ACS).
- Private or leased line services.
- Wideband analog services.
- Bell's Dataphone Digital Service (DDS).
- Satellite channels.

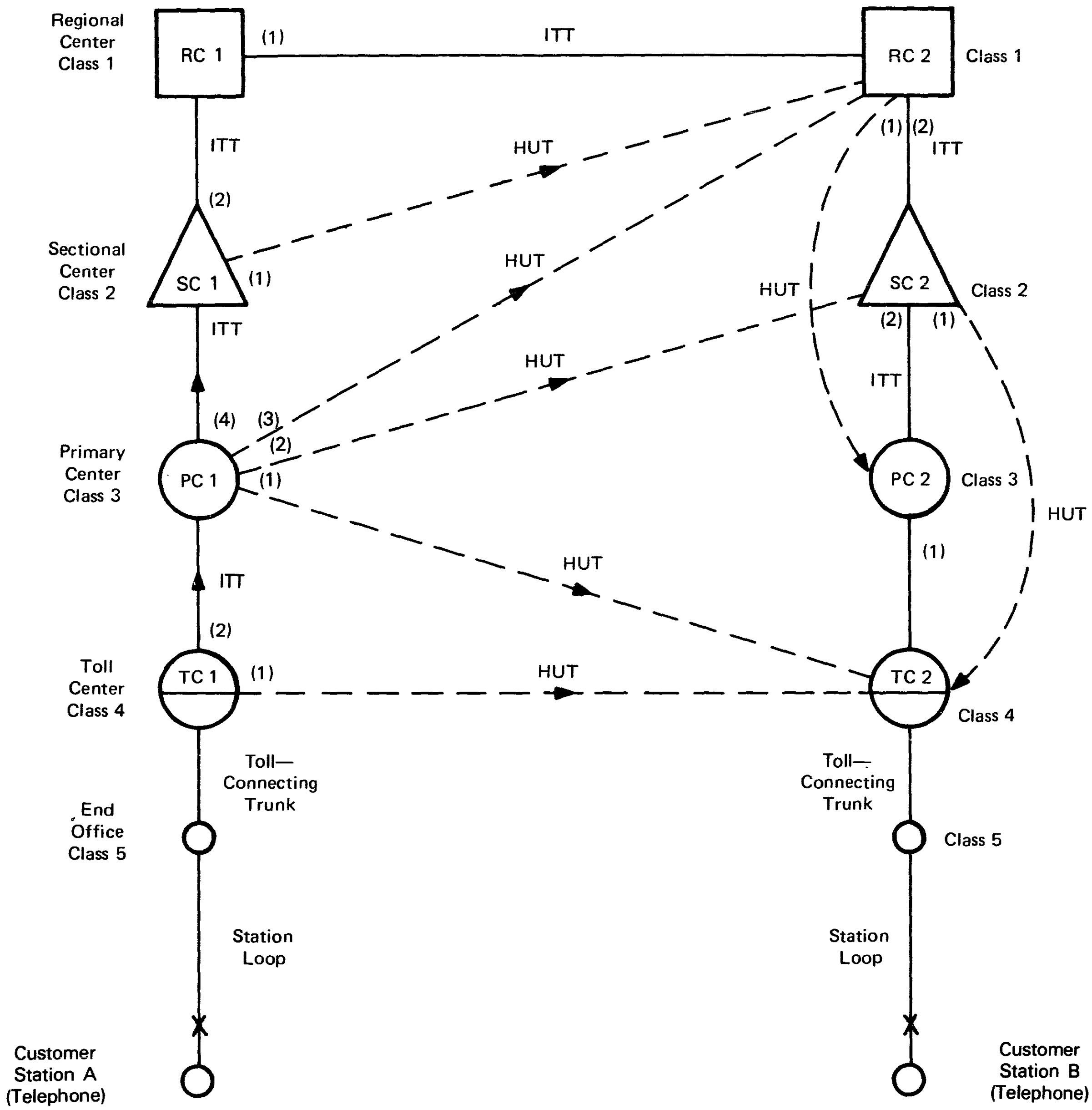
BELL'S SWITCHED OR DDD SERVICE

Figure 1 illustrates some of the elements in Bell's Direct Distance Dialing (DDD) network. Subscriber stations (telephones) are connected via a station loop to their serving central office. The latter is referred to as an end office or local control office, meaning the end or last office in the switching hierarchy. If, as is shown in the figure, station A wants to talk to station B, the entire call is handled within the same end office. For A to talk with C, which is connected to a different but nearby end office, an inter-office trunk is used. When different cities are involved, such as with stations A and D, the call from A is passed from A's end office via a toll-connecting trunk to the toll office serving A. From there the call is routed to the toll office serving D via an intertoll trunk. Then the call is forwarded to D via his end office.

The intertoll trunking forms an elaborate and extensive nation-wide switching hierarchy, shown schematically in Figure 2. At the bottom we see the loops, end offices, and toll connecting trunks of Figure 1. The toll offices and intertoll trunk, however, have been replaced by four levels of switching and many intertoll trunks. The call from A to B will proceed up the hierarchy until it can find a path towards B's toll center. Notice that there are shunts called HUT's (high usage trunks). They are established where large inter-office traffic flows warrant them. They expedite call completion and relieve network congestion.

The many possible variations in call routing have been a speed-limiting factor in the use of the DDD network for data transmission. Where a private line

Current Transmission Facility Characteristics



1. Numbers in ( ) indicate order of choice of route at each center for calls originating at Station A.
2. Dashed lines indicated high-usage paths.
3. ITT—Intertoll Trunk.
4. HUT—High Usage Intertoll Trunk.

Figure 2: Choice of routes for an assumed call on the telephone network

channel has been established between two points, the path is the same from day to day, and the path transmission characteristics are essentially the same (or usually change slowly). Data set technology can rely on this and maintain a higher bit rate than could be obtained if the channel characteristics changed from day to day. On the DDD network, the path changes from call to call with variations in circuit characteristics. Different through-network delays may also occur according to the path established for the particular call. Interfacing between successive trunks, varying points of amplification, and other factors also add to variations between calls.

Modems must be capable of operating over a typical or average path, which will generally limit operation to a lower bit rate than private line channels. Nevertheless, transmission at 4800 bps is for many users a standard for day-to-day operations on the switched network. The Bell System itself offers a modem, the Model 208B, for use on the DDD network at 4800 bps.

Current Transmission Facility Characteristics

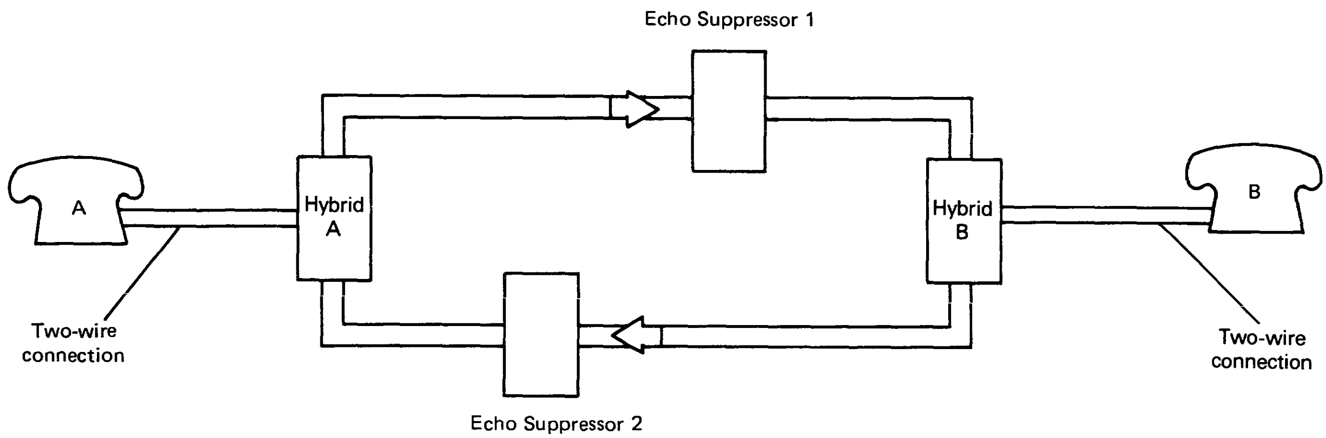


Figure 3: Schematic of a telephone network connection

A second limiting factor of the DDD network is that it is constrained to half-duplex operations for bit rates of 2000 bps and up. Figure 3 is a schematic representation of a connection through the DDD network. The local loop is two-wire but this is interfaced to a four-wire transmission path via a device called a hybrid. While the transmission path could clearly carry data both ways at once, the two different signals could not survive the hybrids and local loops. This applies to present modem technology at 2000 bps and up. Lower speed modems such as the Bell 100 series commonly operate full-duplex over the switched network. A number of vendors, including Vadic and General DataComm, offer a full-duplex modem that operates at 1200 bps over a two-wire circuit.

However, it has not been so much the half-duplex aspect of the DDD network that has limited its use for data, as it has been the time required to reverse transmission directions. Historically, most data transmission has been half-duplex. A typical procedure involves sending a block of data, awaiting the acknowledgement of the block, and sending the next block. In the DDD network the time to reverse transmission direction or "line turnaround" may be as much as 0.2 second. Two required turnarounds per block yield 0.4 seconds of delay for each block transmission. One could send 100 characters in this time using a 2000 bps Model 201A modem. If you are sending one 80-character record per block, the transmitting efficiency is in the range of only 44 percent. The long turnaround is mostly due to the time required to turn off the echo suppressors shown in Figure 3. When the data flow is from A to B, echo suppressor 2 acts to block any of the A to B signal that leaks through the hybrid at B. When B is to transmit back to A, however, echo suppressor 2 must be turned off to allow normal signal flow from B to A.

Yet another drawback for use of the DDD network for data is the call establishment delay. The first component of this is dialing. The delay between pickup and getting the dial tone ranges from 100 milliseconds to 500 milliseconds. Thereafter, automatic calling units can dial at the rate of 10 pulses per second plus a 600 millisecond pause between digits for a rotary dial line. For a tone-dialed line (Touch Tone), the rate would be 10 digits per second. A number such as 609-764-0100, then, would take 12.7 seconds to pulse dial or 1.0 second to tone dial plus the delay to obtain a dial tone.

Once the call has been dialed, there is a delay while the switching hierarchy completes the connection. According to Bell, the delay is about 11 seconds for 0-180 miles; 16 seconds for 180-275 miles, and 18 seconds for 725-2900 miles. Variance for each group is 5 to 7 seconds. In late 1975, AT&T announced a new signaling structure called CCIS for Common Channel Inter-office Signaling. With CCIS, the information needed to establish and route calls is carried on a separate data link from the call circuit. This allows faster signaling and leads to a reduction in call setup time to about two seconds, according to AT&T. CCIS is AT&T's version of the CCITT signaling system No. 6, an international specification for common-channel signaling.

Finally, the DDD network is usually thought of as being a fountainhead of data errors due to passage of the signal through switching centers and over variable routing paths. In the most recent study (1970) of switched network data performance, the results showed an order of magnitude improvement over the prior (1959) study. The 1970 study suggested that some 95 per cent of 10,000-bit blocks would get through error-free at 2000 bps, and 90 per cent at 3600 bps. Shorter blocks, which are more common, result in better performance. (Nothing should be inferred about performance at 4800 bps, because such modems utilize a different modulation scheme and may perform comparably or better.) An error detection and retransmission scheme can substantially reduce the residual error rate. A polynomially

Current Transmission Facility Characteristics

checked code, such as a cyclical redundancy check, might raise the error-free block rate to something like 99 per cent.

VALUE ADDED NETWORKS

A value added network, or VAN, is a special carrier service based on the Advanced Research Projects Agency communications network (ARPANET), which has formulated to provide shared access among 30 research facilities across the U.S. using a system of computer switches and a concept known as packet switching. This concept established a uniform method of packaging data into segments or packets such that packets can be forwarded over any available path in the network to the desired destination. This method distributes the total loading throughout the circuits within a network.

Essentially, a switched network is formed by leasing lines from other carriers and by installing switching computers at the appropriate nodes. By not restricting the necessary delivery service to a single communications path, as a leased line does, traffic can be balanced across the available facilities, thus increasing their utilization.

The first value added network proposal approved by the FCC (in 1973) was submitted by Packet Communications Inc. (PCI). However, PCI never completed its proposed network and has ceased operations.

Three companies are currently licensed by the FCC to operate a value added network for data communications: Graphnet Systems Inc., Telenet Communications Corp., and Tymnet. In 1977, Graphnet was authorized by the FCC to expand its network, which has been used primarily for the transmission of facsimile, into a general-purpose digital communications service.

The chief advantage that seems to be offered by the value added services is an improvement in error rates along with the potential for reducing communications network development complexity and possibly cost. The computers used in the switching operations can also be used to control error detection and recovery—functions which you have to provide in your own computers now. In general, the approach seems to be for the VAN carriers to lease facilities from conventional carriers and provide the switching computers that tie the network together.

The International Consultative Committee for Telephones and Telegraphs (CCITT), part of the United Nations International Telecommunications Union (ITU), has adopted a recommended interface, X.25, to be used in packet-switching networks worldwide. This standard specifies the interface between data terminal equipment and data circuit termination equipment for devices operating on these networks. Telenet Communications Corp. is the first U.S. packet-switched network to implement this standard. Canada's Datapac network is also using this standard. Public packet networks in England, France, and Japan plan to use X.25. Report CS93-501-101 contains detailed information on the X.25 Packet Switching Interface.

BELL's ADVANCED COMMUNICATIONS SERVICE (ACS)

On July 10, 1978, AT&T filed a petition with the Federal Communications Commission for approval of a new data communications service, designated Advanced Communications Service (ACS), as a common carrier offering. FCC acceptance of this petition would precede the filing of a formal tariff, which is anticipated in mid-1970. The earliest date that this service could be offered to users would be within 90 days after the filing of the tariff.

ACS is described in the petition as a public packet-switched network offering customer access for terminals and computers. The network is to be composed of nodes, which provide processing power to control data traffic and to accommodate a wide range of terminal and computer protocols; tandem switches, which provide an enhanced level of switching; and 56,000-bps digital channels, which connect the nodes.

Two basic classes of transmission alternatives will be offered in ACS: Call and Message. Call provides a bi-directional transmission path between an originating station and a destination station; it is intended for applications that require real-time transactions. Message provides a one-way path; it is intended primarily for data entry and remote batch applications that do not require an immediate data response.

AT&T proposes to initially support five classes of terminals and three modes of host computer access. Terminal and host computer access to the network can be accomplished only through the nodes. Terminal classes supported include asynchronous up to 1800 bps and binary synchronous at up to 9600 bps; contention and polled arrangements will be supported. Logical circuits are established on an immediate or deferred basis, depending on the type of service requested.

Within ACS, a user can establish private subnetworks, with access controlled by passwords. The customer can assign specific terminals for managing the network. Some terminals can be assigned to be Customer Network Information Centers for sched-

Current Transmission Facility Characteristics

uled, on-occurrence, or on-demand reporting of maintenance, configuration, equipment, service, usage, and performance information.

Access to ACS depends on whether the customer's premises are within or outside of an ACS serving area. When the customer premises are within this serving area, access is via a digital or analog access channel. When the customer premises are outside the ACS area, access is accomplished via private line service, via DDS, or via the public telephone network through dial access ports. Dedicated access facilities are designed to connect terminals operating asynchronously at up to 1800 bps or synchronously at 2400, 4800, and 9600 bps. Host computer access will be provided at 56,000 bps; AT&T plans to have up to 100 ACS service areas available when the service is initially offered.

It is too early to evaluate the costs for specific ACS communications applications and to compare them with the costs of the additional processor power, main memory, software, and personnel support required to perform the equivalent tasks entirely within a user's own equipment. Rate structures and charges will determine whether it is more cost-effective to utilize ACS or to use conventional carrier-provided facilities with user-provided network control. Depending on the actual rates, ACS may have something to offer both small and large users.

PRIVATE LINE SERVICE CHARACTERISTICS

Private line services have traditionally been offered in three principal categories. These are referred to as low-speed or narrowband; medium-speed or voice-grade; and wideband or broadband. In analog systems, all such channels are derived using frequency division multiplexing (FDM). The carrier begins with bread-and-butter voice channels and combines 12 of them, via FDM, into a "group." In turn, five such groups are combined into a supergroup, the equivalent of 60 voice-grade channels. Finally, 16 supergroups are combined into a line group, representing 960 voice-grade channels. Such groupings, expressed in bandwidth, comprise wideband offerings.

Low-speed offerings are obtained by using FDM to divide the voice channel into smaller channels. In a manner of speaking, a voice channel can be thought of as a group of narrowband channels. The narrowband services—Series 1000 in AT&T parlance and Low Speed Service in WU parlance—are intended for the low bit rates of keyboard/printer terminals. The older Baudot code teletypewriters operate at 45.5, 50, and 75 bps, while the newer ASCII keyboard printer terminals operate at 110, 134.5, 150 and 300 bps. The 150 bps channel is presently the fastest low-speed service offered by most carriers. An exception is Southern Pacific Communications Company, which offers a 300 bps service, suitable for 30 cps keyboard/printer terminals.

Narrowband channels from AT&T and WU are offered for half-duplex service or for full-duplex service (usually at a 10 per cent premium). The half-duplex service is a four-wire circuit, however, permitting the quick turnaround traditionally, associated with full-duplex service. The 10 per cent premium is only incurred if the application involves true simultaneous data transmission both ways. The specialized carriers (SCC's) make no rate distinction between full- and half-duplex use.

There has been a steady trend away from the use of narrowband services for data communications. The services serve teletypewriter-based message systems, and that continues to be their main application. Since the carrier's main unit of capacity is the voice channel, it is the most cost effective. A voice-grade can actually cost little more than a narrowband line, and sometimes less. When one combines the cost with greater capacity, the voice channel is normally more cost-effective. The greater capacity also translates to faster response-time performance, an important factor in many applications. To exploit the advantage of voice channels, keyboard/printer terminal development led to buffered versions. These could still print at economically low speeds and accept the slow manual input, but their buffers enabled them to send and receive at the higher rates of the voice channel. Hence, the cost-effectiveness of voice channels has resulted in the narrowband channels being little used for data communications.

Private line voice-grade channels are highly used in data communications. They may be arranged for two-point or multipoint service, and are generally used at up to 9600 bps. In today's commercial practice, multipoint systems are generally constrained to 4800 bps. This is because of the training time needed for high-speed modem signaling. In multipoint systems, only one remote terminal may have its modem carrier (transmitting signal) on at a time. If two or more were to be on, the mixture would be undecipherable by the computer. Hence, the usual line control procedure is to have the computer invite or poll each remote terminal in turn to transmit. A terminal recognizes its particular poll (address) and turns on its modem carrier. The computer's modem must now adjust itself or "train" to the terminal's carrier. The remote terminal waits through a fixed delay called the Request-to-Send/Clear-to-Send (RTS/CTS) delay for the computer's modem to train before it begins to send data. For modems at bit rates above 4800 bps, the effect of the training time on response time has thus far been too great to allow

Current Transmission Facility Characteristics

economic multipoint operations. Codex recently introduced an interesting concept: 9600 bps from the computer and 4800 bps from the terminal. In the outbound direction the computer keeps its carrier on constantly, so the RTS/CTS delay is not a problem, and the full 9600 bps can be effectively used. However, for the most part, present limitations restrict 7200 and 9600 bps to two-point applications, typically for remote batch, remote job entry, or data collection.

One interesting variation on the present ceiling rate of 9600 bps is the use of two voice channels to achieve 19.2K bps. Devices are available which will split a 19.2K bps bit stream into two 9600 bps streams and feed each into a separate voice channel. At the remote end the device combines the two signals again into the original 19.2K bps stream.

Communication circuits produce two major characteristic effects on a modem's carrier signal. They attenuate (reduce) the signal amplitude (intensity) as a function of carrier frequency, and they delay the signal as a function of its frequency. Conditioning is a process of changing a circuit's characteristics to the point that the attenuation and delay are within values manageable by the modem. These values will be a function of the modem's modulation and demodulation techniques and its operating bit rate.

The Bell System currently provides three types of conditioning: B conditioning, C conditioning and D conditioning. The recently announced B type is for 3002 channels up to 5 miles in length and only when limited distance modems are used. It is available in two types; B1 is for channels operating at 2400 or 4800 bps; B2 is for channels operating at 9600 bps. Type C conditioning establishes delay distortion and attenuation limits to the line. Types C1, C2, and C4 are the most common C conditioning. There are other C grades of conditioning for specialized or military applications, but they are rarely used in commercial practice. Type D conditioning establishes limits on the noise-to-signal ratio and harmonic distortion of the line. There are two types of D conditioning: D1, which pertains to point-to-point lines; and D2, which pertains to two- or three-point channels. Type D conditioning is accomplished primarily by facility selection rather than adding equalizers. Type D conditioning is designed for use with modern high speed modems. These devices automatically equalize the circuits they use, thus duplicating the function of C conditioning. Less sophisticated modems benefit much from C conditioning.

Because of the steady improvement in both network plant and modem technology, conditioning requirements have been much relaxed. For many years, Bell practice called for type C1 even on 2400 bps service. Today, for the newer Bell data sets operating at up to 4800 bps, conditioning is recommended (type C2) for only the 2021 when it is operating at 1800 bps. At 9600 bps, Bell requires type D1 with its sets, and the independents recommend it.

WIDEBAND ANALOG CHANNELS

As noted earlier, wideband channels are implemented via group, super group, or portions of line group transmission facilities. The basic (AT&T) wideband service is designated as Series 8000. This is a two point only service and consists of a channel bandwidth of 48 KiloHertz (KHz), or the equivalent of 12 voice channels. The service is offered for operation at 19.2K, 40.8K, or 50K bps. The 19.2K bps service uses only half of the capability, and the remainder can be used either for a second 19.2K bps service, for up to six voice channels, for PBX tie-lines, or for foreign exchange channels.

The next step up would be to a super group of 60 voice-channel capability. This is commonly referred to as Telpak C, and in AT&T tariff terminology as Series 5000 (Type 5700).

The full bandwidth can carry 230.4K bps. Alternately, the capacity can be organized into units equivalent to 6 voice channels for 19.2K bps, or 12 voice channels for either 40.8K or 50K bps. The unused capacity can be used as voice channels or narrowband channels (at a rate of two per voice channel).

At the line group level, AT&T offers Type 5800 channels (Telpak D) with an equivalent capacity of 240 voice channels, or one-fourth of a line group. This offering is simply more total capacity without any higher per channel offering than the 230.4K bps cited for Telpak C.

DATAPHONE DIGITAL SERVICE

Dataphone Digital Service (DDS) is an all digital transmission service introduced by Bell in 1974. Already available in 28 cities, AT&T is planning to extend it to 96 cities. As discussed previously, digital transmission is less error prone and Bell has backed that with a guarantee of average performance exceeding 99.5 per cent error free seconds. This is independent of speed of service, which may be either 2.4K, 4.8K, 9.6K, 56K, or 1544K bps. The service is available on a two-point or multipoint basis. The actual availability of service in a particular city is limited to certain central offices, all of which are enumerated in Bell's DDS tariff 267.

While DDS has obviated the modem, its physical presence has been replaced with a unit called the Data Service Unit (DSU). This unit serves to interface the

Current Transmission Facility Characteristics

transmission service to the business machine as did the modem, but it is substantially less complex. As with modems, the DSU can be customer supplied if so desired.

DDS channels may be interconnected with Bell analog channels. A terminal in Macon, Georgia, for example, could be routed via an analog voice channel to Atlanta and enter the DDS network there. At present such a connection would have to be made from a customer's location. The possibility of such "off-net hubbing" being performed at Bell facilities has been rumored, but is not tariffed at this writing.

The absence of modems has also suggested elimination of the RTS/CTS delay, thus improving response performance. In fact, there is a modest RTS/CTS, ranging from 8 milliseconds at 2.4K bps to 0.4 milliseconds at 56K bps. More important, however, is that DDS introduces another delay not encountered in analog systems. As with analog transmissions, DDS is implemented via a hierarchy of multiplexing. In DDS, however, the hierarchy is one of time division multiplexing rather than frequency division. In a TDM hierarchy, bits may wait for short intervals until their assigned time slot occurs in the next level of the hierarchy. The effect of this waiting is to cause delays comparable in size to the RTS/CTS delay. Estimated polling response delays, from issuance of a poll to receipt of a response, range from about 30 to 90 milliseconds at 2.4K bps, and from 12 to 70 milliseconds at 9.6K bps. Actual values depend on the mileages between the stations. The unfortunate aspect of these delays is that they occur on every transmission to and from a terminal, including those after polling. The RTS/CTS delay on analog systems applies only to the first response in any series from the polled terminal.

SATELLITE SERVICES

Satellite channels are offered for two-point service in units of voice-grade channels and multiples of these channels. They have thus far found limited use for data communications because of three factors: their essentially two-point character; their limitation to a handful of terminal cities; and their intrinsic, long propagation delay.

Their limitation to two-point service and availability in a few cities are interrelated. The technology exists to provide what would amount to a multipoint service. With the few terminal cities, however, it is not particularly pertinent. Nor are the satellite carriers anxious to become involved with the costs and problems of such services. This is because the existing voice service market (tie-lines, etc.) is much larger and much easier to sell and serve.

A dominant characteristic of satellite data transmission is propagation delay. To achieve a geostationary orbit (always above the same point), the satellites are positioned some 22,300 miles from earth. For a signal to propagate (travel) from an earth station to the satellite and down to another earth station requires about a quarter of a second. Traditional data communications line control procedures would require two such delays per transmitted block of data (one for the block, and one for its acknowledgment) or about a half second of wasted time per block. This would translate into about 33 percent efficiency for unit record size (80 characters) blocks at 2400 bps. Such performance is comparable to that provided by DDD service with the first generation type 201 modems. The error rate performance of the satellite channel would be quite superior, however.

To properly exploit satellite channels for data communications, a different line control procedure is required. Instead of the traditional send-and-wait procedure, a "go-back-N" procedure is required. This means the transmitter continues to send data without pause. If the receiver detects a block with an error, it notifies the transmitter to "back up" and resend that block. This means the transmitter must be able to continually store the last "N" blocks it has sent, where N is dependent on the system parameters. Codex Corporation has developed equipment, which, when inserted into a send-and-wait link, makes it operate as a go-back-N link; American Satellite Corporation has also developed this type of equipment and offers it with its satellite service.

The new high-level data link control procedures (ISO's HDLC, ANSI's ADCCP, IBM's SDLC, etc.) all use go-back-N procedures. Consequently, their appearance in systems will make satellite channels much more effective for data communications.

In the summer of 1977, RCA American Communications successfully demonstrated the use of satellite communications facilities to distribute new wire services. The demonstration involved the transmission of wire and news picture services from an RCA earth station located on the east coast to the company's Satcom I satellite, 22,300 miles above the equator in a geostationary orbit, and finally to a recieve-only earth station located in California. This successful demonstration may hasten the use of satellite facilities for the distribution of news services to many newspapers, each with a receive-only antenna on its roof, located throughout the country.

Also in 1977, a joint venture among IBM, COMSAT, and Aetna Life and Casualty produced a hybrid company called Satellite Business Systems (SBS). SBS shows enormous promise as a 'dumb' carrier for transmission among intelligent nodes (just the re-

Current Transmission Facility Characteristics

Carrier	Switched services			Private line services		
	Low-speed	Voice-grade	Wideband	Low-speed	Voice-grade	Wideband
merican Satellite	—	—	—	Satellite channels	Satellite channels	Satellite channels
T&T	—	DDD, WATS	Dataphone-50, DSDS	Series 1000	Series 2000, Series 3000, DDS	Series 5000, Series 8000, DDS
raphnet	—	VAN	VAN	—	—	—
ICI	—	Measured usage	—	—	Leased or measured usage	Leased or measured usage
CA	International telex	—	—	Satellite channels	Satellite channels	Satellite channels
PCC	—	Datadial, Sprint	Datadial, Sprint	Leased or measured usage, satellite channels	Leased or measured usage, satellite channels	Leased or measured usage, satellite channels
elenet	VAN	VAN	VAN	—	—	—
CTS (Canada)	Datapak	Datapak	Datapak	—	—	—
ymnet	—	VAN	VAN	—	—	—
Inited States Transmission ervice (ITT)	VAN	—	—	Leased usage	Leased usage	Leased usage
Vestern Union	TWX, Telex	Broadband Exchange Service	—	Low Speed Service	Series 2000, Series 3000, satellite channels	Series 5000, Series 8000, satellite channels

Other international telex carriers include Western Union International, ITT World, TRT, and French Cable Company.
/AN—Value Added Network or packet switching.

igure 4. Spectrum of carrier services

verse of the ACS approach) but is currently stalled in the regulatory morass of the FCC and the courts.

CARRIER SERVICES

The dimensions of the network services offered by the common carriers, the specialized common carriers, and the satellite carriers present a formidable task of evaluation and selection from the many alternatives which can satisfy a given business situation.

The major facilities available are summarized in Figure 4. Detailed information about selected services is presented in the individual reports throughout this segment.

The service charges of the carriers in the following reports are presented for your convenient reference; they should not be construed as official though every attempt is made by Datapro to keep them current and accurate. The user must obtain the official rate structure and charges from the carriers themselves. The information may be used for general planning purposes.

The master guide to communications costs and available facilities is the tariffs themselves. These documents are available in the business offices of the various carriers. Be sure the set you look at is up to date. Each tariff carries a flyleaf that lists each page and its current revision. This does not tell you, however, whether the latest set of revisions (which includes a revised page list) has been received. If you use this source, expect to spend some time with them, as the tariffs are complex. It is much easier if you are looking for information on a specific service than if you are trying to find all the services that will handle a particular task. The tariffs, in effect, form the contract between the customer and the carrier.□

Users' Views of Transmission Facilities

Problem:

The many problems of communications systems analysis and design can be partially alleviated by talking with other people who have already faced and perhaps solved exactly the same problems. This is one of the real benefits of belonging to a users' group. Unfortunately, we can't set up a dialogue for you in this service with experienced users, but we can give you a consensus of various users' views and opinions through responses to Datapro's surveys solicited from our subscriber lists. This report gives you just such a consensus of a recent user survey on transmission facilities. We hope it will be a useful preplanning guide by placing the many facility offerings into an experienced-user perspective. Refer to Table 1 in this report for a comprehensive summary of users' ratings by facility.

Solution:

To assess the current level of user satisfaction with the various communications links and to develop a feel for the types of facilities in use, the pattern of usage, and problem areas, a Reader Survey Form was included in the January 1979 supplements to both DATAPRO 70 and DATAPRO REPORTS ON DATA COMMUNICATIONS. By the editorial cut-off date of March 27, a total of 326 users had responded. The survey forms were carefully examined to eliminate any duplications resulting from the dual mailing, and the responses were tabulated.

There were three parts to the questionnaire: Network Description (usage patterns), Network Problems, and Facility Sources and Ratings.

USAGE PATTERNS

The 30 questions we asked under Network Description can be grouped into the following categories: network configuration, types of facilities used, transmission speeds, number of lines used, length of lines, and degree of utilization. The answers supplied by the users are summarized here in the same order. The information presented includes the actual number of responses and the corresponding percentage of the total number of replies (326). Also presented for comparison purposes is the corresponding percentage for the equivalent question from the 1978 survey based on the total number of replies in 1978 (413). Not all the questions we asked in 1979 were asked in 1978. Because not all the users who replied answered the questions in this segment of the survey form, the results are a little lower than actual usage among our subscribers, but the differences are probably small.

Network Configuration

	Number of Responses	% of Total Responses	% of Last Year's Responses
Single computer in network	174	53	56
Multiple computers in network	128	39	37
Use programmable front end	163	50	46
Use programmable remote concentrator	25	8	8
Use intelligent multiplexer	34	10	6
Use hard-wired multiplexer	76	23	26
Use multi-point lines	148	45	42
Use point-to-point lines	233	71	63

Users' Views of Transmission Facilities

The majority of the responses to this year's survey fall within the range of normal variation from those obtained in 1978. However, there is a noticeable increase in the use of point-to-point lines, which seems strange at first glance, since multi-point lines are generally less expensive. There is also a significant increase in the use of intelligent multiplexers, and there may be a correlation. Most of the intelligent multiplexers coming into use are of the statistical variety and are equipped with adequate storage facilities and ARQ operating procedures to effect sophisticated error detection and recovery techniques. In the past, many multipoint circuits were used in a polled environment, and this application is sometimes troublesome to statistical multiplexers. Terminals on a polled circuit, particularly those of older vintage, are not equipped to respond to a poll with the terminal address because it is not necessary and contributes overhead in the response message. The polling device (front end, etc.) knows which machine has been polled and consequently can associate the returned message with a certain terminal. Once the statistical multiplexer is introduced into the system, the returning message(s) are stored until an acknowledgement is received from the receiving multiplexer. Without a terminal address in the response message, the polling device cannot identify the responding machine(s) in the event of a retransmission, because a resent frame may have several responses. The inherent error control procedures of statistical multiplexers overshadow this drawback, however, and users may be reconfiguring with point-to-point lines to solve the polling problem and still enjoy the benefits of intelligent multiplexers.

Types of facilities used

	Number of Responses	% of Total Responses	% of Last Year's Responses
Full period private line service	266	82	77
Partial period private line service	13	4	4
Switched private line service	69	21	19
DDD (telephone network)	188	58	57
INWATS	80	25	19
OUTWATS	109	33	32
TWX/Telex	89	27	21
DDS	48	15	8
DSDS	0	0	1
Packet switching service	25	8	5
Line/modem patch/switch centers for network backup reconfiguration	73	22	22
Share business voice and data communications over same line	98	30	28

Except for Bell's beleaguered DSDS offering, which is still in the throes of the regulatory process regarding rates, this year's results show a normal growth pattern with emphasis toward digital and packet switched services. Next year may not show increases quite this dramatic, as users ponder the benefits and outcome of AT&T's Advanced Communications Service (ACS), which could be in the infant stages of operation for the 1980 survey.

Transmission speeds

	Number of Responses	% of Total Responses	% of Last Year's Responses
150 bps and slower	77	24	31
300 to 1800 bps	187	57	58
2000 2400 bps	174	53	55
4800 bps	208	64	60
9600 bps	159	48	38
Over 9600 bps	46	14	10

As has been the case in past surveys, the trend toward higher speeds is repeated again this year. These statistics support the increased usage of digital facilities previously mentioned. Also, the market flood of microprocessor-based modems appears to have placed 9600-bps analog transmission firmly in the limelight.

Number of lines

To gain additional insight into use of communications facilities, we divided the responses into two groups: large users and small users. The arbitrary division point selected in previous years (use of 100 or more communications lines, a network mileage of 20,000 miles or more, or both) was retained for this year's tabulations. The first group of large users consists of all those who had 100 or more lines and answered either the network mileage question, the hour usage question, or both.

	1979 Survey			Last Year's Survey		
	Large Users	Small Users	Total	Large Users	Small Users	Total
Users responding	27	227	254	42	322	364
Lines reported	6,792	3,634	10,426	9,699	3,086	12,785
Average number of lines per user	252	16	41	231	9.6	35

Line length

A second group of large users is represented by those who answered both the questions about number of lines and network size.

	1979 Survey			Last Year's Survey		
	Large Users	Small Users	Total	Large Users	Small Users	Total
Users reporting	19	191	210	31	253	284
Network mileage reported (thousands of miles)	684	897	1,581	1,167	579	1,746
Lines represented	5,410	3,071	8,481	6,943	2,980	9,923
Average line length (miles)	126	292	209	168	194	176
Average user network size (miles)	36,015	4,695	7,528	37,650	2,290	6,150

Users' Views of Transmission Facilities

Carrier and Service	Number of Responses	Planning Assistance					Ease of Implementing Plans					Reliability of Operation					Quickness to Troubleshoot and Correct Problems				
		WA	E	G	F	P	WA	E	G	P	P	WA	E	G	F	P	WA	E	G	F	P
AT&T/Bell System—																					
Leased telegraph-grade lines	27	2.2	2	10	7	8	2.7	4	13	9	1	3.1	3	23	1	0	2.5	3	10	12	2
Leased voice-grade lines	247	2.4	26	87	87	43	2.6	33	105	81	23	2.8	53	156	32	4	2.5	28	107	79	30
Leased wideband lines	47	2.6	9	16	14	7	2.5	7	17	16	7	2.8	11	29	7	0	2.7	7	24	12	4
DDD (telephone network)	154	2.3	13	48	54	33	2.6	18	71	45	14	3.1	42	87	19	4	2.6	25	62	48	16
DDS	8	2.1	0	3	3	2	3.1	2	5	1	0	3.6	5	3	0	0	3.3	3	4	1	0
WATS	4	2.8	1	1	2	0	3.3	1	3	0	0	3.5	2	2	0	0	2.5	0	3	0	1
Subtotals	487	2.3	51	165	167	93	2.6	65	214	152	45	3.1	116	300	59	8	2.6	66	210	152	53
Other telephone companies—																					
Leased telegraph-grade lines	3	2.0	0	1	1	1	2.7	0	2	1	0	3.0	1	1	1	0	2.7	0	2	1	0
Leased voice-grade lines	37	1.8	1	6	13	14	2.0	1	8	16	11	2.5	2	17	14	4	2.0	1	9	16	11
Leased wideband lines	6	2.0	1	0	2	2	2.0	1	0	3	2	2.7	1	2	3	0	1.7	1	0	1	4
DDD (telephone network)	23	1.8	1	3	8	9	2.0	1	5	10	6	2.4	2	9	7	4	2.1	1	7	7	7
Other and unspecified	5	2.5	0	2	2	0	2.5	0	2	2	0	2.6	0	3	2	0	2.4	0	3	1	1
Subtotals	74	1.9	3	12	26	26	2.1	3	17	32	19	2.5	6	32	27	8	2.1	3	21	26	23
Western Union—																					
Leased telegraph grade lines	20	2.3	1	9	5	5	2.4	1	9	6	4	2.5	1	11	5	3	2.1	1	7	4	7
Leased voice-grade lines	13	2.2	1	5	3	4	2.3	0	5	5	2	2.5	2	5	4	2	2.1	2	3	2	6
Satellite channels	8	2.8	1	4	3	0	2.6	0	5	3	0	3.1	2	5	1	0	2.8	1	5	1	1
TWX/Telex	3	2.7	0	2	1	0	2.3	0	1	2	0	3.0	0	3	0	0	3.0	0	3	0	0
Other and unspecified services	5	2.6	1	1	3	0	2.5	0	2	2	0	3.3	1	3	0	0	2.8	1	2	0	1
Subtotals	49	2.4	4	21	15	9	2.4	1	22	18	6	2.7	6	27	10	5	2.3	5	20	7	15
MCI—																					
Leased voice-grade lines	10	2.5	2	4	1	3	2.7	0	7	3	0	3.2	3	6	1	0	3.1	2	7	1	0
Other and unspecified services	7	2.6	2	2	1	2	2.6	0	5	1	1	2.7	2	3	0	2	2.6	2	2	1	2
Subtotals	17	2.5	4	6	2	5	2.6	0	12	4	1	3.0	5	9	1	2	2.9	4	9	2	2
SPCC—																					
Leased voice-grade lines	13	2.5	2	5	3	3	2.8	3	6	2	2	3.0	5	5	1	2	2.7	4	4	2	3
Leased wideband lines	4	2.8	1	2	0	1	3.0	0	4	0	0	3.8	3	1	0	0	3.5	3	0	1	0
Other and unspecified services	3	2.0	0	1	1	1	1.7	0	1	0	2	1.7	0	1	0	2	2.0	0	1	1	1
Subtotals	20	2.5	3	8	4	5	2.7	3	11	2	4	3.0	8	7	1	4	2.8	7	5	4	4
RCA satellite channels	6	2.8	1	3	2	0	3.0	1	4	1	0	3.3	2	4	0	0	2.8	1	3	2	0
Telenet packet network	17	2.8	1	11	5	0	2.8	2	9	5	0	2.9	3	9	3	1	2.5	2	5	6	2
Tymnet packet network	6	3.3	2	4	0	0	3.3	2	4	0	0	3.2	1	5	0	0	3.3	2	4	0	0
Privately built facilities	28	3.4	13	10	2	0	3.5	16	11	1	0	3.6	17	11	0	0	3.3	13	9	6	0
Other and unspecified services	7	2.7	2	2	2	1	2.7	2	1	4	0	3.3	3	3	1	0	2.6	0	4	3	0
Canadian facilities	15	2.3	0	8	4	3	2.4	0	7	7	1	3.1	3	10	2	0	2.4	1	6	6	2
GRAND TOTALS	726	2.4	84	250	229	142	2.6	95	312	226	76	3.0	170	417	104	28	2.6	107	295	212	101

Legend: WA—Weighted Average; E—Excellent; G—Good; F—Fair; P—Poor.
Weighted Average is calculated by assigning weights of 4, 3, 2, and 1 to the four ratings, respectively.

Table 1. Users' ratings of communications facilities, overall summary

Line usage

A third group of large users consists of those who met the arbitrary "large" criteria and answered the number of lines and line usage questions.

	1979 Survey			Last Year's Survey		
	Large Users	Small Users	Total	Large Users	Small Users	Total
Users reporting	27	222	249	39	325	364
Average usage reported (hrs./line/day)	11.0	9.7	10.4	9.8	10.0	9.9

This year's survey continued the trend shown in our past surveys of an increasing number of lines per user. Average network size stayed at a large value, and line usage remained at nearly the same average level. Any significant variations in these parameters over the past few years in our surveys indicate either a varying group of subscribers who respond or variations in their use of communications facilities. We do not present these figures as an accurate indication of what

Users' Views of Transmission Facilities

a "typical" data communications user is doing, but rather as an insight as to trends.

Network Problems

The number of users who indicated that they had encountered problems is tabulated below according to the types of problems they reported.

Problem	Number of Responses	% of Total Responses	% of Last Year's Responses
Local loop	134	41	47
Interchange channels	62	19	25
Modems	180	55	66
Multiplexers	42	13	10
Terminals	168	52	60
Host computer hardware	93	29	30
Host computer software	121	37	41
Front-end hardware	72	22	24
Front-end software	84	26	27

In all but one category, the percentage of complaints decreased from last year's survey. The exception, multiplexers, is understandable because the increase is directly proportional to the increase in the number of subscribers who indicated use of these devices.

As was the case in previous surveys, a puzzling inconsistency exists between the number of complaints recorded in this survey and the ratings provided by our subscribers in the individual reports covering the various types of equipment. In each case, the ratings given on the items in the individual equipment surveys indicate a higher degree of satisfaction than the complaints recorded here would lead one to conclude. A brief summary of weighted averages derived from recent Datapro user surveys is as follows:

	Number of User Responses	Overall Satisfaction	Reliability
Comm. Processors/Controllers (Front Ends)	235	3.4	3.5
Batch Terminals	311	3.1	3.0
Display Terminals	742	3.4	3.3
Teleprinters	455	3.3	3.1
Modems	1039	3.5	3.5
Multiplexers	80	3.4	3.2
Transmission Facilities	726	–	3.0

It would seem that our subscribers are more tolerant of hardware problems when they are not associated with the overall transmission facility.

User Ratings

Many of our 326 reporting subscribers rated more than one service. Consequently, a total of 726 responses were generated for rating particular services. These are summarized in the accompanying "Users' Ratings" table. We have also included a summary of the responses from Canadian subscribers. Not all of the users who identified a particular type of service in the first part of the survey gave ratings on these services in this part of the survey. Therefore, the number of ratings for particular types of services do not agree exactly with information presented in the usage pattern tables.

The categories rated this year differed from past surveys in that planning assistance and ease of installation were rated separately. The results may be giving a message to users as well as to carriers. The majority of written comments supplied by our subscribers had to do with these categories and also with maintenance. Almost across the board, users rated the planning assistance provided by the carriers as inferior to the actual installation effort, which is hardly glowing in its own right. The users seem reasonably satisfied with the facilities once they are operational, but continue to give the maintenance service relatively poor marks when an interruption occurs.

Many subscribers indicated extreme dissatisfaction when operating in multi-vendor environments. The familiar finger-pointing game continues, with each participating vendor claiming that his service, line segment, etc., is operating properly. Unfortunately, it appears that this situation will be with us on a continuing basis as long as no single carrier takes turnkey responsibility, which doesn't seem to be in the cards unless the carrier is AT&T's Long Lines. Even though several operating companies, not necessarily Bell companies, may be involved in providing the service, Long Lines makes all the installation arrangements, cuts over the service (usually on or very close to the projected turn-up date), and then operates the service, taking responsibility for determining where problems exist and taking corrective action as required. When multiple vendors are involved without one that has overall responsibility, each carrier checks the portion of the service for which he is responsible and leaves it at that. Quite frequently, the check is not end-to-end but test-center-to-test-center, which admittedly tests out the bulk of the circuit, but not 100 percent of it; also, the actual interface connection between vendors is not usually checked.

For the most part, the carriers are conscientious, but all too often the smallest participant or the last vendor in the network gets the lion's share of the blame, and it becomes his inherited responsibility to determine what is at fault and take steps to correct it. This is no small task when the problem is in someone else's area and that someone is not only a competitor but one that is much larger.

Users' Views of Transmission Facilities

The test boards in use by most of the carriers are still related to voice services; and, even though data transmission has been around in bulk for many years, qualified data technicians are still in short supply. The test equipment vendors have made some significant strides in recent years, and now a firm doesn't have to be one of the Fortune 500 to utilize sophisticated on-site test/monitor devices. The trend toward user-owned test centers continues to increase.

The ability to isolate problems may be the only way a user can avail himself of the price/performance advantages of some of the new offerings and still protect himself against prolonged outages due to finger-pointing. One user found that the only feasible solution to the service problem was to make friends with the installation and maintenance personnel. Another could only obtain satisfaction by bringing a personal relationship with a vice president into play after months of poor service. However, as the ratings indicate, not everyone was dissatisfied with the service aspect of the carriers—but many of the Excellent and Good ratings came from users that had on-site test/monitor hardware.□

Topic Index—Section CS20

General Topic	Report Title	Report No. CS20-	See Also Report
Need Assessment Planning Overview	—How to Determine Your Data Communications Needs	-110-101	
	—A Digest of the Latest Communications System Planning Techniques	-115-101	
	—Gaining a Better Perspective of Communications System Applications	-120-101	
	—Planning Ahead for Your Data Communications Applications	-121-101	
	—A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs	-150-101	CS50-815-101
Terminals	—A Short Checklist of Terminal Planning Factors	-205-101	CS10-210-101; CS20-215-101; CS20-210-101; CS25-220-101; CS20-235-101; CS25-220-101; CS60-221-101;
	—How to Calculate the Number of Terminals You Will Need in Your Communications Network	-210-101	CS25-615-101
	—How to Measure a Terminal's I.Q.	-215-101	
	—Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path	-235-101	CS10-230-101
Transmission Facilities	—Planning Considerations for Dial-Up Versus Lease-Line Facilities	-325-101	CS15-710-101; CS15-711-101
Office Automation	—An Office Automation Study	-400-101	CS20-610-101; CS20-710-101
EFT	—High-Velocity Money—Preparing for EFT	-410-101	
	—Planning Guidelines for an EFT Audit and Control System	-412-101	
	—The Economics and Other Issues of EFT	-415-201	CS60-160-101
Distributed Processing	—How to Conduct a Distributed Processing Feasibility Study	-510-101	CS60-510-101; CS60-511-101; CS60-170-101
	—Organizational Design for Distributed Processing	-515-101	CS20-510-101
	—System Architecture for Distributed Data Management	-520-101	
	—A Language for Distributed Processing	-530-101	CS50-510-201; CS15-170-101; CS60-820-101
	—Maintaining Data Order and Consistency in a Multi-Access Environment	-540-101	
Electronic Mail	—Planning for Electronic Mail	-610-101	CS60-160-101
Facsimile	—Facsimile—How to Analyze Its Role in Your Organization	-650-101	CS20-710-101
Teleconferencing	—Planning for All-Electronic Multimedia Corporate Communications	-710-101	CS60-850-101
MIS	—The Communications Framework of Management Information Systems	-940-101	

Topic Index—Section CS20

General Topic	Report Title	Report No. CS20-	See Also Report
Need Assessment Planning Overview	—How to Determine Your Data Communications Needs	-110-101	
	—A Digest of the Latest Communications System Planning Techniques	-115-101	
	—Gaining a Better Perspective of Communications System Applications	-120-101	
	—Planning Ahead for Your Data Communications Applications	-121-101	
	—A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs	-150-101	CS50-815-101
Terminals	—A Short Checklist of Terminal Planning Factors	-205-101	CS10-210-101; CS20-215-101; CS20-210-101; CS25-220-101; CS20-235-101; CS25-220-101; CS60-221-101;
	—How to Calculate the Number of Terminals You Will Need in Your Communications Network	-210-101	CS25-615-101
	—How to Measure a Terminal's I.Q.	-215-101	
	—Providing User-Oriented System Interfaces	-220-101	CS20-205-101
	—Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path	-235-101	CS10-230-101
Transmission Facilities	—Planning Considerations for Dial-Up Versus Lease-Line Facilities	-325-101	CS15-710-101; CS15-711-101
Office Automation	—An Office Automation Study	-400-101	CS20-610-101; CS20-710-101
EFT	—High-Velocity Money—Preparing for EFT	-410-101	
	—Planning Guidelines for an EFT Audit and Control System	-412-101	
	—The Economics and Other Issues of EFT	-415-201	CS60-160-101
Distributed Processing	—How to Conduct a Distributed Processing Feasibility Study	-510-101	CS60-510-101; CS60-511-101; CS60-170-101
	—Organizational Design for Distributed Processing	-515-101	CS20-510-101
	—System Architecture for Distributed Data Management	-520-101	
	—A Language for Distributed Processing	-530-101	CS50-510-201; CS15-170-101; CS60-820-101
	—Maintaining Data Order and Consistency in a Multi-Access Environment	-540-101	
Electronic Mail	—Planning for Electronic Mail	-610-101	CS60-160-101
Facsimile	—Facsimile—How to Analyze Its Role in Your Organization	-650-101	CS20-710-101
Teleconferencing	—Planning for All-Electronic Multimedia Corporate Communications	-710-101	CS60-850-101
MIS	—The Communications Framework of Management Information Systems	-940-101	

How to Determine Your Data Communications Needs

flowing in the proposed network. The volumes are best expressed in bits per unit time and need to be identified in terms of origination and destination locations. The designer will need to know not only average values of the volumes flowing on any given link (that is, connection between a source and a destination) but also the peak volumes, normally expressed as bits per busy hours or bits per busy week, etc. The steps in such an estimation process generally can be made for each link as follows:

- Estimate the number of users generating demand at a given location, both for sending data and receiving it.
- Estimate the number of messages transmitted and received per user at the location. If users vary considerably in the volume they generate, as is usually the case, assume a normal distribution based on sample checks and use the mean value.
- Calculate the total number of messages transmitted and received at the location.
- Estimate the number of bits per message in each direction. This is done by estimating the number of characters per message and then applying the character-to-bits factor, usually in the range 8—10, depending on the code used. It is useful to take the following factors into account when estimating the number of characters:
    - A message is an uninterrupted stream of bits, that is, it may represent several separate logical units (for example, sales orders) that have been grouped together for convenience. In practice, the method of estimation is, however, likely to be based on the expected number of orders, rather than messages, and then the grouping of orders to form messages is assumed to produce an estimated over-all message size.
    - Not only printed characters should be counted, but also spaces between words, control characters (for example, line feed and carriage return), etc.
- Calculate the average volume of bits per time flowing in each direction. This is computed by multiplying the average number of bits per message by the average number of messages in each direction.
- Estimate the peak load. Most business systems experience a busy hour at about 10:30—11:30 in the morning, with a secondary, lower peak hour at around 14:45—15:45 in the afternoon. Estimating the actual load, however, is a difficult step, because even when sampling methods are used (for existing systems), it is usual to find major discrepancies between measurements. Judgement has to be applied in each situation, but in the absence of any information about expected peaks, the following two approaches have both been applied in practical situations.
    - The first approach is based on the assumption that about 20 percent of the average daily traffic occurs in the busiest hour of an 8-hour day. The corresponding figures for a 12-hour day are 16 percent and are 14 percent for a 24-hour day. However, in most situations encountered by the authors, the actual peaks have been considerably higher than the above figures would indicate. Hence, the following method is recommended:
    - Assume that the traffic in the busiest hour of day is 2.5 times the average hourly level of traffic over the whole day. This indicates a higher peak load than that suggested by the first method but seems to correspond more closely to real world experience.

It is also helpful to make estimates of the peak loads during other time periods, notably seasonal or annual variations—for example end-of-year accounting or stock requirements. All such estimates of traffic, based on expected user demand, are key input to later design work in estimating transmission times.

A sample calculation of traffic estimation for a given link for an on-line inquiry type of application is presented in figure 1. Traffic loads vary widely, however, by the type of system. For example, some typical values of traffic handled by terminals are indicated in figure 2.

Estimated number of users at remote location	40	
Estimated number of enquiries per user per day	2	
Estimated number of characters per inquiry per user	90	(to the computer)
	700	(from the computer)
Number of bits per character	10	

Daily volume of traffic is therefore calculated to be

(40 x 2 x 90 x 10) bits per day, to the computer
(40 x 2 x 700 x 10) bits per day, from the computer

that is

72 000 bits per day, to the computer
560 000 bits per day, from the computer
632 000

Estimated traffic in busy hour of day

Method 1: 20% of 632 000, i.e. 126 400 bits

Method 2: $2.5\left(\frac{632\,000}{8}\right)$, i.e. 197 500 bits

Figure 1. Sample traffic calculation for inquiry-response application

How to Determine Your Data Communications Needs

Type of System	Typical Line Speed Used (bits/second)	To the computer	From the computer
Off-line	1200	1 million	4 million
Remote batch	2400	4 million	10 million
On-line data collection without updating	200	0.5 million	0.1 million
Enquiry—response	200–2400	0.05 million	0.5–2.0 million
Real-time	200–4800	0.05 million	1.0 million

Figure 2. Typical traffic volumes (in bits) by a terminal during one day

RESPONSE TIME

Great care has to be taken when assessing a user's apparent response-time requirements. He or she may genuinely believe that an interactive system, for example, should provide the information within, say, 3 seconds. Deeper investigation, however, often shows that it is not the information itself that is required so quickly, but rather just an acknowledgement from the system that the user's request is being processed. A strict definition, therefore, of response time is: that interval of time that elapses between the completion of the user's last keyboard (or other) entry at the terminal and the display at that terminal of the first character (or the output form) received from the computer system or network in response to the command just initiated.

Such an interval of time is of course made up of several independent elements, notably the time taken for the two-way data communication itself (possibly via several nodal computers), the availability of processing time at the destination computer, and then the efficiency of the data retrieval process itself. In the case of off-line transmission, such elements are readily analysed separately, but when on-line operation is concerned, it is usually one of the most difficult of all design tasks to meet the required response time with any degree of certainty.

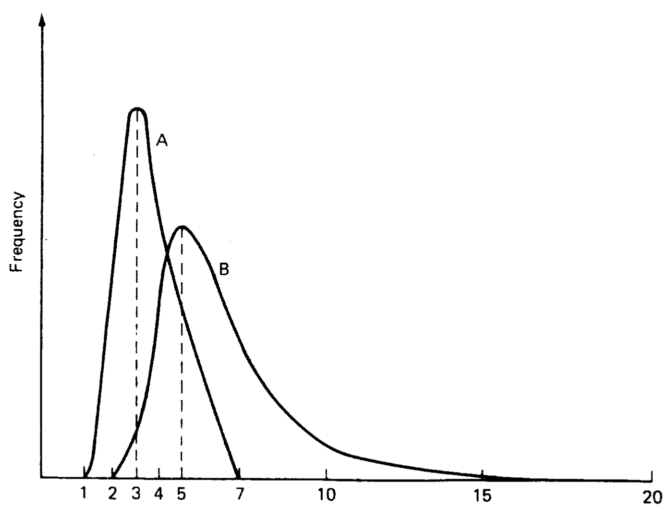


Figure 3. Response-time distribution; two sample curves

Because of this difficulty, it is well worth spending considerable time with each category of user to find out their true response-time requirements. Unrealistic and unnecessarily short response times cause excessively high cost both during design and operation. Moreover, it is common for a design specification to make such statements as: 'at least 90 per cent of the response times must never exceed 7 seconds'. The implications of such requirements can be enormous for the designer, who has to take into account the distribution of expected response times. Figure 3, for example, illustrates two distributions: curve A is that imposed in a given application by a requirement not to exceed 7 seconds. The curve shows that the minimum respone time achievable is 1.5 seconds and the median value is about 3 seconds. Curve B is much less demanding on the designer, because it allows a small percentage of response times to be very high—10—20 seconds. Yet the median value for curve B is 4.5 seconds, acceptable for many applications. The startling point is that the cost difference between two such systems is likely to be very high, the system represented by curve A being perhaps double the cost of that represented by curve B, because of the need for higher speed lines, modems, back-up facilities, etc. Hence there is every incentive to establish the realistic requirements of users with regard to response times.

TYPE OF SERVICE PREFERRED

One of the key factors to be established when evaluating user needs is the type of service best suited to each user. Some users will simply want a routine report of well-defined information at regular intervals; some will want immediate access to information on an inquiry/response basis; some will want hard-copy of any information they obtain, while others may want only selected items; etc.

Some of the points to be determined are as follows:

- Is hard-copy required? If so, can it be supplied later or must it be immediately available? Will multiple copies be required?
- What security requirements does the user want for access to his data and terminal?

How to Determine Your Data Communications Needs

- What language does the user need to work in? By this is meant true spoken languages (English, French, German, etc.), not computer languages. More and more today systems involve crossing national borders and some minimal facilities (for example, in the command language of an on-line system) can easily be provided.

- Is the user interested in a Selective Dissemination of Information (SDI) service, whereby he receives indications of new information lying within his particular field of interest?

- Does the user need to have processing facilities to back up his information retrieval or input work?

- Does the user need some kind of graphical facilities, either on-line on a display screen or perhaps using an off-line plotter?

- Does the user need to access more than one data base and, if so, does he need guidance, to be referred to the appropriate database?

- What kind of error detection and error correction will be needed to support the user?

Users are generally very reluctant to move directly to a purely on-line inquiry/response system from a situation where they have access to full hard-copy files produced on a routine basis. It takes some time (usually months) of regular use of an inquiry/response system before users are prepared to dispense totally with their routine hard-copy files. However, one practical approach to help minimize this period when confidence is being gained by the user is to offer a "nested" data approach, that is, the user can ask for progressively deeper but selective analyses of the same item merely by specifying a code offered to him at the terminal.

SECURITY REQUIREMENTS

The requirements for security of a data communications system cover a wide range of interests. Although this report is primarily concerned with determining the needs of the users of the system, any security measures installed must be integrated so as to protect the interests of the organization as a whole. Hence, the following aspects need to be considered in parallel the users' interests:

- normal commercial management prudence

- auditors' requirements

- interface with or by data processing systems linked to the system (possibly operated by third parties)

- corporate security arrangements (for example, physical access to areas where terminals, back-up equipment, etc. are kept)

- back-up facilities in the event of breakdown.

The definition of requirements for security measures for the users' terminals should therefore include consideration of physical access controls, passwords, coding, etc.

In investigating user requirements for security, the above aspects should be considered, but it always remains the designer's responsibility to provide overall security in terms of reliability of the system, protection against accidental or deliberate data destruction, planned maintenance of hardware (including terminals), etc.

TYPE OF TERMINAL PREFERRED

The type of terminal is determined largely by the factors of cost, the nature of the application, and the type of system being designed. For example, an off-line system transmitting batch data assembled from many different locations that will not interact directly with the user is likely to be one capable of handling punched cards or paper tape. Large batch on-line terminals, likewise, will be determined by technical and cost requirements and not by the end-user. It is when the end-user makes direct use of a terminal that special care has to be taken of end-user requirements. For example, preferences for visual display screens or teleprinter terminals, keyboard design, screen brightness, etc. should be taken into account.

TRAINING REQUIRED

In gathering information about users' needs in the context of a data communications system, special note should be taken of the potential abilities of the different types of user to become fully effective when the system goes operational. Persons who are used to operating keyboards, for example, typists, shop assistants, calculator operators, etc., will normally have little difficulty in operating a standard kind of terminal keyboard. People not familiar with keyboards or working with machines may need a considerable adjustment period. Even experienced keyboard operators sometimes find it disconcerting initially to have a machine answer back almost instantly and demand further input. Hence, part of the fact-gathering of user needs should include some evaluation of the expected training requirements and acceptability of the system by those who will actually interact with it.

METHODOLOGY FOR STUDYING USER NEEDS

The first part of this report has been concerned with identifying the type of information that the designer of a data communications system needs to know about the potential users of the system. However, the most thorough investigation in the world will only indicate likely requirements and will contain real elements of risk for the ultimate design, for example, will the response time be acceptable? Are the estimates of traffic volumes realistic or will additional capacity be required? Has some key output data been omitted?

It is important therefore to minimize such risks by undertaking as thorough an analysis as possible of what the users really need. The naive approach to obtain such information is to go and ask users what they need! This widely used approach is moderately effective for traditional data-processing systems, but it is entirely inadequate and can be positively misleading for data-communications-based systems. How can a potential user describe his requirements for an on-line system, for example, when he has never in his life experienced on-line operations? What charging scheme should be applied when the future level of usage is entirely unknown? The difficulty in overcoming problems like these meant that many of the early data-communications systems were built on an act of faith, a belief that, when confronted with a real new tool, users would make good use of it. Such an approach often paid off; but sometimes it did not. Therefore it is prudent to follow some basic steps to gather as much information as feasible about real user needs and expected user reaction.

The following approaches are all relevant and useful, but are best not relied on as a sole means of indicating user needs. Rather, a combined approach is necessary. It is recommended that all steps be followed in sequence so that together they form a real methodology for estimating user needs.

Studying Other Systems

Before any analysis of the immediate problem area is undertaken, it is an excellent idea to take a broad look at the experiences of others with the same general problems. The two basic steps are:

1. a survey of the literature

2. visits to selected operations systems.

Clearly, there is a danger of spending too much effort, with rather uncertain benefits, on this sort of activity, but it is almost always well worth while to pick out from the literature the key results, costs, and problems encountered in other systems that appear to have tried to achieve roughly the same goals as the one being initiated. Where feasible, a visit to a few (say three) installations that have been operational for six months or more is likely to be very instructive. Such visits should concentrate on:

- cost and time of implementation (actual versus plan)
- main problems encountered (technical, human, outside factors)
- user satisfaction
- level of usage achieved
- opinions on what could have been done better.

Interviews with actual users are essential; reliance on the system designers' and operators' opinions of user satisfaction can be very misleading.

Interviews and Questionnaires

This is the traditional approach, widely used in systems analysis. As such, when applied by skilled practitioners, interviews and questionnaires can provide a valuable means of data gathering about new needs. Plenty of guidelines exist on the conducting of interviews and questionnaires, and these will not be repeated here. Rather, the following items are meant to highlight the special points to watch with regard to data communications systems.

1. Use the interviewee's or questionnaire-filler's language, that is, avoid technical jargon completely. Reference to terms such as response time, terminal, line speed, etc. are guaranteed to silence the average interviewee.

2. Wherever possible, provide a demonstration. Even if the actual application is rather limited, the simplicity of the average terminal is likely to encourage the potential user to think of his own possible use of it.

3. Never rely on questionnaires alone. Even the best-designed questionnaire will be misunderstood in parts. A combined interview—questionnaire-filling session can be very effective, however, because it has a ready-made structure.

4. Avoid Yes/No questions. Wherever possible, the respondent should be given a spectrum of choice from which to indicate his stand-point. For example, rather than the question. 'Do you need immediate hard-copy of the answer to your inquiries?' it would be better to ask: 'How long after you have obtained answer to an inquiry do you need, if at all, hard-copy?'

How to Determine Your Data Communications Needs

The heart of the problem in all such user research is to know what assumptions have to be made and presented to the potential user doing interviewing or questionnaire-filling. The potential user must be given some indication of the final form of the system he might be using in order to express his opinion and give relevant facts. Of particular importance to him are:

- What exactly will the service be?
- What will I have to pay for it?
- How will it affect my existing operations?
- How reliable will the service be and will it give me more than my present means?

These are difficult enough basic questions for the analyst/designer to answer, so it can only be expected that a potential user's answers to detailed questions on data volumes, response times, and so on, have to be interpreted with great caution. For this reason, the use of pilot systems is a particularly valuable way of evaluating users' needs and their response to a real data communications system.

Pilot Systems

The following paragraphs will indicate how pilot systems may be used to gather information on user needs. The essence of a pilot system is that it provides a testing ground with limited investment and yet allows direct experience of all the aspects of an operational full-blown system. The danger in interpreting the results of pilot systems is that the users themselves know that the environment is an experimental one. Their reactions may be different—sometimes more favorable, sometimes more critical—from when they are exposed to a truly operational system. Nevertheless, designers of data communications systems learn a tremendous amount about user requirements when they observe a pilot system in operation. Users, of course, learn about data communications systems too, and accept a subsequent system more readily if they have been involved in its design.

The key points about the use of pilot systems for user research are as follows:

1. The objective of the experiment has to be very clearly defined and should not be too ambitious or it will become very costly. A definite time limit should be set.

2. The data to be gathered during the operation of the pilot system should be systematically collected and analysed at defined intervals during the experiment. Both qualitative and quantitive data will be required, as outlined in the first part of this report.

3. Care should be taken to choose an 'average' user group. The temptation is to select an enthusiastic group of volunteers, but these should only be used if there is no other reasonably cooperative 'more average' group available. Of course, if it is possible to mix, or treat in parallel, two such groups, this represents an even better approach.

Forecasting Models

It is common to expend considerable effort on trying to determine the needs of users as they are today or will be when the proposed system is first installed. Rarely are attempts made to forecast the growth of use of the system after it becomes operational. It is not that system designers do not feel the desirability of such projections; it is simply that such information is thought unobtainable or so uncertain as to be useless. Consequently, most user studies do not make forecasts of needs beyond the immediate phase of system implementation.

Clearly, the topic of forecasting in general is outside the scope of this report and the reader is referred to some of the excellent treatises on the subject. It is the dual aim of this report, however, to draw attention to the need for making forecasts in the field of user needs and to identify some simple approaches to deriving such forecasts.

A fundamental point to be born in mind is that the purpose of forecasts is not to predict what will happen; they can only indicate what could happen on the basis of a range of assumptions. Such assumptions rarely hold exactly true in practice, but the best forecasts are those that are not highly sensitive to a single assumption. As it is unlikely that all or even most of the assumptions will be wrong, the general validity of the forecast can then remain, even if — as is likely — some assumptions are wrong.

A further basic principle is that the more distant the time-horizon of the forecast, the greater will be the degree of uncertainty about its validity. This leads to two further guidelines. First, it is pointless to make a forecast for a time horizon that is too distant to be able to attach any real confidence to it. In the area of user needs for data communications, this normally means that a 5-year horizon is reasonable, but that 10 years is the maximum feasible forecast period. Beyond 10 years, there will be far too much uncertainty about the technology to be used, about the economic state of the organization, etc. Secondly, it makes sense to concentrate one's efforts on the near-term (say 5 year) forecasts and to use a cruder model for the longer term. Hence the technique used to forecast the near term may be based on a careful analysis of the existing situation followed by an assessment of the effects of many different factors on that base situation, whereas for the long term it may be considered adequate to assume, say, a simple 10 per cent annum growth after the initial 5-year period. The data communications system designer is usually eager to provide a system that will meet the needs of users up to 5 years, but expects that the system will anyway be overhauled in some way within 10 years. Hence the near-term forecast has great influence on his immediate design, while the long-term forecast helps him more in formulating his strategy towards providing flexibility of the system.

There are two broad types of growth that should be forecast, each needing to be handled in different ways.

Expansion of Use via the Original Users

This represents the 'natural' growth of the system and is normally the main element considered in forecasting the future use of traditional data processing systems. Hence the usual key factors here are as follows:

1. Growth in the volume of business or organizational activity (giving rise to more sales orders, invoices, inquiries, production achieved, etc.).

2. Intensification of use by users themselves. As users become familiar with a given system, there is usually a slow trend to increase their use of it. This is particularly true of inquiry/response systems, where a doubling of activity per person sometimes occurs over a 3-year period, following the initial settling-in process.

An example of a forecast based on the above factors might be as follows. In a system being designed to connect remote sales offices to a central system, it is determined that the current volume of sales order for a given branch office is 80 per day. Each order contains 120 characters of leader information and an average of six lines of items ordered, each line containing some 60 characters. Hence the total volume of sales order characters that would currently be transmitted if the system were operational is (80 x (120 + (6 x 60))) = 38 400. In addition, the computer responses, errors, re-transmits, etc. would bring about another 10,000 characters daily. Furthermore, the employees at the sales office today find that, although they receive a weekly printout of customers accounts, they have to make about one telephone inquiry to the head office accounts department for every fourth order. It is estimated that such an inquiry could be handled via a terminal on the proposed system with about 50 characters being transmitted to the central system and 200 characters response, that is, 250 characters per inquiry.

It is desired to estimate the total daily traffic at the time of implementation, 2 years hence, and at a period 4 years after that. The financial department estimates that during that time turnover will grow at about 7 per cent per annum in real terms (not counting inflation). The sales department believes that the average of six lines of items per order will tend to increase, say to eight lines over the next few years. A reasonable approach to preparing such a forecast would therefore be:

- Current daily traffic (if the system were operational)
  = 48 400 + (0.25 x 80) 250
  = 53 400 characters

- Two years hence—
  Expected sales orders = 80 compounded at 7 per cent per annum for 2 years
  = 92
  Expected daily traffic = {92 x [120 + (7 x 60)]} + (10,000 compounded at 7 per cent for 2 years) + (0.25 x 92) 250
  = 49 700 + 11 500 + 5750
  = 67 000 characters (rounded)

- Four years after implementation—
  This is calculated similarly to the previous forecast, except that here it is assumed that the inquiry/response facility will be used more intensively, say once for every three orders (instead of four).

  Therefore,
  Expected sales orders = 80 compounded at 7 per cent per annum for 6 years
  = 120
  Expected daily traffic = {120 x [120 + (8 x 60)]} + (10,000 compounded at 7 per cent for 6 years) + (0.33 x 120) 250
  = 72 000 + 15 000 + 10 000
  = 97 000 characters

Such an approach could of course take into account many other factors, such as new product coding (changing the number of characters per line), differences in turnover growth by product type, etc. But usually the broad indication of an expected traffic volume for a given link of 67,000 characters daily, rising to 97,000 over the following 4 years is adequate for the systems designer without further refinement. It often surprises designers to see how much an apparently modest growth of turnover of 7 per cent per annum, combined with factors such as more lines per order and some increase in the inquiry/response volume, can give rise to a traffic in 6 years at a level over 80 per cent lighter than that estimated for the current level. It thus further emphasizes the impor-

How to Determine Your Data Communications Needs

tance of making forecasts, however simply obtained, of the future growth in usage of a planned system.

Extension to New Users

The preceding example and discussion were concerned with the growth of the planned system via the initial user base. In many cases this may be quite sufficient for the forecasting of system growth, but there are some general application areas where considerable expansion of use takes place among entirely new users. Indeed, sometimes the major growth element is to be found there, not among the initial users nor as a result of the natural growth of the organization. This type of application is typically one where a service is offered to a wide range of potential users and is normally of the inquiry/response type, with perhaps some additional services such as issuing of copies of standard reports on demand only.

In such applications, the designer is faced with the classical market research problem, namely, what penetration of the potential user base will be achieved by a new service. Forecasting this kind of system usage is always very difficult, but in the absence of any such estimates the designer's job becomes even more difficult! It is therefore well worth deriving some estimates using the following principles.

First, every possible use should be made of hard data. For example, in interviewing or surveying the user's needs, specific questions should be put about their expected usage of the proposed system. Experience of others in offering similar services should also be sought in order to obtain as much useful data as possible.

Second, an examination should be made of possible substitution effects. These are among the most reliable of forecasting techniques and are used when a new service is expected to replace, perhaps only partially, an existing one. In the sample forecast given earlier in this report, the replacement of the sales office telephone queries to head office about customer accounts by on-line terminal inquiry is an example of a simple assumption about substitution (in that particular case, a one-for-one substitution).

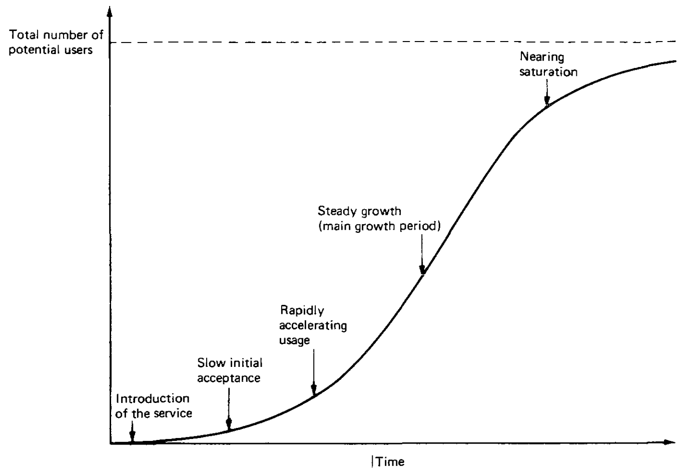


Figure 4. Typical pattern of growth of usage of on-line services

Third, the general pattern of penetration of on-line information services is usually that of an S-shaped curve, as illustrated in figure 4. This type of penetration curve is common to many types of market penetration, but it is of course difficult to select the actual curve most applicable in a given situation.

As much hard data, such as the growth of similar systems implemented in the organization as possible, should of course be used, and by judicious use such forecasting models can be a major aid to the designer of a data communications system.☐

A Digest of the Latest Communications System Planning Techniques

Problem:

Once you have assessed, analyzed, and quantified your needs for improved communications facilities, then you can start to think seriously about hardware, software, and transmission facilities. But you can't just plunge ahead and buy this thing because you like its color, that thing because it's got the right nameplate, and the other thing just because it's handy. Unlike buying a computer, in which there is really a somewhat limited set of variables to account for, buying a communications system touches on every single aspect of an information processing complex and even laps over into non-EDP areas like corporate structure, the corporate management style, and corporate growth plans.

This report gathers all these pieces together into a succinct digest that will give you at least the framework of a structure you can use for your formal planning and development program. All of the reports in this and the next two sections are essentially elaborations of each of the points highlighted in this report.

The practical and important techniques for planning, evaluating, and optimizing data/computer communication networks are presented for managers and designers. Managers must be aware of the range of work required to select the necessary people and to provide the necessary resources. Designers will find the concisely presented information on traffic and requirement analyses, performance criteria, communication device and transmission facility selections, network architectures, and design techniques, tools, and tradeoffs to be very useful.

Solution:

The demand for data/computer communication network usage has been continually increasing, and no end is in sight. This demand is a direct consequence of the advance in microelectronics, the development of wideband digital transmission facilities, and the desire to extend computing resources to more remote locations. The availability of microelectronic components allows the development of new data communication applications and new data communication technologies that would otherwise not be economically feasible. The availability of wideband digital transmission facilities opens the door for applications that require

"Planning and Design of Data Communications Networks" by Wushow Chou, North Carolina State University. From the AFIPS 1974 PROCEEDINGS of the National Computer Conference. Reprinted by permission.

inexpensive, low error rate, high bandwidth facilities, such as high speed digital facsimile. The economy and convenience of extending the use of computing resources, such as database and software packages, have promoted the development and expansion of communication-based computer systems, which, of course, use data communications.

This trend of rapid growth further stimulates new applications and increases the number of users. As a result, the rate of innovations in system design concepts, hardware features, and transmission services is accelerating. In turn, new applications and new demands are induced. The cycle continues. Eventually, there will be data terminals in every home and office, just as there are now telephones. The data terminals, or terminals combining data and voice transmission capabilities, will outnumber the telephone sets limited to transmitting only voice.

Data/computer communication networks are becoming increasingly confusing, large, and complex as innovations increase. Two major problems are faced during network design: sorting out the confusing alternatives and acquiring a set of practical techniques for evaluating and optimizing the design.

The confusion in selecting an appropriate design stems, in part, from the difficulty of recognizing and considering a large number of available application and design alternatives. The alternatives are constantly evolving and expanding. Also, it is not easy to understand the limitations and applicability of all alternatives. There are multipoint polled teleprocessing network architectures, store-and-forward packet switching architectures, ring switching network architectures, and others. Line control procedures are asynchronous and based on characters, synchronous and based on characters, or based on bits. Cost saving device options include multiplexers, statistical multiplexers, intelligent multiplexers, programmable concentrators, and front-end processors. Where modems were required to transmit digital information by analog signals, now digital services and digital interface units are the more natural approach. While we are still struggling to understand and distinguish among the various terrestrial transmission services offered by common carriers and special common carriers, we are given the additional alternative of satellite communication services. Even terrestrial alternatives are expanding as packet radio, cable TV, and fiber optics are becoming available for data transmission. Journal articles and conference presentations suggest that data communication networks can handle many other functions such as electronic message service, teleconferencing, and electronics fund transfer. Cryptic terminologies, like system network architecture, distributed communication architecture, and X.25, add further to the confusion for many trying to design a system.

Techniques for analysis and design are not readily available. There has been little incentive for anyone to refine and present such techniques to others. Developing a new concept brings more fame than playing with analytic techniques. Network design can be more expeditious without a thorough cost/performance analysis and optimization. The performance constraints that are defined and the least cost design that is set as the goal in the planning process are often not taken directly into account during the actual design process. Rules of thumb almost always prevail. With the increasing size and complexity of communication systems, rules of thumb are no longer always adequate or even applicable. We need better techniques and tools to determine the best system concept, to evaluate the design strategy, and to select the most effective cost/performance communication devices. We need to know the impact of new technologies, new devices, new services, new tariffs, and new applications.

Managers want to know particularly what must be done now to plan a new or upgraded data communication network. Designers and analysts must somehow handle all the complexities involved in analyzing user requirements, setting design constraints, and evaluating performance. Users or others must select from among many communication devices and transmission facilities to determine the ones with the best cost/performance for tradeoffs to determine their applications.

The following sections give the techniques needed for planning and designing data communication networks. Particular emphasis is on new techniques developed and used by leading experts in the field to handle the difficult problems in modern networks. Information is given about common planning pitfalls, important traffic characteristics, performance definitions, communication devices, transmission facilities, network structures, and techniques for design, optimization, and performance evaluation.

MANAGEMENT AND PLANNING

In most of the large operational communication networks, costs can easily be reduced by 15 percent or more with only minor alterations in the network. The cost savings can often reach 30 percent or more if the whole network design can be reoptimized. Furthermore, the cost reductions can be realized without degrading performance. In fact, both cost and performance can often be improved simultaneously —all of which amounts to a strong indictment of most existing network designs.

One factor contributing to inefficient design is that (voice) communication systems and computer systems are often managed as two separate entities,

A Digest of the Latest Communications System Planning Techniques

especially in large corporations and government agencies. When the operation of a data communication network requires the knowledge of both communications and computers, it is generally left either to communications management personnel, who lack computer background, or to computer management personnel, who have no training in communications. Within either management hierarchy, the data communications problems end up in the hands of middle- or low-level managers.

The middle-level technical managers face many difficulties in planning and operating data communication networks. These network managers are usually technically oriented and often lack sufficient training in management. They are likely to have certain characteristics common to many technical people—they tend to overestimate their ability and to overcommit themselves for technical projects.

The middle level of management usually does not have enough power to handle the job. Approval from top management for buying communication equipment and paying line costs might be relatively easier than obtaining funds for less tangible expenses, such as developing programs and seeking expert assistance during the planning and design process. Middle level managers must often make compromises within the limitations of inadequate budgets. They do not want to risk user dissatisfaction, so they tend to overdesign networks. They might use more lines and equipment than are actually necessary and still design a network that does not guarantee satisfactory performance.

Network managers do not want to risk a blunder because of their relatively low level in the organization. Usually they hesitate to initiate any innovation or change unless it is absolutely necessary. Lack of coordination is another difficulty. Different groups must exchange technical assistance in a broad data processing system or in the larger corporate communications system of which the data communications network is a part.

A significant factor contributing to inefficient designs is a general underestimation of the complexity of data communication. Data communication networking is an evolving, new technical field. Mastering this field requires knowledge of computer software and hardware, communications hardware, transmission facilities, line tariffs, human psychology, queueing theory, statistics theory, communications theory, advanced computational techniques, and the most advanced network optimization techniques.

Datapro Comment:

We have included selected reports on each of these topics throughout this service even though some of them, like queuing theory, quickly resort to advanced mathematics. But there is no easy, simplified route you can take to solve the problems of communications system design, and we firmly believe that any attempt to avoid or bypass the technical nitty-gritty will produce a poor system. We are delighted to find someone else who agrees with us.

Some of the advanced techniques are the result of the most current research. The planner must be aware of the new developments in the field. In addition, he must be clever and think innovatively. Not many people possess all these capabilities. Unless they are aware of the qualifications and situation, a big corporation's management is likely to think that someone in the company should be able to develop expertise in the area of data communications networks, or at least someone with this expertise should be easy to hire. Assistance from consultants is sought out of desperation in some cases. However, technical competence does not always play enough of a role in the selection of a consultant when management thinks that any of the many data communications consultants can do the job. A consequence of their failure to recognize what is necessary is that management might not get the best design for all of the company's needs.

TRAFFIC AND REQUIREMENT ANALYSES

Successful implementation of a data communication network depends largely on the thoroughness of the data traffic and user requirement analyses.

The most tedious part of planning is gathering traffic information. Traffic information, including current measurements and future projections of traffic, should be collected from every user. Information must be determined for each terminal and for each type of message or transaction in the form of distributions of the number of transactions per unit time during the peak hour, average day, and peak day; input message length distribution (number of characters per message); output message length distribution; and priority.

Accurate measurements and projections are usually impossible to make, but continued attempts to update the information can be a valuable exercise for network users and managers. It is often helpful if the planner visits the users to help them and to validate the information supplied by them. The planner must work with what is available, so some factors in the analysis might not always be fully defined.

Even more uncertain than the traffic statistics and projections are the users' requirements for network performance. In general, users do not know exactly what they want. Sometimes they demand a level of performance that is practically impossible to achieve. Other times, they demand a performance level they might not need and must pay a high price to attain. If cost were not a concern, any user would like to see negligible response time and almost perfect network reliability. Apparently, no one can afford to pay for such a level of performance. But users usually do not have much feeling for the cost/performance relationship. It is the responsibility of the planner to educate the users, to show them the relationship between cost and performance, and to assist them as they modify their performance requirements to reasonably obtainable levels. (Cost and performance tradeoffs are described in the last portion of this report.) The traffic information and the users' performance requirements form the constraints for the design.

PERFORMANCE CRITERIA

A general goal in designing a data communication network is to design a minimum cost network satisfying performance requirements or criteria. A common slogan in the field is to improve the cost/performance ratio. What is performance? It means different things to different people. For a well designed network, the performance should be measured by the following criteria: blocking probability or message response time, traffic capacity or throughput, network reliability, transmission error rate, and sensitivity to variations in traffic level. Knowing the traffic bottleneck is also important information, especially for future expansion and upgrading. However, it is not generally used to measure current network performance.

Blocking Probability

This criterion is used to measure how promptly a data communication network responds to calls from dial-up terminals. It can be defined by any of the following three factors:

1. At least A percent of calls obtain access to a computer port within B minutes. This definition is used for systems where the terminal dials to a switchboard and waits while an "operator" attempts to connect the terminal to an empty computer report.

2. At least C percent of calls obtain access to an empty port on the first attempt.

3. D percent of calls obtain access to an empty port with no more than two attempts, three attempts, etc.

The parameters A, B, C, and D are constants determined by user requirements or network planners, for example, A = 99, B = 5, C = 95, and D = 99.

Message Response Time

This criterion is used to measure how promptly a system responds to terminals connected to the system by leased or private lines. Message response times have different definitions at different parts of a data communication network. As far as the users are concerned, terminal response time and overall response time are most meaningful. The terminal response time is defined as the time required from the instant the transmit key or equivalent key on a terminal keyboard is depressed until the moment the reply message begins to appear at the terminal. This is the most commonly used criterion. However, it has serious drawbacks when used as a measure of the service promptness. A user might have waited a long time at the terminal before his message is keyed in. The overall response time is the elapsed time from the instant that a user or a message arrives at a terminal to the moment the user is completely served or the reply to the message is received. The response time is usually defined by requiring average response time to be no more than X seconds, or response time for at least Y percent of the transactions to be no more than Z seconds.

System Capacity or Throughput

Capacity is sometimes taken to mean the maximum amount of traffic, in terms of transactions per second, characters per second, etc., that a system can carry. Unfortunately, such a definition is unrealistic in that a user who attempts to send messages into a system operating at this capacity would experience intolerably long delays between message input and response.

A more practical definition of capacity is the maximum traffic that a system can carry while satisfying the blocking probability or response time criteria.

Network Reliability

While the failure rates, mean time to failure (MTTF), and mean time to repair (MTTR) of the equipment and lines are often beyond the control of network planners, the network's reliability can usually be strengthened with proper network structures. Again, the definition of reliability varies according to usage. For most purposes, at least one of the following definitions is applicable:

A Digest of the Latest Communications System Planning Techniques

1. Percentage of time a terminal can communicate with the central computer.

2. Percentage of time an office can communicate with the central computer. (The office can have more than one terminal.)

3. Percentage of time a terminal can communicate with any other terminal. (This definition applies when there is direct inter-terminal communications.)

4. Percentage of time an office can communicate with any other office.

5. Average number of terminals or offices connected to the network.

6. Average number of equipment failures or the average man hours required for repair during a day or other unit of time. (This definition is useful for equipment maintenance crews.)

Sensitivity

A network can behave properly when the traffic volume is within the projected range but then break down entirely and disastrously if the traffic exceeds the volume for which the system was designed. Thus, a good planner must be concerned with the effects on the system of having the actual traffic above the projection. He should make sure that a small variation in the projection does not create intolerable response times or blocking probability. He should create a curve like the one shown in Figure 6 (early reference) during the planning process.

Transmission Error Rate

The transmission error rate is a function of message size, line conditioning, and hardware characteristics. It is more critical in a centralized data communications environment than in a distributed computer network environment because the centralized environment does not usually have error detection or correction features but the decentralized environment does. For example, suppose the error rate is not allowed to be more than one in every thousand characters. The planner should calculate the error rate for his candidate system or network. If it is less than 0.001, he is safe. If it is not, he has two choices. One is to add error detection and correction features to the terminals or terminal controllers. The other is to redesign the network, i.e., use lower speed modems, use fewer terminals per multidrop line, etc.

Traffic bottleneck

After a network has been designed to satisfy specified traffic requirements, the traffic bottleneck of the network is not immediately a characteristic of interest. However, if the network has to be upgraded to handle more traffic in the future, knowledge of the traffic bottleneck can help in estimating the incremental cost of expanding network traffic handling capacity. For example, if the bottleneck is the high speed line between a concentrator and the central computer, it is quite simple to upgrade the network capacity, either by adding a line or by using modems of higher speed. On the other hand, if the bottleneck is at the central computer, the upgrading would be costly.

Communication Device Selection

Communication devices are used in data computer communication networks to interface the data terminal equipment (DTE) with the transmission facilities, to convert the data signals to signals that can be handled by the transmission facilities and vice versa, to the reduce cost of the transmission facilities, to relieve the load on the main computers by handling the communications overhead, or to switch and control communication. Some of the devices available to do these various jobs are as follows:

1. Interfacing with Transmission Facilities—Modems, Baseband Modems, Access Controller

2. Saving Communication Costs—Multiplexer, Concentrator

3. Reducing Processing Load—Front-end processors, Concentrators, Terminal Control Units

4. Switching and Communication Control—Ring Switches, Packet Switches, Message Switches

With advances in solid state electronics, communications devices are more versatile and generous with options. Numerous possible combinations are available for improving performance, reducing communications costs, and satisfying special requirements. However, these goals are not easily achievable. It is important to know also what a vendor has not said, what devices are most effective for specific network structures or performance requirements, and how many of each particular device should be used.

A high speed modem can increase throughput, improve response time, and might even reduce costs, but the transmission error is much higher on a line without a higher speed modem. Sometimes, this high error rate prohibits the use of high speed modems with nonintelligent terminals. A modem sharing unit can save costs in a system if several terminals can be located in one office and can share the same multidrop line, but network reliability will be lowered. On the other hand,

A Digest of the Latest Communications System Planning Techniques

a combination of modem sharing units and port or line sharing units can improve reliability as well as reduce costs. A multiplexer or a concentrator can reduce overall communications costs under some traffic and operational environments, but without proper planning, the introduction of multiplexers or concentrators will degrade network performance and could even increase the costs.

TRANSMISSION FACILITY SELECTION

All data communication networks need transmission lines. The planner must choose between dial-up and dedicated lines, must choose the right line speed, and must calculate and compare line costs.

In the dial-up case, a terminal is connected to the network only when there is a need for actual communication. In the dedicated line case, a terminal is always connected to the network via a leased or private line even if no transmission is taking place. Dial-up connections are used when terminal locations are not fixed (like traveling salesmen's portable terminals), when terminals do not belong to the same organization (like in a time-sharing environment), when terminals are used in a remote batch environment with low utilization, or when terminals are sparsely located with low utilization. Presently, the dial-up arrangement is either by direct distance dialing (DDD), foreign exchange, or WATS lines.

The choice of terminal speed depends mainly on the application, performance requirements, available line tariffs, and cost. It can be determined by rules of thumb, experience, or standard speeds of terminals for the specific application. A simulation program can be used to determine response time, given specific line speeds, to see whether a specific speed satisfies a given requirement and to evaluate cost/performance tradeoffs using different line speeds. In the same network, different terminals and communication devices can be connected to lines with different speeds.

Transmission facilities are commonly obtained from AT&T (or PTT in Europe). However, sometimes less expensive, albeit less convenient, alternatives are also available. These include private facilities (wires, coaxial cables, microwave facilities, radio waves, lightwaves), CATV (Cable Television) facilities, and special common carriers (satellite companies, public packet switching networks.)

NETWORK ARCHITECTURES

Network costs and performance depend greatly on the structure the planner chooses. In general, there is no easy way to determine the best structure. Repetitive simulation and design cycles are required. The following are the most commonly used network structures:

1. Point-to-point connection via dial-up.

2. Point-to-point connection with leased lines.

3. Multipoint tree-structured connection (Figure 1). The tree structure usually terminates at a multiplexer, concentrator, message switching processor or central computer. In general, the tree network and terminals are connected so that when one device is transmitting a message, the message is also transmitted to all the other devices connected in the tree. (This is in contrast with the ring structured network.) However, only the addressed devices would actually receive the message. (Figure 2).

4. Ring structured connection (Figure 2). In this structure, terminals or devices form a ring or loop. When terminal A is sending a bit to terminal B, no other terminal knows anything about this bit. (This is in contrast with the tree network.) If this bit information is not for B, B then passes it to C, and so on. At least one of the devices on the ring is a computer which is the central computer or a message switching computer capable of switching a message from one ring to another.

5. Multiplexed structure. Several low speed point-to-point lines or multipoint tree structured lines are terminated at a multiplexer; the multiplexer is then

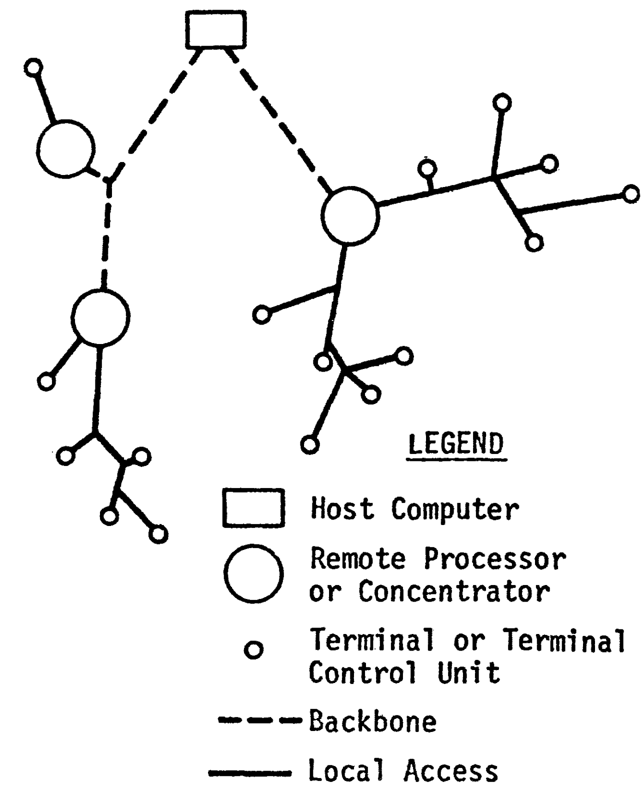


Figure 1. Tree-shaped hierarchically controlled data communication network architecture

A Digest of the Latest Communications System Planning Techniques

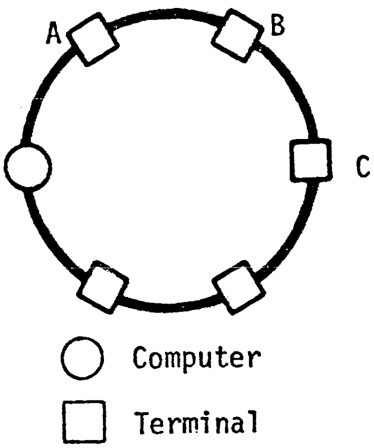


Figure 2. Ring architecture

connected to a distant multiplexer through a high speed line. By this arrangement, every low speed line is connected to the distant site as if it had its own line, but the line is actually shared.

6. Store-and-forward or packet switching structure (Figure 3). In a packet switching structure terminals or computers are connected to a packet switch (PS), and the PS's are interconnected according to traffic requirements. Usually the interconnection has a mesh shape, but theoretically, it can have any shape. If messages from the computers or terminals are too long, they are broken into small pieces, called packets, before transmission. (This is the reason for the term packet switching.) Frames and segments are terms that have also been used. When a PS receives a packet, it first stores the packet in the main memory. After analyzing the destination of the packet, the PS routes the packet to an output line for transmission to the next PS along the path it calculates is the fastest path for the packet to reach its destination.

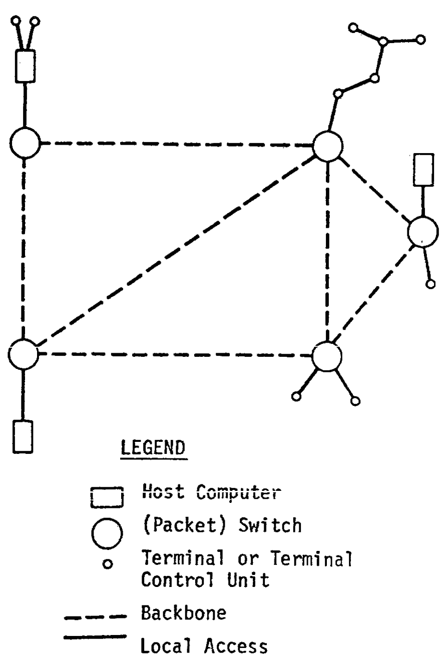


Figure 3. Packet Switching Architecture

7. Single-channel-multiple-access structure. In this structure, devices are connected to a common wideband channel. They communicate with each other by contending for the common channel according to a multiple access scheme. The wideband channel can be cable, satellite channel, or ground radiowaves.

8. Hierarchical structure. Terminals have local access connections to concentrators or message switching computers. These computers are then connected either directly to a central computer or to one another as the backbone network. Figure 3 shows an example of this kind of structure in a packet switching context.

DESIGN TECHNIQUES AND TOOLS

Techniques that can be used for design, optimization, and performance evaluations include mathematical topological optimization algorithms; heuristic topological optimization algorithms; analytic models and formulas, such as queuing formulas; simulations of analytic models; descriptive simulations; and regression analysis. Hand analysis and design becomes almost impossible because of the complexity of many data communications networks. Computer programs for various analysis and design functions are essential. The usefulness of such programs relies heavily on how inexpensive and convenient it is to run the programs repetitively. Thus, efficiency in running time is as important as accuracy. The following computer programs are very useful tools for the design of data communication networks. Depending on the complexity of the network and the generality intended, these programs can range from very simple to quite complex. The more simple programs can even be implemented on a handheld calculator. The more complex programs might need 200-380K bytes of computer memory.

Contention Analysis Simulation: Determining blocking probability.

This program is useful for dial-up terminals. The output of the program should give the distribution of blocking probabilities for each of the following: number of users (terminals), number of ports, distribution of number of calls per unit time at each terminal, call holding time distribution at each terminal, and distribution of time intervals between two consecutive attempts to obtain a channel for the same call.

Central Processor System Configuration Program: Verifying performance for specified CPU configurations

A Digest of the Latest Communications System Planning Techniques

For specified CPU and peripheral device types, the program should indicate whether the specified configuration satisfies throughput or response time requirements. If the requirements are met, the total time span spent by a message in the CPU and peripheral devices will be given. If they are not met, the bottleneck causing the oversaturation will be indicated. Analytic queuing models, as opposed to brute force simulation and analytic closed form formulas, should be used to develop this module. Brute force simulation of the CPU is too complicated and too time consuming with respect to both development and execution.

Network Simulation Program: Simulation of the whole data communications system.

Given a network configuration and traffic requirements, the module should be capable of supplying terminal response time statistics to provide a response time/throughput relationship as a function of work configuration. This is done by simulating a whole system, including regular and intelligent terminals, multidrop lines, concentrators, trunk lines, and CPU's. For effectiveness and efficiency, simulation, analytic formulas, queuing models, and empirical distributions should be judiciously mixed into the program. Specifically, intelligent terminals, concentrators, and CPU's are to be simulated by queuing models or are to be described by empirical distributions. Line and terminals are to be simulated.

Network and Design Program: Concentrator and multiplexer allocation, terminal clustering, multidrop line topological design, and economical analysis.

Given the locations of terminals, and CPU's with their basic characteristics; with the traffic characteristics, and with the line utilization requirements, multiplexer and concentrator locations are then selected. Terminals are cost effectively connected to the proper concentrator or CPU via a multidrop line. Thus, an important design goal is to implement optional heuristic algorithms in the program so that the program size and running time can be proportional to the number of terminals. (In general, the size and running time grow quadratically or as the cube of the number of terminals.)

Network Reliability/Availability Program: Calculation of network reliability.

Given element failure rates (or MTTF's and MTTR's), the module calculates network reliability criteria. A combination of simulation and analytic techniques should be used to ensure the effective determination of reliability for networks with thousands of terminals within reasonable computer time.

DESIGN TRADEOFFS

It is the network planner's responsibility to assist users in defining their requirements and to design a least cost network while satisfying the requirements. To do a good job for a large network, the planner needs to develop a set of curves to weigh and compare the tradeoffs for cost/performance and for design alternatives. The design tools described in the preceding paragraphs are extremely useful for this purpose.

Evaluation of Design Alternatives

By choosing some of the network structures given earlier and by using different line speeds, the planner can develop several sets of curves as shown in Figure 4 and 5. For a specified response time requirement, each curve in Figure 4 represents the cost/throughput relation for a specific network structure. For a specified network throughput, each curve in Figure 5 shows the cost/response time relationship for a specific network structure. From these curves, the planner can determine the most cost-effective network structure. Figure 4 indicates that the choice of the least cost network structure depends on the throughput requirement.

Evaluation of Cost/Throughput Tradeoffs

Figure 6 shows the relationship between cost and throughput for a fixed response time requirement. With these curves, the network planner can help users to decide how much they are willing to, or must, pay for throughput in the network.

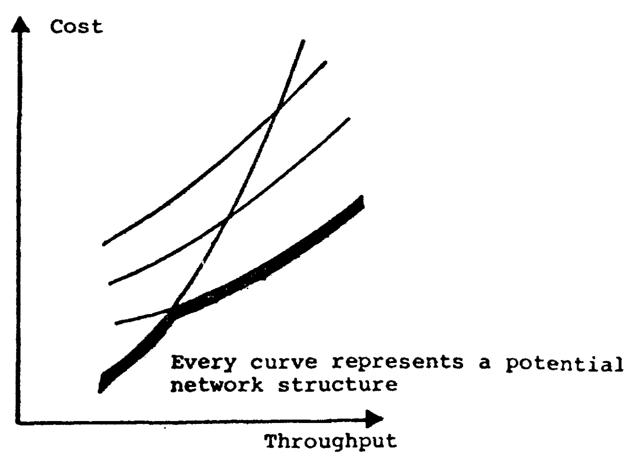


Figure 4. Cost vs. throughput for a fixed response time (useful for evaluation of network structure alternatives)

A Digest of the Latest Communications System Planning Techniques

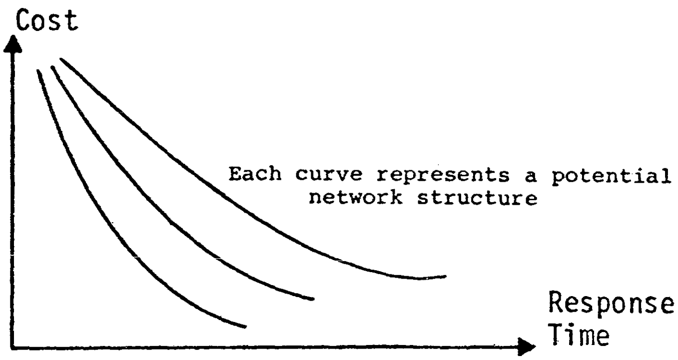


Figure 5. Cost vs. response time for a given throughput requirement (useful for evaluation of network structure alternatives)

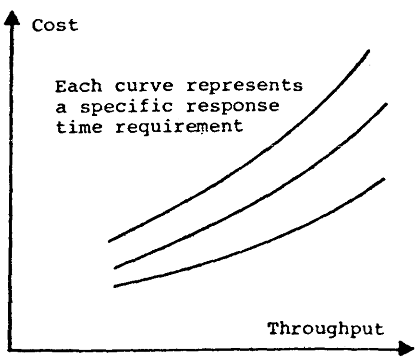


Figure 6. Cost vs. throughput with response time as a parameter (useful for evaluation of cost/throughput tradeoffs)

Evaluation of Cost/Response Time Tradeoffs: (CPU response time, network response time, or CPU and network response time).

In Figure 7, each curve represents the cost/response relationship for each specified throughput requirement. These curves help the user to determine how much he is willing to pay for the response time that he will get and help the planner to determine the best combination of CPU configuration and line configuration.

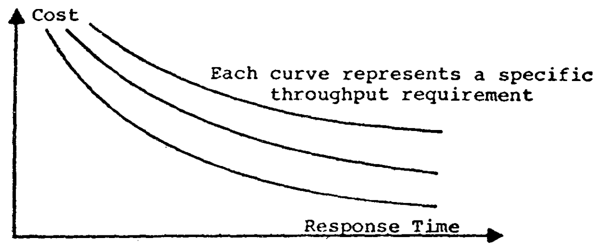


Figure 7. Cost vs. response time with throughput as a parameter (useful for evaluation of cost/response time tradeoffs)

Evaluation of Cost/Reliability Tradeoffs

Reliability is another term that users do not quite know how to define. There may be many schemes for improving network reliability. To evaluate them, the planner must develop a curve to show the incremental cost for improved network reliability.

Derivation of Response Time/Throughput Relationship

Since traffic estimates are rarely accurate and future growth is even harder to predict, network planners need to know how sensitive network performance is to traffic variations from projected levels. Figure 8 shows the relationship between performance and throughput with performance measured in terms of response time. □

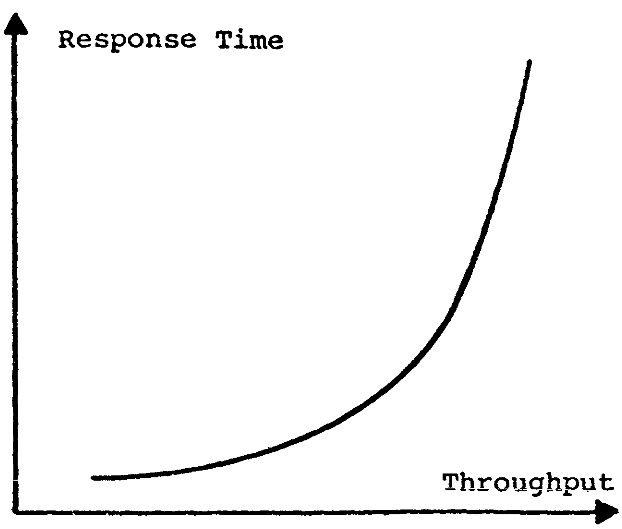


Figure 8. Response time vs. throughput (useful for performance sensitivity evaluation)

Gaining a Better Perspective of Communications System Applications

Problem:

It is sometimes difficult to recognize that a communications system has two distinct aspects. The first aspect is very specific. The system is designed to solve particular problems—generally to eliminate bottlenecks in the way information flows throughout a company and between the company and the outside world. The second aspect is very general. A more efficient communications system offers a framework for other applications that frequently had no direct relationship to the original problem-solving purpose of the system simply because the new system substantially improves the quality and quantity of communications linkages among men and machines. Unfortunately, many of these related but indirect applications are difficult to perceive in the planning phase of a system because the planners are invariably too caught up in the urgency of finding quick solutions to pressing problems. Potential applications are thus overlooked until after the system is designed and built, at which time it may be too costly to modify the system to accommodate the applications.

This report is offered to enlarge your applications perspective at the very outset of systems planning. It describes seven significant applications areas and demonstrates how much each application has in common relative to a communications system. This report will help you to see how certain planning steps taken now could make all the difference between "too costly" and "feasible" when it comes time to consider new applications for your fully developed system.

Solution:

We explore the range of applications of data communications networks by considering some typical examples in greater detail. The following industries will be considered:

- Banking and finance
- Insurance
- Retailing
- Government
- Health care
- Manufacturing
- Telephony

"Data Communication Application from *Distributed Processing and Data Communications* by Daniel R. McGlynn. Chpt. 8, pp. 219-252 © 1978 John Wiley and Sons. Reprinted by permission.

BANKING AND FINANCE

Banking applications of data processing are more intensively concerned with data than many other industries and illustrate the gradual evolution from batch processing to on-line processing and distributed processing through data communications. Moreover, the increased interest in the implementation of electronic funds transfer (EFT) through telecommunications networks is an even more significant area.

On-Line Processing

A typical on-line processing system is the IBM 3600 Finance Communication System. The IBM 3600 interfaces with a System/370 in an integrated network that permits tellers in remote locations to enter, update, and retrieve customer data and financial information. The IBM 3600 consists of several components:

- IBM 3601 finance communication controller
- IBM 3604 keyboard display terminal
- IBM 3610 document printer
- IBM 3612 passbook and document printer
- IBM 3618 administrative line printer
- IBM 3614 consumer transaction facility, or self-service teller terminal

In operation, the teller enters the transaction information on the 3604 keyboard display. The information entered is sent to the 3601 controller, which automatically performs any required computations. The customer's passbook is then inserted into the 3612 passbook and document printer, and the controller supervises the printing of the updated information in the passbook. The 3601 also relays the transaction to the host computer so as to immediately revise the centralized files, thereby enabling any teller at any remote station to obtain the updated customer data.

The 3618 administrative line printer is used to obtain running totals or balances for all transactions for management report or auditing purposes.

Distributed Processing

The next stage in the evolution of data processing in the banking and financial industries is the transition from on-line processing to distributed processing and decentralized processing. Such a transition is only possible if the organizational environment makes such processing advantageous over traditional centralized processing.

The banking and financial industry is one example where the operational environment of branch offices makes distributed processing feasible. Data entry, verification, processing, hard-copy output for the customer, internal records, and auditing purposes may all be economically performed at the branch office level rather than through a centralized location. The branch offices may be "loosely coupled" in a data communications network so that a customer may make use of a branch office that is not identified as his own branch office for certain transactions. The branch offices may also be interconnected to a supervisory office, which periodically collects pertinent summary data from the branch offices for management purposes.

Since the transition from an on-line centralized data processing system to a fully decentralized system requires significant organizational and operational coordination and planning, we review the experience of one such financial institution making such a transition—Citibank N.A., in New York.

CITIBANK DECENTRALIZED DATA PROCESSING NETWORK. The Citibank decentralized data processing network is a reflection of the bank's decentralized management policy rather than any single network architecture philosophy. Each operational unit is responsible for its own activities, including all data processing operations associated with its operations. Since the management within that operational unit is most familiar with the data characteristics and processing requirements for its own activities, it therefore selects its own computer configuration most suited to its needs. Although such processing is done independently of central corporate management, financial control and management auditing and review are achieved by interacting with the computer system of the operational unit over a communications network.

The most important consequence of the decentralized management concept is the complete decentralization of the bank's database and the gradual downgrading of the centralized database and computer processing facilities. The comparative efficiency of a single large-scale, multiprogrammed, multiprocessing computer system compared to a heterogeneous mix of smaller computers at remote locations was carefully considered before the bank committed itself to decentralized processing. Some of the key considerations in evaluating the comparative efficiency of centralized and decentralized should be noted.

First, the bank observed that, to be effective, large-scale, centralized data processing systems require a high level of operational knowledge of the business. The management of a large-scale data processing operation is practically a business in itself, and the experienced professionals in management of data processing operations generally do not have the business understanding to effectively design the computer configuration to satisfy the user's requirements.

Stated in other terms, the data processing managers of large data processing systems may be able to produce highly efficient computation, measured in

Gaining a Better Perspective of Communications System Applications

terms of throughput, virtual memory, optimized compilers, and performance monitoring, but may not always be able to supply the specific information that is valuable to the bank's operation managers in an efficient fashion.

Of course, the experience of Citibank in evaluating the efficiency and effectiveness of its centralized data processing facility is not necessarily applicable to other organizations or even to other banks. Indeed, for many organizations the economics of scale and centralized management of a single large-scale data processing operation could possibly be more efficient than a distributed or decentralized data processing network. However, very large organizations (i.e., those using several large-scale computer systems) may well consider whether several smaller data processing operations may be managed more effectively and provide more useful information to the end user than a single large-scale, centralized data processing operation.

The second consideration of importance to Citibank concerns the logistics of system planning, development, and implementation. The planning and development of large-scale computer installations, or the upgrading of existing installations, required much longer lead times than was desirable. The bank was unable to rapidly respond to changing customer requirements and found that when system designs were finally implemented they were no longer suited to user requirements.

In shart, large-scale computer systems require much longer planning, proposal, and delivery schedules than the rapidly changing banking business justified. Once such systems were installed, they were not flexible enough to adapt to changing market conditions and consumer preferences.

The third consideration in evaluating the comparative efficiency of centralized and decentralized data processing operations concerned the installation and facilities management of large-scale computer systems. The environmental requirements of back-up power, air conditioning, raised floors, and security devices all contribute to the complexity and decreased effective efficiency of centralized data processing operations.

The final consideration was the increased technical complexity of large-scale computer facilities. Business problems and user requirements were not simply translatable into the hardware/software capabilities of complex computer systems. The business managers did not speak the jargon of priority queues and partitions, and the systems analysts could not always relate such technical issues to the basic user needs.

The Citibank decentralized data processing network is implemented primarily with minicomputers of various manufacturers, including Digital Equipment Corporation (DEC), Interdata, Data General, and Hewlett-Packard. Over 50 such minicomputers were installed in 1975, and over 100 were installed in 1976.

Citibank's minicomputers operate in different corporate divisions, performing different applications and are not intended to routinely communicate with each other. The heterogenous makeup of the network is a consequence of the decentralized management decision to implement a particular computer system that is most efficient and effective for its particular operation, regardless of the types of computers used by other divisions or operations. Compatibility is not necessary since the computers do not share a common database, and information between divisions or operations is exchanged on a special request basis, rather than through an automatic procedure. Financial control and management review is achieved at present by magnetic tape output or a simple data communications transfer to another processing facility.

One example of such a system is the letter-of-credit channel. Four CRT terminals are used as input devices, while the processor supervises the issuance, amendment, and payment of commercial letters of credit. The processor maintains its own database of letters of credit outstanding, and hands off on magnetic tape the pertinent information for the customer accounting system and the loan liability system for account and loan liability updating.

Another example is the Lockbox Automation System, which utilizes an Intel 8080 microprocessor controlling a display and a Burroughs Check Encoder. The Lockbox system captures the customer check information and amount at the same time the check is being encoded and updates the customer files by a magnetic tape handoff to the customer accounting system.

Both of these systems are implemented on a DEC PDP 11/70 configured with two tape drives, three 88-megabyte disk drives, and two 300-line/minute printers. Operators interface with the system through some 64 communications ports and 60 terminals.

Other on-line applications at Citibank include:

- Foreign bank accounts
- Foreign exchange processing
- Commercial loan data entry
- Domestic money transfer
- International money transfer

- Federal Reserve Bank interface automation
- Automation loan processing
- Clearing House interface automation
- Corporate accounting and information

The decentralized data processing applications described above will eventually be tied together through a data communications network, thereby automating the financial control and management review functions now performed by the physical transport of magnetic tape financial summary listings. Such a network would use a general line control procedure so that machines from different manufacturers in the network would be able to communicate with each other.

Another application of such a data communications network would be for direct customer communications with the bank. Large corporate customers would have a terminal for entering transactional information into the network. The network will connect to a Transaction Processing Computer, which will handle the transaction and perform the routine accounting and general advice function. Finally, account and MIS data will be sent to the head office for generating internal reports.

Electronic Funds Transfer

Electronic Funds Transfer (EFT) refers to the concept of a checkless, cashless transaction mechanism by which funds are transferred automatically by electronic means. EFT is an extremely controversial and complex issue in data communications, and it would be worthwhile to consider some of the issues in greater detail. The following basic issues are considered in the sections below:

- EFT background
- Economic issues
- User concerns
- Regulatory and consumer interests

EFT BACKGROUND. There are many different types of EFT systems, depending on the application and environment. These include:

1. Preauthorized banking services, which provide for the automatic deposit of a payroll or Social Security funds in a designated account or the automatic debiting of mortgage or loan payments from a designated account.

2. Customer-directed payment systems, which enable a bank customer to utilize a telephone or EFT terminal to direct the bank to pay funds to a designated account.

3. Point-of-sale systems, which use a terminal on the premises of a merchant that enables the customer to authorize payment from his account to the merchant's.

One could also include as EFT systems the following:

4. Credit verification/authorization systems, which enable a merchant to check the credit of a customer by automatic query of a centralized bank or credit card company computer.

5. Automated banking facilities, such as automatic cash or travelers' check dispensers providing 24-hour service to the customer.

EFT must be considered as an evolutionary step in society's payment media. As technology changes, and as the nature and cost of transaction processing change, there has been an increased interest among large-scale transaction-processing organizations, such as the government and financial institutions, to use the new technology to improve the transaction processing mechanism. Some of the early experiments in EFT were also carried out by smaller financial institutions such as the Hempstead Bank on Long Island, New York and the First Federal Savings and Loan Association of Lincoln, Nebraska in the early 1970s.

ECONOMIC ISSUES. The basic impetus to EFT is an economic one. Present-day transaction media consist of cash, checks, and credit cards. The costs associated with such media are very high:

- Currency ($71 billion in circulation)—an annual cost of over $4 billion for printing, security, and so forth
- Checks (32 billion clearing annually)—an annual cost of over $7 billion in processing
- Credit cards (over 6.5 billion transactions, involving $85 billion)—an annual cost of $3.9 billion

The total cost of all funds transfer is over $22 billion for handling some 325 billion transactions annually (i.e., a cost of over 6 cents per transaction). Such costs are expended for personnel, document processing equipment, losses from forgery or fraud, security and protection, and so on.

Gaining a Better Perspective of Communications System Applications

It is difficult to compare the cost of electronic funds transaction systems with present-day systems: there are simply too many variables that affect the cost an order of magnitude or more. However, with standardization, widespread market acceptance, and high volume use, EFT costs measured over suitable periods (e.g., 10 years) can be very attractive.

In the short term, however, EFT offers a number of distinct features:

1. Simplicity of establishing new EFT outlets or terminals

2. Increased merchant competitiveness by offering consumers a new, more convenient payment technique

3. Minimized merchant handling costs and credit losses

4. Increasingly lower cost of processing technology (microprocessors and related components).

USER CONCERNS. One of the major problems in EFT is the competitive implication of such systems and the attendant user concerns. We can identify two major competitive areas:

1. Competition among financial institutions

2. Competition between financial institutions and merchants

EFT systems are a high-risk, high-capital-investment item. Large financial institutions are in a much better position to enter the field and experiment with EFT systems than small ones. Furthermore, EFT systems are highly visible to the customer, and the intense competition for the retail depositers makes the availability of EFT systems an important competitive factor. EFT complicates the competitive balance between different financial institutions: large versus small institutions, urban versus rural and suburban institutions, and commercial banks versus savings banks and other depository institutions.

EFT systems open up a new prospect of competition between financial institutions and merchants. As in many other interdisciplinary and interindustry activities, the basic question is: Where does the "banking" function end and the "retailing" function begin? Associated with one industry or the other are unique regulatory, labor, and legal interests and customs that cannot be assumed by the other.

In addition to competitive issues, EFT raises significant technical and administrative issues that must be answered:

- Equipment compatibility
- Code and protocol standardization
- Security and system integrity
- Compatibility with related systems (e.g., UPC)
- Ownership of systems
- Supervision and control of systems

REGULATORY ISSUES. The regulatory and legal issues may not be the most significant, but they are clearly issues that must be settled before EFT systems can be implemented.

There are many forms of regulation that relate to EFT, ranging from state and Federal regulation of banks and the legal infrastructure of commerce such as the Uniform Commercial Code (UCC), to consumer law. Fundamental changes must take place in all these areas before total implementation of EFT can be possible.

The National Conference of Commissioners on Uniform State Laws is moving forward in one area—automated stock transfers—by amending the relevant statutes to allow corporations to establish and maintain stock ownership data by computerized records rather than the issuance of engraved certificates.

Already, new mutual funds—for example, cash management and municipal bond funds—do not routinely issue certificates, but instead send their shareholders a monthly or quarterly statement of their shareholdings. Such recordkeeping particularly simplifies automatic dividend reinvestment procedures.

There are of course some who oppose such automated reform. The cost of manual stock transfer is high—several dollars per transfer—and such costs mean significant revenues for banks or stock transfer companies. New regulations may also impose periodic shareholder notification requirements that may be burdensome on companies with many small shareholders or infrequently traded securities.

Discussion of consumer law aspects of EFT is also important. Some of these issues are:

1. Recordkeeping and account information: when is written documentation to be supplied to the user concerning an EFT transaction and what minimum information should be contained therein?

2. Error control: what means should be established to correct for errors, or provide recourse similar to a stop payment order? What effect does EFT have on well-established legal doctrines such as the "holder in due course" (UCC 3-305), which provides that the holder in due course of an instrument (such as a check or its EFT equivalent) takes the instrument free from all claims to it on the part of any person?

Gaining a Better Perspective of Communications System Applications

3. Standardization: does EFT require standardization of technology, coding, use, and disclosure, and what degree of regulation is appropriate or sufficient?

INSURANCE

Insurance applications of data communications networks are quite different from banking and financial applications. Such applications are concerned with the preparation and processing of individual insurance policies at agency locations throughout the country. It must be emphasized that, unlike banking transactions, the preparation of insurance policies are tailored to the individual background and requirements of the customer. The data communications network could be used to implement an on-site preparation of insurance policies at remote agency locations by interacting with the centralized files at the host computer location.

In the implementation so described, the independent or company agent enters the required customer information on a terminal in his local office. The centralized host computer processes the application by examining claims history and other characteristics, and calculates the premium. The actual policy itself may be printed out on an on-line printer associated with the terminal, so that copies of the policy are supplied to both the agent and the customer for signature on the spot. Meanwhile, the host computer establishes a billing cycle for the customer, and credits the local agent with the sale. Renewal notices are also generated automatically.

RETAILING

Retailing is one of the most significant applications for data communications networks and specialized terminals. The data generated by department stores and mass merchandisers in the ordering, handling, inventorying, and selling of goods is extremely important for the efficient management of transactions involving vast quantities of assorted merchandise. More importantly, the analysis and interpretation of such data enables managers to react even more quickly to consumers' demand patterns.

One typical retailing data processing system is the IBM 3650. The individual components of 3650 system include:

- IBM 3651 store controller, which controls all other devices in a store network and communicates directly with an IBM System/370 host computer
- IBM 3653 point-of-sale terminal, which includes a sales-slip printer, customer and operator guidance displays, a data entry keyboard, and an optional "wand" for reading magnetically encoded data
- IBM 3657 ticket unit, which prints, encodes, and reads information from price tags
- IBM 3275 display station and 3284 printer, which attaches to the 3651 store controller for administrative message handling or special customer services

By integrating both purchasing and selling activities into a single communications loop, a distributed processing network for a retailing operation makes possible a more coordinated, efficient flow of merchandise through the firm. A narrative review of how such a system would be implemented in a retailing environment is worthwile discussing here. A more detailed discussion of a "management information system" will be postponed to the section on manufacturing applications, although it must be realized that a management information system could also be implemented in the retailing industry.

The first step of the retailing cycle processed by the system is the buyer's original purchase order, which may be entered at a distribution center, store, or administrative headquarters through the 3275 terminal, for example. The record of the purchase order is entered into the host computer storage, and the original copy is printed out at the 3284 printer for mailing to the vendor.

When the shipment arrives at the receiving dock, the foreman utilizes the 3275 terminal to access the host computer storage. The purchase order is then displayed on his screen, enabling him to check it against the shipper's bill of lading. If the merchandise is in order, he accepts it by making appropriate entries on the 3275 terminal. The use of a remote terminal or distributed processor for such functions is often called a "point-of-purchase" system.

The host computer updates the outstanding purchase order file, updates the appropriate inventory file, records the amount due the vendor on the accounts payable record, and alerts the ticketing file that price tickets should be prepared for the received merchandise.

When the merchandise is received at the ticketing station, the 3657 ticket unit prints out the appropriate price tag, which also indicates merchandise class, size, color, or model for inventory control purposes.

The final step of the retailing cycle is the "point-of-sale," at which the customer brings the merchandise to the cashier. Data from the price tag is either entered manually on the keyboard of the 3653, or read by the optional wand. The inventory control file is updated, and the printer produces a receipt and journal tape for auditing purposes. Special services, such as credit sales, back-ordering of a particular size or color, and similar requests may also be implemented through the system.

Another important application of a distributed processing network is the airline reservation and ticketing system. Although such networks may be built around a large host computer, it should also be noted that they can be built around minicomputer networks. The AEROFLOT network is such an example.

The AEROFLOT network, installed by the Compagnie Generale de Construction Telephonique (C.G.C.T.) is based on 23 DS6-400 hardware/software modules built around the SPC-16/85 General Automation minicomputer. The network includes five major switching centers in Moscow, Rostov, Novosibersk, Alma-Ata, and Sverdlovsk. In addition to message switching, the centers also handle real-time file interrogation and data base management. The centers are expected to handle some 4 million messages, or about 1 billion characters, every day. Each of the processing modules is designed to handle at least eight messages per second.

GOVERNMENT

In this section a specific operational function will be discussed in greater detail—the function of patent searching as part of the statutory responsibility of the government in granting an inventor a patent—but bear in mind that many general administrative and operational functions of government may be assisted by information-handling facilities.

Patent Searching

Patent searching is the process of examining a patent application for novelty and unobviousness in view of previously published "prior art," such as previously issued patents. The patent searcher or examiner analyzes the application to determine the particular classifications of technology the application is related to, and physically "searches" the Patent Office files in those classifications for similar or related patents and other publications.

The process of patent searching is a good example of the potential application of distributed processing computer technology to an important government function. The U.S. Patent Office processes more than 100,000 patent applications per year, and the searching process is a time-consuming, manual task which limits the effectiveness and efficiency of patent examination.

The first step to a distributed processing patent searching system was taken in 1975 with the introduction of a high-speed computer controlled microform search system (CCMSS).

The CCMSS is not a true on-line system, but it demonstrates the capabilities of such a system in the user environment. Every issued patent and many important publications are classified into one or more technological classes or subclasses. In performing a search, the searcher considers the various classes that the application is related to, and searches both classes. If the invention is concerned with both classes of technology, the searcher must still search both individual classes.

The distributed processing capability of the CCMSS enables at least some portion of the searching process to be performed more efficiently through the use of logical operators, which may be specified during the search. By "logical operator" we mean a function such as AND or OR applied to two or more classes. If the searcher only wishes to examine those references in classes 250 and 283, for example, which are classified in both class 250 and class 283, he may enter the search command "250 and 283" on the terminal console. The system then performs the logical AND operation on all elements of classes 250 and 283, and stores the reference numbers of those which are common to both classes. The searcher may then enter a DISPLAY function on the console and sequentially examine each of the references.

HEALTH CARE

The application of distributed processing and data communications networks to health care delivery systems is expected to be a much more significant use of data processing technology than the suggested "remote medical literature searching," "remote medical diagnosis," and "electrocardiogram analysis" projected for 1990. Although ideally the world of 1990 should be able to allocate more of its resources to medical research than the present, it is very doubtful that such resources be extensively applied to medical literature searching as opposed to basic or clinical research.

The application of distributed processing to hospital and medical services is similar to other "delivery systems," in which billing for a wide variety of services is generated, with specific items being prorated for third-party billing. Futhermore, extensive documentation and reporting to federal and local agencies must be supported.

Gaining a Better Perspective of Communications System Applications

Remote Medical Diagnosis

The prospect of remote medical diagnosis through the use of data communications facilities is significant where diagnosis can be performed at a location remote from the patient. One actual example of remote medical diagnosis and data processing is the use of computed tomography x-ray equipment at different locations connected by a data link.

A computed tomography (CT) scanner is an x-ray device that records several projected views through a patient's cross-section, and then mathematically "reconstructs" the cross-sectional image on a television screen using a computer.

CT scanners are particularly suited for implementing remote diagnosis because of the high cost of the equipment, the distinct types of operations needed to be performed by the equipment, and the digital nature of the processing itself. As a high-cost item (costing upwards of $500,000), not all hospitals would be justified in acquiring such equipment based on their potential patient load. However, because of the modular nature of the equipment, several hospitals over a given region may have sufficient patient load to justify a "distributed" arrangement of CT scanning gantries, connected by data links to a centralized processing facility. The processing facility does the mathematical reconstruction of the image, copies are made for the referring physician or hospital, and more detailed analysis and diagnosis can be made at the remote location.

MANUFACTURING

The application of distributed processing and data communications in the manufacturing environment is similar to that in the retailing industry; however, the different data inputs and processing variables presents a more sophisticated processing task.

One simple application has been to link sales offices with manufacturing plants through the use of intelligent terminals and a data communications network. The salesmen can immediately inquire about the production schedule of the particular unit ordered, so that the customer knows exactly when he can expect to receive the goods.

The salesmen can also use the terminal and network to his advantage to provide special customer services. If, for example, another model of the desired product is available from a different manufacturing plant, or is available from stock at a different location, the salesman would be able to so advise his customer.

When the goods are ordered, the transaction is entered through the terminal, thereby generating a customer invoice within 24 hours. The system deletes the goods ordered from the master inventory file and credits the salesman with the order.

A wide variety of intelligent terminals can be used in a data communications network in the manufacturing environment. The IBM 3770 Data Communication System is just one example of a family of operator-oriented remote terminals that connect to a host computer. The IBM 3770 family includes various types of terminals with different keyboard, printer, and on-line storage combinations.

These intelligent terminals are connected to the host computer through a communications processor, such as the 3704 or 3705 in the case of the IBM 3770 system. The host computer maintains the master inventory file, as well as records of the individual salesmen orders.

Another more sophisticated application of distributed processing is in a management information system (MIS). The MIS is the most detailed and comprehensive data processing system that could be implemented in a manufacturing industry. As the name implies, an MIS is concerned with information that can be presented to management on demand.

MIS should not be confused with a management decision system. A management decision system is one that takes given information and performs specific decision-making algorithms on that information to derive characteristics or parameters concerning such information that is useful for management decision-making. A simple example of management decision-making algorithms in the area of financial analysis is the calculation of discounted cash flow (DCF) and return on investment (ROI).

An MIS, on the other hand, is directed at the initial task of providing the user—a designated manager—with certain up-to-date information on demand. Once that information is displayed to the user, he may then call decision-making subroutines which will operate on that data. The essential point is the obtaining of information, not the processing of it.

TELEPHONY

Telephony may not seem at first to be a particularly significant application of data communications until two specific applications are considered:

- The use of the public telephone network for the transmission of digital information
- The gradual conversion of certain analog facilities of the public telephone network to digital facilities

Gaining a Better Perspective of Communications System Applications

Data communication presently accounts for about 15% of the public telephone network. The public telephone network, being an analog system, was never really designed for digital communications, and the demand for such services has not been fully met to complete satisfaction. However, the growth of specialized common carriers catering to data communications users has been a difficult one, and there is still no clear indication that such alternative carriers will be viable over the long run. Some estimates project the data communication usage of the public telephone network at about 30% by 1980. It must be pointed out, however, that projections of usage for specialized communication services have always been overly optimistic, both inside and outside of the Bell System and other carriers. In any event, the public telephone network is and will continue to be the most important transmission medium for data communications.

The second aspect of digital communications is the conversion of the analog facilities of the telephone network into digital facilities. The use of digital facilities for voice transmission is economic for long-haul transmission; and as the volume of such intertoll communication increases, more and more digital facilities will be implemented.

The significance of the conversion to digital facilities is that data performance is greatly enhanced through digital facilities compared with analog facilities. Analog facilities are susceptible to a variety of impairments such as amplitude and delay distortion and phase jitter that significantly affect the error rate.

The introduction of the ESS No. 4 (Electronic Switching System) in 1976 was a major step forward in the implementation of digital facilities in the common carrier telephone network in the United States. ESS No. 4, is a digital switching system as opposed to previous analog systems. ESS No. 4 is able to handle 550,000 calls per hour, also significantly more than the previous analog systems. The switching element is only one part of the telephone network, and the implementation of digital carriers is equally as important.

The Bell System in the United States already utilizes a digital short-haul carrier known as T1. The T1 system utilizes pulse code modulation (PCM). Each voice frequency channel is sampled 8000 times per second, and converted into a seven-bit PCM word (7-bit quantization). An eighth bit is added for signaling information. Twenty-four of such voice frequency channels are combined into a single time-division multiplexed frame of 193 pulse slots or bits. The 193rd bit, as shown in the representation of a single frame in Figure 1, is used for framing information. With a frame rate of 8000/second, the total bit rate of the carrier is 1.544 Mb/second (megabits per second).

The T1 carrier is only the first of a hierarchy of digital transmission systems shown in Figure 2. As shown, four T1 carriers are multiplexed through the M12 multiplier to form one T2 carrier, which operates at 6.312 Mb/second, and carries 96 voice frequency channels. Similar multiplexing is performed to form the T3 and T4 levels.

PABX

The PABX (private automated branch exchange) equipment industry gained importance in the United States after 1968, when the FCC ruled that private companies may connect their own terminal equipment to the exchange network. Although PABX systems are principally used for telephony, many PABX utilize digital encoding and switching, and could be used for data switching as well.

PABX systems may be analyzed and classified according to a number of distinct characteristics:

- Type of service
- Type of control
- Technology
- Type of switching cross-point
- Number of station lines
- Number of trunk lines

The type of service refers to the operational characteristics and features of the PABX system, such as direct inward dialing, identified outward dialing, conference calling features, and other specialized services.

The type of control refers to the use of relay logic, solid-state wired logic, or solid-state stored program control for performing the common control function of the PABX system. The type of technology refers to the type or level of technology used in the common control unit; electromechanical, or electronic (discrete circuits, integrated circuits, large-scale integrated circuits/microprocessors). The common control system may also be designed as either a time- or space-division multiplex system.

The type of switching cross-point refers to the use of crossbar, reed, or solid-state switches for making the switched connection between lines.

Gaining a Better Perspective of Communications System Applications

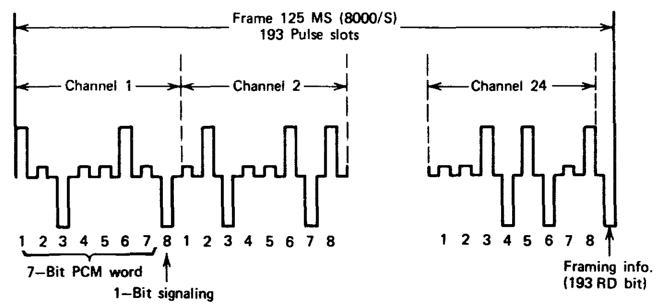


Figure 1. T1 PCM frame

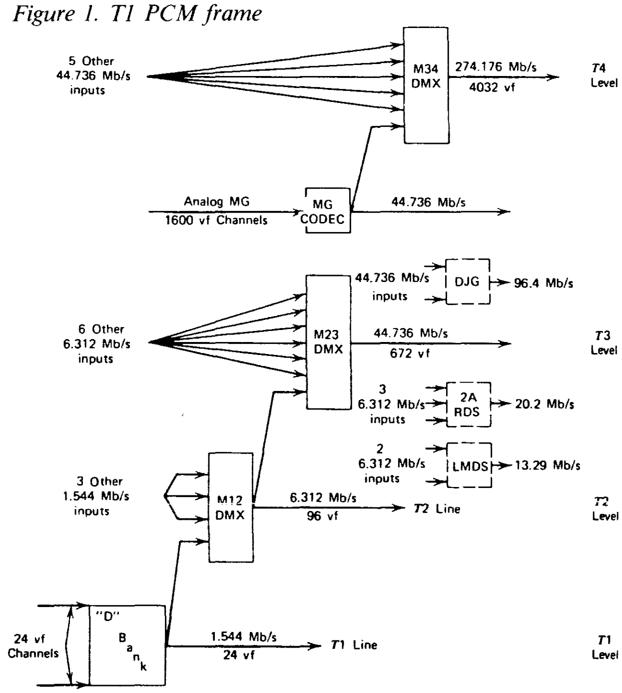


Figure 2. Digital hierarchy

The number of station lines refers to the number of extensions, while the number of trunk lines refers to the number of external lines or "trunks." Some of the features that are important in analyzing such characteristics of PABX systems include trunk-trunk connection capabilities and TDM-trunk connections.

IBM 3750

The IBM 3750 is an integrated network for voice and data communication switching presently being marketed in several European countries. The 3750, successor to the earlier 2750, was developed and manufactured in France, and announced in 1972.

The 3750 was primarily designed to offer a range of data communications capabilities operating either as a stand-alone unit or linked to a computer system for performing more advanced functions. Pushbutton telephone can be used for data collection and inquiry purpose, as well as standard telephone voice communication facilities of a conventional PABX. The IBM 3750 is configured with up to 2264 station lines and 192 trunk lines.

The voice communication features of the 3750 include facilities for three-party conference calls, access at each extension to a paging system, and direct inward dialing to extensions from the public network without operator intervention. Additionally, users can dial three-digit codes for frequently used outside telephone numbers rather than the number in full.

The system can automatically record the particulars of outgoing calls made from each extension, including the number called, caller's extension, call duration, and cost information. The system can also provide an analysis of traffic density on both trunk and internal extension lines to assist in efficient administration of the network.

For data collection, the 3750 can assemble, check, and identify each input message and store this information on its disk. This information can be transferred directly to a computer via a telephone line, or via punched paper tape. When the IBM 3750 is connected to the computer via communications lines, pushbutton telephones can be used, for example, as inquiry terminals with audio response.

Via contact sensing and operating lines, the system can monitor the condition of devices such as alarms, or the open-or-closed status of doors or valves.

The basic IBM 3750 configuration consists of an IBM 3751 control unit, IBM 3752 network and IBM 3753 line units,from 2 to 18 IBM 3755 operator desks, and an IBM 3557 keyboard-printer unit for system control. The control and switching hardware is all-electronic and incorporates solid-state monolithic technology.

The switching function and other related operations such as scanning, path marking, and signal distribution are handled by the duplexed controller. It operates under stored program control. In addition to controlling the basic switching operations, these programs control supervisory, traffic analysis, and diagnostic functions. Diagnostic routines monitor the system's own logic, automatically signaling any malfunction that might occur.

The switching system stored program for the 3750 is generated by IBM from a master program to meet the specific requirements of each individual user. Routine changes in program-controlled functions are handled by the customer. For example, the reassignment of telephone extensions can be made rapidly by updating tables in the machine memory rather than through hardware or program modifications. Only major changes—such as the addition of new

Gaining a Better Perspective of Communications System Applications

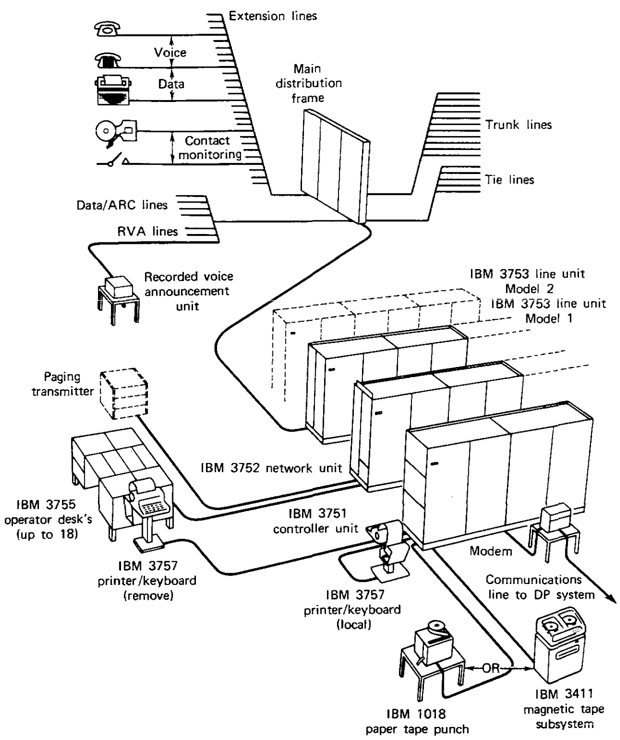


Figure 3. IBM 3750 switching system

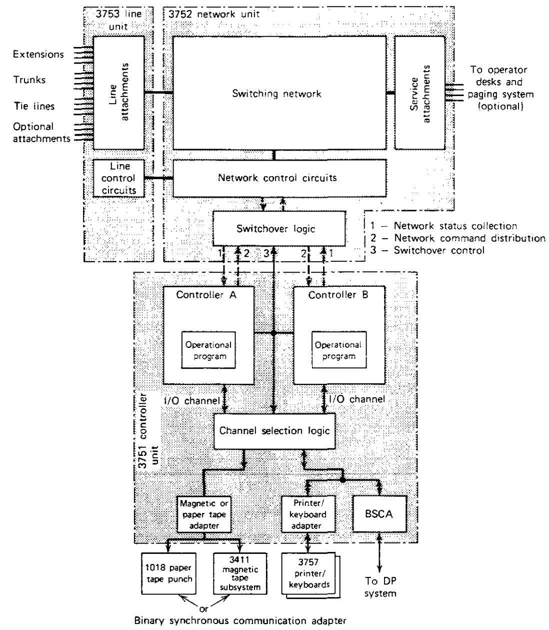


Figure 4. IBM 3750: functional organization. (Reprinted courtesy IBM World Trade Corp.)

features—require the preparation of new switching system stored programs or the installation of additional hardware. This is archived with little or no system interruption.

Figure 3 is a schematic of the basic IBM 3750 configuration.

The functional organization of the system is shown in Figure 4. Some of the services performed by the 3750 are represented by specific dialing codes which are tabulated in Figure 5.

The operators of the 3750, handle all connection and call supervision procedures that are not carried out automatically on the 3750. These functions include:

- Connecting incoming calls from the public switched network to extensions
- Making connections between extensions and tie-lines
- Performing supervisory functions

Another important terminal for use with the 3750 is the 3221 Numeric multifunction device. The 3221 has been designed as a low-activity data collection and inquiry device, and contains a voice data switch which enables alternate use of the extension line for either data or voice communications. The 3221 contains a keyboard of 16 punch buttons, a reader, a speaker, and an amplifier for audio response.□

Gaining a Better Perspective of Communications System Applications

Codes for Voice Functions

Code Name	Code Type[a]	Function	Used for
Call to 3750 Operator	P	Connects extension to any free 3750 telephone operator	Calls to operator
DOD	P	Connects extension to public switched network	Direct outward dialing (DOD)
Personal DOD	P	Connects extension to public switched network, and instructs the 3750 to record call duration for charging as personal call	
Selective DOD	P	Connects extension to a specified exchange on the public switched network	
Public exchange operator recall	S	Connects extension to public switched network operator	External calls
Tie-line access	P	Connects extension to a remote 3750 or other switching system via a tie line	Tie-line calls
Inquiry to extension	S	Holds a call while inquiry is made to another extension and, if required, transfers the call to the other extension	Hold, shuttle, transfer
Inquiry to 3750 operator	S	Holds the call while inquiry is made to a 3750 telephone operator and, if required, transfers the call to the operator	Hold and transfer to operator
Return	S	Enables an extension user to return to held party after making an inquiry to an extension or to the telephone operator. Also enables an extension to shuttle between two parties after making an inquiry or accepting a camp-on request	Hold and shuttle calls
Accept	S	Enables extension to accept a camp-on request	Camp-on
	S	Enables extension to accept a stacked call	Stacking
	S	Enables extension to accept an external call transferred from another extension	Transfer
Camp-on request	S	Signals a busy extension that another extension is waiting on the line, and connects both extensions when the busy extension either dials the accept code or hangs up	Camp-on, shuttle call
Protection	S	Prevents intrusion into a confidential conversation. Same code as used for data protection	Protection
Intrusion	S	Enables extension to break into established call	Intrusion
Short-code dialing	P	Enables associated extensions to call each other	Manager/secretary call assistance
	S	Initiates inquiry and transfer with one digit	
Bar extension	P	Causes all incoming calls to be diverted to an associated extension (for example, the secretary's)	
Free extension	P	Terminates diversion mode established by the bar extension code	
Reclaim	P	Reclaims a rerouted call	Call assistance

Figure 5. IBM 3750: dialing codes. (Reprinted courtesy IBM World Trade Corp.)

Gaining a Better Perspective of Communications System Applications

Code Name	Code Type[a]	Function	Used for
Store external number	S	Stores in the 3750 an external number that is busy, answers, or does not answer, allowing later recall by the repeat external number code	External number repetition
Repeat external number	P	Causes the 3750 to pulse out an external number that has been stored by the store external number code	
Add-on third party	S	Introduces a third party into a call established between two other parties	Add-on third party
Paging request	S	Enables extension to set paging device into action	Dialed paging
Paging answer	P	Connects paged party to caller	
Abbreviated dialing	P	Enables an extension to call selected external numbers by dialing only three digits	External abbreviated dialing
Night service answer	P	Enables extensions to accept incoming calls when the 3750 is in Special-Night-Service Type-2 status	Special night service type 2
Codes for Data Functions			
Protection	S	Prevents intrusion while data is being transmitted from an extension equipped with talk/data switch	Data transmission
Data collection	P	Enables data to be sent from an MF terminal through an extension line for storage on the disk file. The code can have up to 30 different values, each corresponding to a different message format (four values only in basic data collection)	Data collection
Autoconnection request	P	Connects local serial terminal to the DP system via data line	Auto-connection
Autoconnection end	P	Terminates link between local serial data terminal and the DP system	
Teleconnection request	P	Establishes transmission link between two internal data terminals associated with extensions	Tele-connection
Teleconnection end	P	Disconnects data transmission link	
Codes for Contact Monitoring Functions			
Contact sensing	P	Checks whether a contact is open or closed	Contact sensing
Contact operating	S	Changes the status (open or closed) of a contact	Contact operating

[a]P: prefix; S: suffix.

Figure 5. IBM 3750: dialing codes. (Reprinted courtesy IBM World Trade Corp.) (Continued)

Planning Ahead for Your Data Communications Applications

Problem:

A very real problem in planning for and designing a data communications system is that you may yield to the immediacy of the need to get more data from point to point without thinking ahead to the kinds of new applications that can evolve from the increased data flow. Part of the solution to the problem will come from a deep and very real understanding of your company's business, structure, and growth directions combined with an equally deep understanding of communications and data processing technologies. If you happen to be this kind of guru, then you can look forward to a top slot in the highest echelons of your company's management . . . but more about this later.

If you have been following our arguments very carefully, you could ask the embarrassing question, "If all of a network's intelligence is concentrated in its nodes, or termini, and the data links are simply high-speed, dumb, point-to-point conduits, then why worry about the communications facilities for applications planning?" Unfortunately, the present data link offerings are not yet the transparent low-cost ideal paths they will become. There are still many different levels of dumbness, horrendously difficult pricing schedules to figure out, and a lot of tradeoffs to consider between the quality of the service versus its accessibility. These simply stated considerations occupy the bulk of this service, but you must begin your evaluations with a clear picture of your future information processing applications before you commit yourself to the tangibles of hardware, software, and transmission facilities.

The present and projected applications of computer-communication networks or information networks include electronic mail, teleconferencing, "the office of the future," management information systems, modeling, "computerized commerce," monitoring of patients, military command and control, home security, education, and news. This report briefly examines 30 applications and the network capabilities they require. It presents a way of estimating the relative importance of various network characteristics and of predicting the suitability of a network or network architecture for a given set of applications. The report then considers several issues that relate to the political, social, and economic impacts of networks. Among the issues are privacy, security, compatibility, impact on productivity, the roles of networks in international technology transfer and economic competition, and the confluence or collision of the fields of computers and telecommunications.

Planning Ahead for Your Data Communications Applications

Solution:

The subject of this report is applications of networks. The networks involve the use of computers, but computation in the narrow sense does not necessarily dominate the applications. The scope of the report includes, no less than computation, computer-based applications in which the main emphasis is on communication among people, on access to information, or on control of systems, organization, or—to mention early one of the deepest though least imminent concerns—societies. The applications of networks that we shall examine include electronic message communication[1-4], electronic funds transfer[5], access to information, computer-based office work and "telework," management of organizations and command and control of operations, education, entertainment and recreation, reservations and ticketing, and several others.

Some of the problems and issues in network applications are mainly technical and some are mainly nontechnical, but almost all are mixtures of the two, and in most of them the technical and nontechnical factors interact strongly. For example, the relative merits of circuit switching and packet switching are mainly a technical matter, but the fact that the electronic switching stations of the existing telecommunications "plant" are circuit switches surely is an economic factor in the circuit-switching/packet-switching issue. The determination of what should be the individual citizen's right to informational privacy is mainly a nontechnical matter, but the pragmatics of providing informational security, the technical basis for assurance of privacy, must enter the decision process. The national telecommunications plant-in-being of the U.S. figures strongly in many of these problems and issues and forces them to involve both technical and nontechnical factors. The plant is valued at something like $120 billion, and most of it was designed to carry analog voice signals, which are quite different in their spectral and temporal parameters and in their requirements for error handling and security, from digital computer signals of the kinds that will flow through the networks of the future. Because of its inherent redundancy, speech remains intelligible even when mixed with considerable amounts of noise, but even a single undetected error—a single bit—can have extremely serious consequences in electronic funds transfer (EFT) or seriously degrade the performance of a network carrying enciphered information.

One of the major motivations for networking is the need to share resources. The main resources that are often advantageous to share are communications facilities, computer facilities, and information itself. The design of a network can make it easier or more difficult to share resources and thus directly influence the amount of resource sharing that will occur. The amount of communication facilities sharing depends upon many design factors, all of which influence how well the network is able to allocate resources dynamically in response to changing needs and availabilities. Though the need for sharing certain types of computational facilities may diminish with the arrival of the age of the personal computer, it is not at all likely that the need to share resources will disappear altogether. Geographically distributed users can share, through a computer network, the costly high-performance computers that are required to solve certain large computational problems. Even those using personal computers to satisfy the bulk of their computing needs may wish to avail themselves through a network of special software services provided by vendors—and they will certainly wish to communicate with one another.

The sharing of information is the most important type of resource sharing. The term "information sharing" immediately conjures up the thought of sharing large data banks of information among many users, but that is only one aspect of information sharing. All the applications discussed in this paper have aspects of information sharing. Applications concerned with communications, management, commerce, government, protection, education, and awareness all involve sharing. The convenience and effectiveness with which sharing can be accomplished and the facility with which information can flow across the boundaries of individual application programs will have profound effects on how well the applications serve their intended purposes.

Many problems and issues arise from the interaction of information sharing with information security. For example, should EFT have a network or networks of its own to simplify the problem of providing secure transmission, processing, and storage of funds data, or should EFT messages be carried over a general-purpose network so that a reservations and ticketing operation can be completed in a single transaction involving the traveler's organization, the airline and the bank.

APPLICATIONS

In the context of information networks, just as in the context of computer systems, an application is essentially the implementation of a purpose. Applications,

"Applications of Information Networks" by J.C.R. Licklider and Albert Vezza. Reprinted by permission from the November 1978 Proceedings of the IEEE., Vol. 66, No. 11.

Planning Ahead for Your Data Communications Applications

like purposes, may be defined narrowly or broadly. When the airlines began to develop computer-based reservations systems and formed a consortium, ARINC, to interconnect several of their systems, the application was narrow: airline reservations. Now, after years of growth and augmentation, one can rent a car, reserve a hotel room, and arrange to be greeted with flowers and mariachi music. The broadened application might be defined as reservations and ticketing for almost anything that flies to or can be purchased at a distant place. One can project the broader definition into the future and envision a general reservations and ticketing application, operating in a nationwide or worldwide common-carrier network, through which anyone could examine the availability of, and reserve or buy a ticket to, almost anything in the broad class to which reservations and tickets apply. But, of course, there is no reason to stop at that particular class. One can expand the scope further and arrive at "computerized commerce," dealing with the whole gamut of things that can be bought and sold. The very broadly defined application would include advertising, dynamic pricing, and computer-based purchasing strategies. It might even make a place for cartels of suppliers and cooperatives of consumers. No doubt there would be vigorous competition among several or many offerers of the application, and perhaps one can imagine even a "meta-market," an over-arching system that interconnects and integrates the competing "computerized commerce applications."

In any event, that introduces the notion of applications of information networks. It might serve to introduce, also, the notion of issues, which involve the interplay of the opportunity and the threat aspects of applications. It is not difficult to imagine the mischief that could be played by pranksters or dissidents in a poorly protected, publicly accessible, nationwide reservations system.

Basic Applications

Computer-communication networks perform three basic classes of operations upon information: transmission, processing, and storage. The earliest recognized applications of networks were essentially separate exploitations of the three basic classes of operations. Transmission of information through a network from a program in one computer to a program in another, of course, requires some processing and some storage (memory), but in simple message communication and file transfer interest is focused sharply on transmission. In every practical computer, processing requires storage (memory or registers), but in early time-sharing services such as Quiktran[6-8], which when introduced did not provide intersession file storage, the (dial telephone) network application was essentially access to processing. In the Datacomputer[9] service available through the ARPANET[10-12], although processing is involved in both storage and retrieval, one of the main applications is essentially access to storage in and of itself: the Datacomputer is a place to park bits.

A fourth essential network function combines the basic transmission, processing, and storage operations to provide access to information—with the focus of interest on the information itself, rather than on any of the three basic elementary operations.

Simple message communication and file transfer, access to time-shared processing, access to storage, and access to information are important as well as fundamental network functions, but they are no longer typical of the activities or services we associate with the term "application." In present-day parlance, "application" suggests something more highly differentiated and specialized and closer to some specific task or mission.

Communication Applications

In the developmental history of the ARPANET, electronic message service was a sleeper. Even before the network included a dozen computers, several message programs were written as natural extensions of the "mail" systems that had arisen in individual time-sharing systems in the early 1960's. By the Fall of 1973, the great effectiveness and convenience of such fast, informal message services as SNDMSG[13] had been discovered by almost everyone who had worked on the development of the ARPANET—and especially by the then Director of ARPA, S.J. Lukasik, who soon had most of his office directors and program managers communicating with him and with their colleagues and their contractors via the network. Thereafter, both the number of (intercommunicating) electronic mail systems and the number of users of them on the ARPANET increased rapidly.

ELECTRONIC MAIL, ELECTRONIC MESSAGE SYSTEMS: It soon became obvious that the ARPANET was becoming a human-communication medium with very important advantages over normal U.S. mail and over telephone calls. One of the advantages of the message systems over letter mail was that, in an ARPANET message, one could write tersely and type imperfectly, even to an older person in a superior position and even to a person one did not know very well, and the recipient took no offense. The formality and perfection that most people expect in a typed letter did not become associated with network messages, probably because the network was so much faster, so much more like the telephone. Indeed, tolerance for informality and imperfect typing was even more evident when two users of the ARPANET linked their consoles together and typed back and forth to each other in an alphanumeric conversation. Among the advantages of the network message services over the telephone were the fact that one could proceed immediately to the point without having to engage in small talk first, that the message services produced a preservable record, and that the sender and receiver did not have to be available at the same time. A typical electronic mail system now provides a rudimentary editor to facilitate preparation of messages, a multiple-address feature to make it easy to send the same message to several people, a file-inclusion scheme to incorporate already prepared text files into a message, an alerting mechanism to tell the user that he has new mail in his mailbox, facilities for reading received messages, and a "help" subsystem. The prospects of electronic mail appear to have caught the attention of computer manufacturers and software and time-sharing firms as well as telephone companies, national telecommunication authorities, and the U.S. Post Office—and most of them now seem to be planning, developing, or even offering some kind of electronic mail service.

Even before electronic mail was well established, it had become apparent that users would need computer aids for scanning, indexing, filing, retrieving, summarizing, and responding to messages. Indeed messages are usually not isolated documents but documents prepared and transmitted in the course of performing complex activities often called "tasks." Within task contexts, messages are related to other messages and to documents of other kinds, such as forms and reports. It seems likely that we shall see a progressive escalation of the functionality and comprehensiveness of computer systems that deal with messages. If "electronic mail" refers to an early stage in the progression, "electronic message system" is appropriate for a later stage and "computer-based office system" or some comparable term for the stage of full integration. At some intermediate point, message service will no doubt be blended with direct user-to-user linking to provide for delay-free conversation to sequential exchange of messages.

DUOLOGUE AND TELECONFERENCING: Although there has not been thus far, very much use of networks for one-on-one interaction between users, it seems likely that some kind of computer-augmented two-person telephone communication will one day be one of the main modes of networking. In order to displace the conventional telephone, "teleduologue" will probably have to offer speech, writing, drawing, typing, and possibly some approximation to television, all integrated into a synergic pattern with several kinds of computer support and facilitation. The two communicators (and their supporting programs) will then be able to control displays in certain areas of each other's display screens and processes in certain sectors of each other's computers. Throughout a duologue, each communicator will be advised by his own programs and will use information from his own data bases and other sources accessible to him. The effect will be to provide each communicator with a wide choice of media for each component of his communication and with a very fast and competent supporting staff.

A teleconference[14-17] is an organized interaction, through a communication system or network, of geographically separated members of a group. The term "teleconference" has been used recently mainly to refer to interactions organized or presided over by or with the aid of programmed computers. In some teleconferences, the members of the group participate concurrently; in others, each member logs in when it is convenient for him to do so, reviews what has happened in his absence, makes his contribution, and logs out, perhaps to return later in the day or later in the week.

During the last five years, a considerable amount of experience has been gained with computer-facilitated teleconferencing, but it is evidently a complex and subtle art, and teleconference programs still have a long way to go before teleconferences approach the naturalness of face-to-face interaction. On the other hand, we note the inefficiency of traveling to meetings and the inefficiency of letting one participant take up the time of $n - 1$ participants when only $m < n - 1$ are interested in what he is saying. As teleconferencing is perfected (especially nonconcurrent teleconferencing), it will become an extremely important technique.

Neopaperwork

"Office automation," "computer-based officework," "the high-technology office," and a few other such phrases refer to the aggregation and integration of several applications of computers and networks in office work. ("Automation" is intended in its weak sense, which includes computer "aiding" and "semi-automation.") Office automation includes everything presently called "word processing" (dictation, document preparation, etc.) plus computer-based filing (information storage and retrieval), communication (electronic mail, electronic message services, duologue, teleconferencing), and modeling (simulation), and it connects the electronic funds transfer, management information systems, and parts—if not all— of computerized commerce.

Office automation is expected to make heavy use of networks, both local and geographically distributed. Much office work is organized in an approximately hierarchical manner, with component desk functions such as transcription, editing, filing, retrieval, scheduling, and telephone answering at low echelons and corporate or divisional functions such as planning, marketing, operations, and public relations at high echelons. Low-echelon functions typically are carried out locally, within a single office or suite of offices, and, when low-echelon functions are supported by minicomputers or microcomputers, local networks will be required in their integration. In geographically distributed organizations, of course, geographically distributed networks will be required as higher level functions are integrated.

TELEWORK: Networks will make it possible for people to do informational work effectively at locations remote from their managers, their co-workers, the people who report to them, and indeed, even from customers and clients with whom they must interact. Such telework will require facilities for duologue, teleconferencing, and all the other aspects of office automation—but little beyond what automation of a nondispersed office will require.

Telework will offer the possibility of saving the hours and the energy spent in commuting. It may burden some families with more togetherness than was contracted for through the marriage vows ("for richer or poorer, but not for lunch"). But its strongest impact on individual lives will surely be felt by persons immobilized by prolonged illness, physical handicap, or children. For many of them, networks will open many doors—including the door to gainful employment.

AUGMENTATION OF THE INTELLECT: The at-a-distance aspect of computer-based work that is emphasized by the term "telework" will be overshadowed, in the opinion of many, by what Engelbart[18-21] has called "augmentation of the intellect." Computers will help people do informational work faster and better by providing fast and accurate tools to supplement such slow and fallible human functions as looking up words in dictionaries, copying references for citation, stepping through checklists, and searching for matching patterns. Augmentation is needed at levels that range from A) helping poor typists who cannot spell to put out neat and accurate reports, to Z) improving the content and style of top-level policy statements. Expectations differ concerning the prospect for significant early contributions from artificial intelligence, but it is clear that relatively unsophisticated augmentation systems can make major contributions. Consider the help provided by descriptor-based and citation-index-based information retrieval systems to a person looking for references pertinent to a particular fact or concept. Or consider the impact that would be made, on writing such as this, by a text editor that automatically displayed the Flesch Count[22] of every paragraph it helped compose.

TASK MANAGEMENT AND COORDINATION: In addition to helping the individual worker, computers will facilitate teamwork. Each office task will have its planned course of actions, involving particular workers at particular projected times. A computer-based task management system will monitor the task as it moves along the course, checking the actions as they are taken, arranging that planned coordinations and approvals are obtained, and revising the plan (or calling for human help) when the schedule slips. In the early days of office automation, the task management process will be mainly a matter of maintaining orderly work queues for the office workers and displaying for them at each moment 1) what needs to be done and 2) the information needed in doing it. What will need to be done will usually be, of course, to solve a problem or to make a decision—most of the preliminary work will have been performed automatically by computers. With the passage of time, as people come to understand the problem-solving and decision-making processes and the supporting information in programmers' terms, computers will chip away at the problem-solving and decision-making substance of office work, but we expect the now-rising wave of office automation to succeed or fail on the measure of its help to human workers and to human teamwork.

Management Applications

Office automation will have its impact upon management, of course, as well as upon the office workers. Management deals almost exclusively with information. (Money is essentially information, of course.) The comptroller's department was computerized early. Electronic funds transfer will be a major application of special-purpose, limited-purpose, or general-purpose network. On-line financial services may burgeon. Inventory, ordering, production, pricing, and planning will all be interrelated with the aid of networks and computer modeling.

MANAGEMENT INFORMATION SYSTEMS: The widespread feeling of disappointment in the management information systems (MIS's)[23-26] of the 1960's and early 1970's had, we believe, a simple basis: the activities that generated the information required to support management decision making had not yet been brought on line to computers, and, therefore, the required information was not available to the management information systems. To some extent, information important to the manager is so global in scope that capturing it all on line is still not possible. (It was not worthwhile to keypunch all the basic operating data just to feed them into the management information systems, for only a small fraction of the totality would ever be used. It was impossible to anticipate just what subsets or aggregations of the basic operating data would be required.) As soon as all the information activities involved in operating an organization are on line, however, the basis for an effective management information system will exist. A few organizations are already approach-

Planning Ahead for Your Data Communications Applications

ing that state, but most are just entering—or just beginning to contemplate—office automation.

Local and geographically distributed networks will make it possible, at a cost, for top management to access all the facts and figures involved in the minute-to-minute operations of a business. Top management should resist the temptation to convert that possibility into actuality. The principle that looks best at present is to let the data of a corporation reside where the managers most conversant with them reside ("keep the data near the truth points") and to have conversant managers "sign off" on the release of data upwards in the corporate tree. Certain data should be abstracted and moved upward according to preset schedules; other data may be queried from above—but queried through an authenticating release process. Of course, the release process may in some instances be mediated by programs operating on behalf of the human conversant manager rather than by the human conversant manager himself or herself.

The foregoing discussion pertains, indeed, to most of the data management functions in office automation. In distributed organizations, data will be distributed, and one of the main uses of networks will be to move data from points of residence to points of use.

MODELING AND SIMULATION: Computer-based modeling and simulation are applicable to essentially all problem solving and decision making. At present, however, modeling and simulation are computer applications much more than they are network applications—and they are far from ubiquitous even as computer applications.

The trouble at present is that most kinds of modeling and simulation are much more difficult, expensive, and time consuming than intuitive judgment and are cost-effective only under special conditions that can justify and pay for facilities and expertise. But those are prime conditions for resource sharing and, hence, networking. Whereas very large organizations will be able to afford their own concentrations of facilities and expertise, small organizations will not. As management grows tighter and more sophisticated, therefore, there may come to be a place for management consultation and service firms that specialize in modeling and simulation and offer very large or special facilities—and deliver their products through networks. Perhaps a glimpse of such a future has been given by the large array-processing computer, Illiac IV[27,28], which has been used through the ARPANET in modeling the world climate and the space shuttle. Similarly, MIND[29], a system accessible through a value added packet network, is being used to design communication networks.

Commerce

Shifting our attention from activities within an organization, such as a business firm, to interactions among organizations, we can see another kind of application for networks.

ELECTRONIC MARKETS: Networks will serve as marketplaces, providing meeting grounds for buyers and sellers. At first, networks will displace telephone and mail, which now serve the marketplace function for most businesses. Later, networks will begin to displace stock exchanges and commodity markets. Ordinary office automation and funds transfer facilities will adequately support negotiations and transactions when the "commodities" bought and sold are purely informational or sufficiently specifiable by words and figures. Wide-band facilities for examining products at the time of purchase ("squeezing the grapefruit") may extend the scope of the electronic marketplace to commodities that must be selected or approved individually by prospective purchasers. We can expect networks to go beyond the role of the mere place or medium for transactions and, with the aid of sophisticated programs, actively to "make a market" in the sense that certain stock brokers make markets in certain stocks.

COMPUTERIZED COMMERCE: Computerized commerce[30] is based on the idea of electronic markets. It goes beyond providing a marketplace and making a market—and back into the primary motivation of the business firm—by using computers to develop and carry out the strategies and tactics of buying and selling. The concept of computerized commerce is applicable to both the wholesale and retail levels. At the wholesale level, a company's buying algorithms, using data from its inventory data base and from its model of its market and its use of supplies and equipment to produce goods for sale in that market, will try to optimize procurement with respect to operational effectiveness and cost—while its selling algorithms, using parts of the same model, will try to optimize sales with respect to profit. Obviously, a high degree of integration of the whole process from input through output will be advantageous. At the retail level, consumers will need the aid of computers to contend effectively against the computer-based sales procedures for merchants. Unless or until personal computers can provide that aid, there may be a place within networks for shoppers', advisory services and shoppers' cooperatives.

It is evident that a buyer, at either the wholesale or retail level, will often wish to screen the offerings and prices of several or many alternative sellers. To make extensive "comparison shopping" economic, either there must be a market data base that is integrated over sellers or the cost of establishing network connections with many separate sellers must be low.

EMPLOYMENT SERVICES: As suggested in the section on telework, networking may change considerably the conditions and dimensions of employment in the informational occupations. Because networking will for some eliminate the time now spent in travel between home and office, it will make it feasible in certain cases for a worker to work on several different tasks for several different employers or clients in a single day. The new flexibility will create a need for employment services that function essentially as electronic markets in human information processing. Such services would use indexing, abstracting, and retrieval techniques as though workers were books or journal articles. They might create a new mobility in employment, favoring free lancers over workers of constant fealty, and the effect might spill over into conventional employment.

Professional Services

The major professions deal heavily if not exclusively with information. The law is information, and the indexing and retrieval of legal information represents an already very important application of computers and networks. Engineering abounds in potential network applications. We shall deal with teaching shortly under the heading Education. Here, let us examine briefly a few medical applications.

Telecommunications have been used to let a physician at the hospital direct the work of paramedics at the scene of an accident and link a specialist consultant to an emergency operating room a thousand miles away. Networks may serve such applications a bit more effectively or economically than the presently available communication channels do, but the most significant impacts of networks will probably be made in other areas such as the monitoring of patients, the monitoring of (the health of) nonpatients, and access to medical data bases and knowledge bases.

MONITORING OF PATIENTS: Microcomputers and networks will make it possible to monitor continually all the indicators that contribute significantly to the determination of a patient's condition—and to do so for the patient in the home almost as well as for the patient in the hospital. It may, therefore, be safe to send certain convalescent patients home from the hospital earlier than is now the practice and to care at home for certain chronic patients who now have to live in hospitals[31]. For the patients who must remain in hospitals, one can envisage intramural networks far more communicative than the typical present-day call button and signal light—but perhaps a patient with a wide-band channel to someone else's attention would demand more attention.

MONITORING OF NONPATIENTS: It is interesting to speculate upon the application of networks to the monitoring of the health of the entire population. One can imagine future situations in which epidemiological control might be much more crucial than it is now to the maintenance of public health. In order to determine the cause of a rapid increase in the incidence of some deadly new disease, birth defect, or form of cancer, for example, it might be necessary to monitor in detail everyone's dietary intake, exposure to radiation, or even contact with other people. The practical and philosophical difficulties inherent in such an undertaking are too profound to address here, but it is easy to imagine the network, together with the millions of input stations and the massive (distributed) data base, as sine qua non. Less easy to imagine, perhaps, but quite as vital, would be the medical detective work—based on the day-and-night scouring of the data base by analytical and hypothesis-testing programs—required to figure out the cause and to project the cure. But such speculation does not define an application of networks that is probable during the remainder of this century. More attuned to our chosen time scale is the prospect that batteries of routine physical and medical tests might be administered by computer via a network at frequent intervals to many people. Computer feedback from the resulting collections of data might be more effective than the chart beside the bathroom scale in motivating people to exercise, watch their calories, and take their vitamins.

MEDICAL RECORDS: Networking promises to make medical records available wherever needed, even if the patient has an emergency far away from his regular physician and local hospital. The medical records application imposes strong requirements for informational privacy and security and for data management on a nationwide or even international basis, but the factor that will probably inhibit the development of a medical records network is the poor quality of most medical records. As long as a patient's record is kept in longhand (or shorthand) by his physician, there is not much pressure for comprehensiveness; many patients go from one doctor to another, and many records are therefore fragmentary. But the computer is bringing, or can bring, a new approach to the creation of medical records[32,33] (data direct to record from clinical tests, prompting by computers on the basis of check lists, entry of patient histories by patients, and contribution to medical records by paramedics), and if it does so, it will make sense for networks to bring a new approach to storage and retrieval of medical records.

MEDICAL KNOWLEDGE BASES: If the technology of knowledge develops in the way that can be projected from recent work in artificial intelligence, medical knowledge bases will probably become the foci of very significant network applications.

Planning Ahead for Your Data Communications Applications

The application that will doubtless be considered most significant will be to make available to every physician, no matter how remote from the centers of medicine, the vast and carefully organized bodies of medical knowledge that will constitute the medical knowledge bases. Knowledge bases will be used as sources in medical education (especially including continuing education) and as consultants or assistants in medical practice[34,35]. Human expert consultation will, of course, be available to supplement the computer knowledge bases, but considerations of cost and availability will almost surely favor the computer. Difficult problems of legal responsibility and liability may have to be solved: advice from a knowledge base may be similar to advice from a book, but a knowledge-based program that controls the administration of an anesthetic would appear to introduce a new factor.

Possibly even more far reaching in its implications than access to medical knowledge bases by physicians is access to medical knowledge bases by laymen. Knowledge bases for laymen would have to be quite different in content and packaging from knowledge bases for physicians, and ideally the two applications would complement each other. The layman-oriented application might deal mainly with the complaints not ordinarily taken to a doctor or with the decision process that determines whether or not to seek a physician's help. In either case, if the knowledge-base program had access to the individual's medical record, and if it could make simple observations such as temperature and pulse rate through the network, it could go rather far beyond the limits of the conventional book of medicine for the layman. Society should examine such incursions by the computer into medicine or even paramedicine very carefully before making up its collective mind about them. They obviously mix benefits with dangers. Unfortunately, they tend to be approached with prejudice.

Government Applications

Actual and potential government applications of networks include military command and control, communications, logistics, acquisition and interpretation of intelligence data, dissemination of intelligence, law enforcement, delivery of government services such as Social Security benefits to citizens, and converting the paperwork of the bureaucracy into bits. Paperwork in the government is rather like paperwork in the private sector, but carried a step or two further into detail. The other government applications, on the other hand, seem rather special. Military command and control, communications, intelligence, and to a considerable extent logistics systems must be able to operate fast, move fast or hide, and function in the presence of physical (as well as other) countermeasures. Law enforcement information systems are in some ways like highly amplified credit reference systems: derrogatory information seems especially crucial, for to be forewarned is to be forearmed, and action must often be taken on the basis of whatever data can be assembled in a few seconds. Serving all the citizens and collecting taxes from most of them requires that certain personal data be held about almost everyone—enough in sum to make a several-trillion-character data base that at least conceivably could be subverted to political or economic exploitation. There are strong lessons about government applications of networks in the recent rejection by the Office of Management and the Budget (OMB), at the well-timed suggestion of several members of Congress, of the proposed new Tax Administration System of the Internal Revenue Service.

MILITARY COMMAND AND CONTROL AND MILITARY COMMUNICATIONS: Military command and control and military communications are prime network applications. Both interactive computing and networking had their origins in the SAGE system (Semi-Automatic Ground Environment for air defense), and many of the military systems used to command forces and control weapons are essentially computer-communications networks. For reasons we do not fully understand—since fast response to a changing situation is the essence of command and control—the World-Wide Military Command and Control System (WWMCCS) is actually not very interactive, and its computers, which use the GECOS operating system[36] designed for batch processing, are not interconnected by an electronic network. But surely WWMCCS will in due course be upgraded. Autodin II is under development and will supplement or replace Autodin I[37], the Department of Defense's present store-and-forward digital telecommunications network, with a modern packet-switching network based on modified and secured ARPANET technology. Networking is being pursued actively, also, in the intelligence community. One of the most significant possibilities for the military that is opened up by advances in information technology is the achievement of a much tighter coupling between intelligence and command and control. One can envision a reduction in the time required for the distribution of intelligence information from days or hours to minutes or seconds. Such an advance would, of course, put pressure on intelligence gathering and processing to operate on faster time scales.

MILITARY LOGISTICS: There is less progress, but also less pressure, in the logistics area, where more than 20 large batch inventory systems can be counted, diverse in respect of hardware, programming language, and data management system. Over the coming years, however, even the logistics situation will probably be brought under control and onto a network. The overall objective is to make the entire

Planning Ahead for Your Data Communications Applications

operation of a military effort responsive to coherent hierarchical command in the light of valid and current intelligence—with security against enemy actions and countermeasures.

THE NETWORK OF THE NATIONAL CRIME INFORMATION CENTER (NCIC): The NCIC is operated by the FBI and connects with state and local police units in most of the states. The NCIC contains, among other things, data on stolen cars and stolen license plates and the police histories of convicted criminals. The case of the NCIC network is an interesting study because, in it, the informational needs of the police and the information-providing capabilities of computers and telecommunications run head-on into Congressional concern for the right of informational privacy. When a police officer stops a speeding car and appoaches it to make an arrest, he would like to know something about the car and driver. Is the car stolen? Does the owner have a history of resisting arrest? Forewarned is forearmed. About two years ago, however, an innocent man was killed by an arresting officer forearmed by forewarning with incorrect information. In the most recent chapter, the Senate Committee on Government Operations caused to be rejected the NCIC's request for permission to acquire a message-switched network to speed up communication with state and local police.

SOCIAL SECURITY: In the U.S., the Social Security Administration (SSA) distributes more than $100 billion a year to more than 20 million people and interacts with millions of clients each year through about 1300 offices manned full-time and 3000 manned part-time. The set of computer processible data bases that support SSA operations contain more than a trillion characters, and it is estimated that in those operations each year several trillions of characters flow from one location to another, about a tenth of a trillion by a network and processing system called the "SSADARS system," and the rest mainly by mail.

In 1976, the SSA began planning the modernization of the process through which it discharges its massive responsibilities[38]. The new process will make even heavier use of computers than does the present one and there will be much consultation and updating of central or regional data bases from local offices. (The present process requires several computer areas, each with multiple mainframes, more than a hundred disk drives, and more than a hundred tape drives—and, in all, approximately 400,000 magnetic tapes). The new communication subsystem will, therefore, be a network of very major proportions, probably a dedicated SSA network operated by the General Services Administration or (improbably in the near term) a part of an even larger and more comprehensive network. Most technologically developed countries will sooner or later have social security networks.

Protection

If military, intelligence, and police networks are reckoned as networks for protection, then protection is a very large category of network applications. There is another member of the class that deserves mention.

HOME AND NEIGHBORHOOD SECURITY: Several of the projected applications of computers in the home relate to security: sentry against intrusion, fire, and gas and water leakage, monitoring the well-being of the elderly and infirm, and "electronic babysitting." In most of these applications, computers will be better at detecting trouble than in correcting it, and there will be a strong requirement for communication with remote persons or agencies. At present, some burglar alarms are connected by dedicated lines to central security offices or police stations and some "dial up" in the event of trouble. If a packet network were available, it would probably be less expensive and it would provide a wider range of options, including absent members of the family, friends, and neighbors as well as security companies and public agencies.

It seems possible that a neighborhood communication medium (with a broader fan-out or faster sequencing of calls than the telephone) might be just what is required for the elderly to help one another achieve a higher level of security and peace of mind. CB radio or house-to-house (or apartment-to-apartment) wiring or a multipurpose packet network could provide the medium.

Probably just conversation of the kind that prevails on CB radio interconnections would go a long way, but it could be reinforced by slightly higher technology. Home computers could be programmed to interpret a variety of indicators of trouble—the sound of a fall, too long a flow of water, the refrigerator door open, prolonged quiescence—and to ask for an "all's well" report whenever there was cause for concern. Failing to be satisfied that all was indeed well, the computer could call for help. It would have a list of participating neighbors and a schedule of probable availability for each, and—by communicating with their home computers—it could quickly find someone to look in and check, or provide assistance. It has been suggested that a neighborhood net could monitor its clients while they were walking on the sidewalks as well as while they were at home, and a small device has been demonstrated that sends out a radio signal when its wearer falls down—or for some other reason becomes horizontal[31]. If neighborhood networks existed, there would probably be no end of inventions to exploit them in the interest of security: heartbeat monitors, breathing monitors, footstep sensors, and so on. And if the present trend of population statistics continues, security applications might constitute a significant sector of the network application pie.

Education and Awareness

Beginning with the last section and continuing now into this one, the focus of interest has moved from the organization—or the individual as a member of an organization—to the individual as an individual in the primary family group in the home. Probably the most important network prospects for the individual in this century lie in education and training.

COMPUTER-BASED EDUCATION AND TRAINING: We assume that advances in computer representation of knowledge and in computer mediation of interactions between people and knowledge bases will advance computer-based education and training far beyond the "expensive page turners" and drill and practice routines that are associated in many minds with the term "computer assisted instruction." We assume that knowledge bases accessible through networks will eventually accumulate more knowledge, in each of many fields of learning, than typical teachers are able to master and retain, and that the knowledge in the knowledge bases will be well organized (by experts in each field) and effectively accessible to students at all levels of mastery and aptitude. However, computer-based techniques for the representation, organization, and exploration of knowledge are at present still topics of research—and even if they were fully developed today, it would still take a decade or two to translate the content of the many fields of learning into computer-processible knowledge bases. During the coming years, therefore, application of networks in the area of computer-based education and training will be preliminary and propaedeutic. Perhaps toward the end of the century it will approach its ultimate volume and significance—and be among the top three or four uses of networks.

NEWS: At present, most people gain their awareness of what is going on in the world mainly through mass media that report on events rather than processes, that select a few news items instead of covering the news, and that give everyone, regardless of his or her interest pattern, the same few selections. Networking has the potential of changing the news into a multidimensional dynamic model of the world that each individual can explore in his own way, selecting for himself the topics, the time scales, the levels of depth and detail, and the modes of interrogation and presentation. Interest profiles and other techniques of selective dissemination may play important roles, but networking in principle removes the necessity of disseminating (with its implication that the initiative lies mainly with the transmitter) and opens the door to self-directed exploration and investigation by the receiver of the news. To provide the multidimensional dynamic model for exploration and investigation would, of course, be a demanding responsibility for the gatherers and organizers of news, but they gather and organize much more even now than they print or (especially) broadcast. There will probably be a long slow evolution from the newspaper/newsmagazine format and the nightly news format through increasing levels of user initiative toward truly user-dominated interaction with a whole-world knowledge base.

REQUIREMENTS IMPOSED UPON NETWORKS BY APPLICATIONS

Now that we have sketched out several applications, we should examine briefly the network characteristics they require. The applications do not all require the same network characteristics, of course. One application may require one pattern of characteristics, while another application may require another pattern. Some of the frequently required characteristics are the following.

1. Bidirectional Transmission: Most applications require two-way communication—if not the capability of sending and receiving simultaneously (full duplex) then at least the capability of alternating between sending and receiving (half duplex).

2. Freedom from Error: One wrong bit may completely change the meaning, especially if numerical data are represented nonredundantly. In such cases, even though the basic communication channel itself is not error free, the end-to-end communications must be made error free through the use of adequate error handling mechanisms.

3. Efficiency Despite Burstiness: A source that transmits short bursts of information and is quiescent between bursts typically does not wish to pay for channel time while it is quiescent. Both human beings and computers are bursty sources.

4. Low Cost per Bit: The cost of network service depends, of course, upon many nontechnical factors as well as upon the technical efficiency of the network in converting its resources into services. But technical efficiency is a very strong and basic factor. This characteristic refers to the cost of transmitting one bit from source to destination. The relation of cost to distance is considered separately.

5. High Connectivity: A source may need to transmit to any one or more of many destinations. A destination (i.e., user) may need to examine many sources.

6. High Information Rate: Wide-band channels are capable of transmitting many bits per second. The criteria for "high," "wide," and "many" vary widely

with type of signal and level of expectation. The 50,000-bits/s information rate of most of the ARPANET channels seems like a high information rate for ordinary interactive computing, but it is too low for convenient transmission of large files (e.g., high-resolution photographs) and far too low for moving pictures, even low-resolution television pictures.

7. Security: Security is a complex of characteristics, some of which provide the technical basis for the protection of privacy. Others have to do with preventing disruption of service and protecting against fraud and theft.

8. Privacy: In the U.S., this complex of informational rights, including but by no means limited to protection against eavesdropping, has been formulated by the President's Commission on Privacy and, to a considerable extent, expressed in legislation in the Privacy Act of 1977. Other nations also have privacy laws, of course, some of them in some respects more stringent than ours.

9. Authentication: A good authentication scheme provides the electronic equivalent of a signature. Ideally, authentication identifies the author of a document and makes it impossible for him to escape responsibility for the authorship. Ideally, also, authentication makes it impossible for anyone to change even one character or bit of the document without destroying the "signature."

10. High Reliability: Low probability that network service, as seen by the application, will be impaired by macroscopic malfunctions. For present purposes, we distinguish between macroscopic malfunctions and microscopic errors in bit transmission.

11. Full-Duplex Transmission: Some applications require, and most are favored by, the capability of sending to another station and receiving from it at the same time.

12. Priority Service: Guaranteed or preferential service, especially when the network is congested, is widely regarded as essential for certain very important functions or for certain very important persons.

13. Speech Capability: Present speech circuits transmit alphanumeric information inefficiently, and most present data networks were not designed to transmit speech. It will be advantageous, however, to integrate speech with data.

14. Pictures: It will be advantageous to integrate pictures, also, into the repertoire. Graphs, charts, diagrams, and simple sketches fit readily into the pattern of data transmission, but high-resolution pictures and, especially, moving pictures require high information transmission rates. This characteristic is essentially a second "information rate"—but scaled in such a way as to be more demanding of very-wide-band capabilities.

15. Insensitivity to Distance: Synchronous satellites and packet switching both tend to make the difficulty and cost of transmission less dependent on distance than they are in traditional communication systems. Rarely is it an absolute requirement that difficulty and cost be independent of distance, but often it is desirable.

16. Short Transit Time Delay: If the sum of the signal-transit time and the signal-waiting-in-buffer time is too great, an application may be slowed down too much or disrupted. The 0.2-s delay introduced by transmission via a synchronous satellite somewhat disturbs two-way speech communication. The delay introduced by transmission from one processor to another may slow down the operation of a multiprocessor that is a network of minicomputers or microcomputers.

17. Uniform Time Delay: In some applications, successive segments of the signal must reach the destination in sequence (or be put back into sequence if they arrive in scrambled order). Note that reordering may cause all the segments to be delayed as much as the most-delayed segment.

18. Broadcast Capability: Some applications require, and some are favored by, the capability of transmitting to many or all destinations concurrently.

19. Mobility: Some or all of the stations may need to move from place to place and may need to communicate in transit.

To obtain rough measures of the requirements imposed upon networks by the several applications, we filled in the body of (an early version of) Table 1.* Into each cell we entered a number to indicate our intuitive rating of the importance of the characteristic for the application. The rating scale we used runs from 0 (lowest) to 5 (highest). For example, we considered connectivity to rate at 4 in importance for mail and message systems because mail and messages typically fan out widely from senders to receivers and fan in to receivers from a wide distribution of senders. We did

*In order to obtain a broader basis on which to think about, and possibly model, the relations between networks and applications, we suggest that you (the reader) photocopy Table 1 and fill in some rows with your estimates of the importance of the characteristics to the applications. We have also provided room for you to define additional applications or characteristics. If you are willing to share your estimates with us, despite the fact we are not bold enough to share our raw-data estimates with you, please post them to J.C.R. Licklider and A. Vezza, MIT-LCS Rm. 219, 545 Tech Sq., Cambridge, MA 02139.

Planning Ahead for Your Data Communications Applications

NETWORK CHARACTERISTICS / NETWORK APPLICATIONS	1 Bi-Directionality	2 Freedom from Error	3 Efficiency Despite Burstiness	4 Low Cost Per Bit	5 Connectivity	6 Information Rate	7 Security	8 Privacy	9 Authentication	10 Reliability	11 Full-Duplex	12 Priority	13 Speech Capability	14 Picture Capability	15 Insensitivity to Distance	16 Shortness of Delay	17 Uniformity of Delay	18 Broadcast Capability	19 Mobility	SPACE FOR ADDITIONAL CHARACTERISTICS	SUBJECTIVE RATING OF IMPORTANCE OF APPLICATION
BASIC																					
Transmission																					
Storage																					
Processing																					
Information																					
COMMUNICATION																					
Mail, Messages																					
Duologue																					
Teleconferencing																					
Speech																					
Encrypted Speech																					
Still Pictures																					
Moving Pictures																					
NEOPAPERWORK																					
Telework																					
Augmentation																					
Task Management																					
MANAGEMENT																					
M.I.S.																					
Modeling																					
COMMERCE																					
Electronic Markets																					
Computerized Commerce																					
Employment Services																					
PROFESSIONAL																					
Monitoring Patients																					
Monitoring Non-Patients																					
Medical Records																					
Medical Knowledge Bases																					
GOVERNMENT																					
Military C^3																					
Logistics																					
NCIC																					
Social Security																					
PROTECTION																					
Home Security																					
EDUCATION AND AWARENESS																					
Computer Based Education																					
News																					
SPACE FOR ADDITIONAL APPLICATIONS																					

Table 1. Importance coefficients of network characteristics for various classes of network applications; each cell entry should be subjective rating of importance on a scale from 0 (low) to 5 (high)

Planning Ahead for Your Data Communications Applications

not assign a 5 because mail and message systems would still be valuable (cf., the plans of Satellite Business Systems) if connectivity were limited to within organizations. In the case of the column 6, information rate, we used a somewhat special scheme. The numbers from 1 through 5 encode five class intervals of information rate in bits/s: 1) 75-300, 2) 300-1000, 3) 1000-10,000, 4) 10,000-1,000,000, and 5) above 1,000,000. For the rough purposes of our analysis, nevertheless, we shall interpret the entries in column 6, as all the other entries, to be estimates of the importance of the (columnar) characteristic for the (row) application.

Figure 1 shows the relative importance of the 19 network characteristics. As one examines the average ratings of the characteristics, it comes as no surprise that bidirectionality is very important. It is the "co" in "communication."

It may be slightly surprising, however, that freedom from error is so important. It is freedom from error as seen by the application, of course. There are bound to be errors in the raw network channels, but they may be detected and eliminated by error-correcting circuits or by retransmission. "Error-free" may in practice mean one bit error in 10^{12} or 10^{14} bits, on the average. The importance of achieving an extremely low error rate stems in part from the fact that many of the applications involve information, such as financial data, in which changing a single character could make a great difference. Freedom from error is required, also, by most cryptographic schemes. Where freedom from error is not required by an application, one can usually find error detecting and correcting mechanisms within the application itself. Such mechanisms are quite evident, for example, in human conversation. But it greatly simplifies most network applications if the network can be counted on to do the error handling.

Ability to handle bursty transmissions efficiently ranks third. The advantage provided by this characteristic translates directly into a cost advantage.

Low cost ranks fourth in importance. It did not rank higher because we recognized that certain of the applications, such as military command, control, and communication, are relatively insensitive to cost. Also, other network applications such as mail and messages are already quite cost competitive with their conventional counterparts and do not demand very-low-cost facilities.

Connectivity ranks fifth in importance. The reason connectivity does not rank higher is that we assigned only a medium score for connectivity to applications that required only connectivity within an organization or within a region, and many applications could function—though perhaps at some disadvantage—with such limited connectivity.

Information rate ranks sixth. We interpret that to mean that very wide-band transmission is not vital to most of the applications and that most of them could be satisfied with an information rate in the range 1000-10,000 bits/s. However, that is the information rate seen by the application. To handle heavy traffic, and to handle a few of the applications, a network should have channels of considerably greater bandwidth than that.

Security, the complex that includes assurance of service when required and protection against fraud and theft, ranks seventh.

Privacy, the complex that includes protection against disclosure of personal information and unauthorized use of it, ranks eighth.

Authentication, a characteristic closely associated with security, ranks ninth.

At the other end of the ranking, mobility (19th) is not required by most of the classes of applications we considered—but, of course, is essential for some applications.

Broadcast capability (18th) was scored low because it is not needed at all in many applications and is needed only occasionally in others such as mail and message systems—and, when needed, usually can be simulated adequately by repeated point-to-point transmissions.

Uniformity of time delay (17th) is important mainly for speech transmission. If speech had been given a weighting proportional to its probable eventual importance in networking, uniformity of time delay would have ranked higher.

The capability of giving preferential treatment to high priority traffic (12th) ranks as high as it does because we viewed priority in the context of present-day systems that may introduce considerable delays into the delivery of some or all of their messages. In the context of future systems in which a whole transmission will take less than a second, priority may be much less important. However, the need for priority is unlikely to vanish. Priority classes are useful in queuing messages for processing by people and in indicating the prioritizer's sense of urgency or importance to the recipient. Moreover, even very wide-band systems tend to be designed just barely to handle expected peak loads, and even such systems can be overloaded—in which case, prioritization might be helpful. On the other hand, it is conceivable that the processing of priorities might slow a system down more than eliminating low priority traffic could speed it up.

Planning Ahead for Your Data Communications Applications

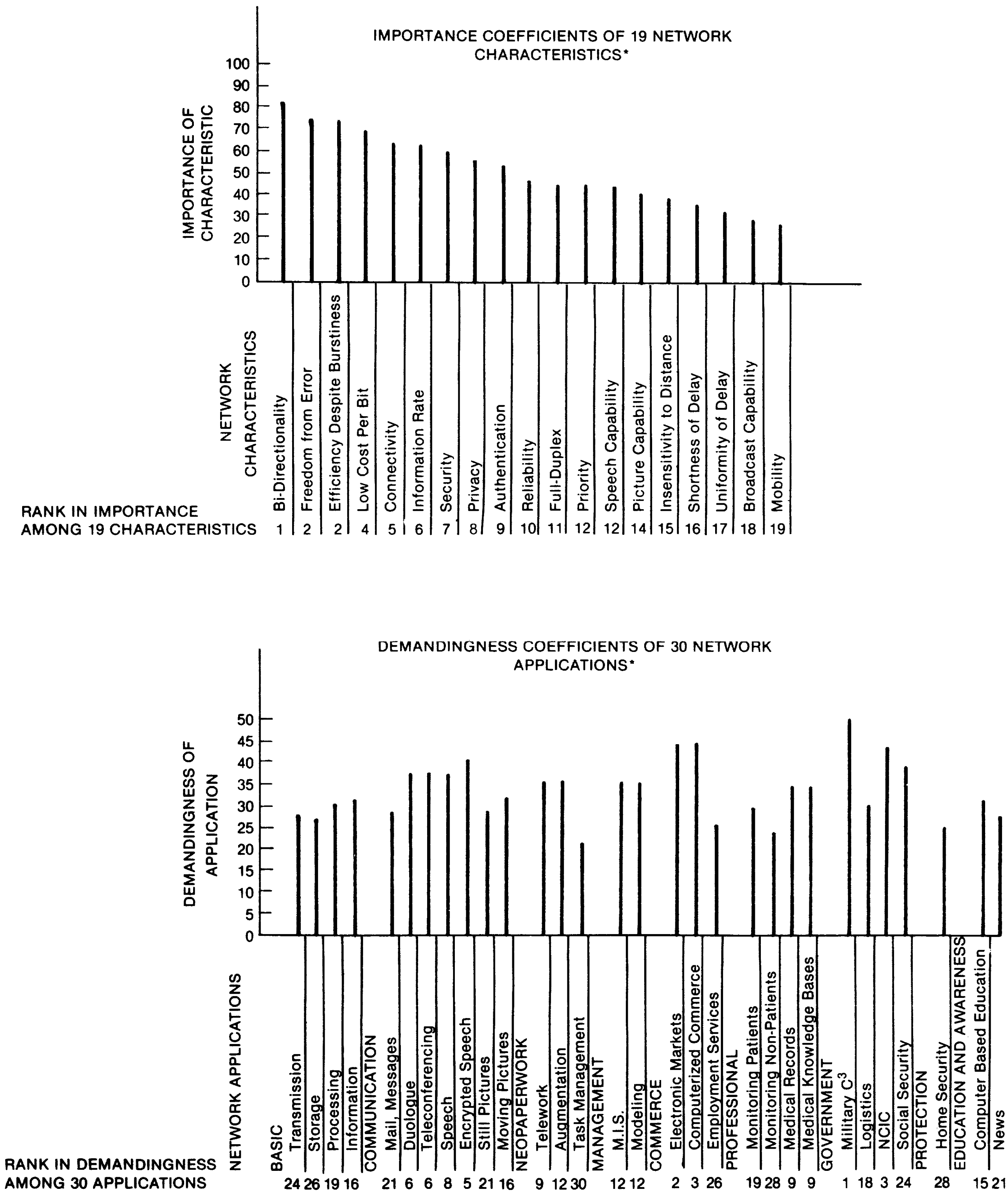


*Figure 1. Importance coefficients of 19 network characteristics and demandingness coefficients of 30 network applications (*Divide each number by 1000 to scale the sum to unity.)*

Planning Ahead for Your Data Communications Applications

We do not want to attribute too much value to our no-doubt-idiosyncratic subjective estimates of the importance of network characteristics to applications, but we would like to carry the analysis another step to illustrate what we think might be a valuable method. It attempts to deal with the relative merits of various networks or network architectures.

The method begins with a table of applications versus characteristics similar to Table 1 except that each application has an importance weight and all its cell values are multiplied by that weight before the columns are totaled. The method assumes, also, a table of networks or network architectures versus network-characteristics such as Table 2. The entries in Table 2 represent our very subjective impressions of the degrees to which the characteristics at the top characterize the networks at the left-hand side. The values are certainly not definitive. In the case of the hypothetical augmentation of the ARPANET, they assume major increases in number of subscribers and in information rate and they assume that advanced provisions are made for security, privacy, authentication, and priority service. They assume, also, that satellite relays are incorporated into the network along with wide-band surface channels and that there is a packet-radio subsystem to serve mobile applications.

In the case of the projected SBS service, indeed, they are based only on the most informal information, and they make rather optimistic estimates about the characteristics of the hypothetical networks that would be developed on the basis of the SBS facilities. The reader is invited to substitute his or her own estimates.

To determine the suitability of a network or network architecture to a set of applications, one simply multiplies each cell value in its row in (the table like) Table 2 by the importance of the corresponding characteristic at the bottom of (the table like) Table 1— and then finds the sum (across the row) of the products.

To illustrate the use of the method, we worked with the four networks of Table 2 and with four sets of applications. Application set 1 was the set shown in Table 1. Sets 2, 3, are 4 were subsets consisting of—set 2: speech and encrypted speech, set 3: still and moving pictures, and set 4: duologue and augmentation. (We gave equal or uniform weighting to the applications in each set—to all 30 in the first set and to both of the two in each of sets 2, 3, and 4.)

We obtained the 16 appropriateness indexes shown in Table 3. The detail telephone network performs best in the speech applications, of course, but it does not appear to do badly in the others. (Giving more weight to cost tends to reduce its scores.) The experimental ARPANET appears to perform well on speech, which surprised us despite the fact that experiments on the transmission of compressed speech over the ARPANET have been very successful, but best on duologue and augmentation, which we expected. In the scoring, the ARPANET suffers because, being experimental, it was not developed in respect to some of the important network characteristics. Assuming such development (Hypothetical Augmented ARPANET) yields the appropriateness scores in the third row of Table 3, which are all toward the upper end of the five-point scale. The hypothetical corporate network based on the projected SBS service appears to perform almost as well as the hypothetically augmented ARPANET, suffering in the comparison only

NETWORKS \ NETWORK CHARACTERISTICS	1 Bi-Directionality	2 Freedom from Error	3 Efficiency Despite Burstiness	4 Low Cost Per Bit	5 Connectivity	6 Information Rate	7 Security	8 Privacy	9 Authentication	10 Reliability	11 Full-Duplex	12 Priority	13 Speech Capability	14 Picture Capability	15 Insensitivity to Distance	16 Shortness of Delay	17 Uniformity of Delay	18 Broadcast Capability	19 Mobility
Dial Telephone	5	2	0	2	5	3	2	2	1	3	2	1	4	2	0	4	4	2	2
ARPANET	5	4	4	3	3	3	1	2	1	4	4	0	2	2	4	3	3	1	1
Hypothetical Augmented ARPANET*	5	4	4	4	5	5	4	4	4	4	4	3	4	4	4	4	4	3	4
Hypothetical Corporate Network Using SBS**	5	4	4	4	2	5	4	4	4	4	4	3	4	5	4	2	4	3	0

*A hypothetical network based on ARPANET technology but with very wide-band ground and satellite channels, very many subscribers, advanced provisions for security, an authentication scheme, arrangements for priority, and a mobile/portable radio adjunct based on the ARPA Radio Net.

**A hypothetical network of the kind that might be based on the projected facilities of the Satellite Business Systems Corporation and used by a large corporation with geographically distributed branches. It is assumed that this network is used only within the corporation and therefore has restricted connectivity.

Table 2. Estimated degrees to which four selected networks possess the 19 network characteristics, the ratings are the authors' intuitive estimates on a scale from 0 (low) to 5 (high)

NETWORKS \ APPLICATION SETS	1 30 Applications of Table I	2 Speech and Encrypted Speech	3 Still and Moving Pictures	4 Duologue and Augmentation
Dial Telephone	2.5	2.8	2.4	2.6
Experimental ARPANET	2.8	2.8	2.5	3.0
Hypothetical Augmented ARPANET	4.1	3.9	4.0	3.8
Hypothetical Corporate Net Based on SBS Service	3.8	3.5	3.9	3.8

Table 3. Appropriateness scores for four selected networks, each rated on four different sets of applications. The appropriateness scores are based on a scale from 0 (low) to 5 (high). The way they were determined is described in the text.

because we assumed for it limited connectivity (intercorporate communication only), no surface channels (and therefore always the 0.2-s satellite delay), and no mobility. Those lacks showed up only in the average over the 30 applications and in the speech applications.

Obviously, the result obtained with the method is no better than the ratings it processes, and we do not make any claims for our ratings. We believe, nonetheless, that the scheme puts into an orderly array some of the basic factors that determine the relative appropriateness of various networks for various sets of applications, and that it leads the users of the method—or at least it led us, as we used it—to consider the factors carefully and to think about how they act and interact. Upon examining the interactions, it soon becomes clear that the linear weighting scheme smooths over many nonlinear logical interactions, and that a more advanced model would have to be more like a computer program than three tables and a pocket calculator. Nevertheless, the first step has to be to survey the variables that are active in networking, and the simple scheme provides a start on that.

NETWORK ISSUES

Brittleness

Brittleness is approximately the inverse of the lauded complex of system attributes: flexibility, robustness, and gracefulness of degradation. Brittleness often arises from a quest for efficiency or economy. If you space the pony express posts as far apart as a fresh pony can run, then the mail does not get through when an emergency forces you to use a tired pony. A socioeconomic unit with a minimum-capacitance supply system avoids the waste inherent in having products stagnate in the pipeline but crumbles in a siege. Networks will almost surely be more efficient than the systems they supplant. Should not the expectation prompt us to ask whether they will also be more brittle?

From an engineering point of view, the preferred approach is to avoid brittleness through judicious choices in the architecture and design of networks. The dynamic routing feature of the ARPANET, for example, permits the network to continue to operate, as far as two functionally connected host computers are concerned, as long as there is some path left intact between them. Indeed, just as it is to reliability, redundancy is the main engineering antidote to brittleness in networks—redundancy of interconnection, redundancy of power supply, and redundancy of information storage. Moreover, sophisticated uses of redundancy, as in restructurable logic and error-detecting codes, provide much greater returns in robustness than do brute force applications that cost the same.

Electronic Imperialism or Technology Transfer to the World

Aviation may have had more impact in technologically not-yet-developed but developing countries with poor roads and few rails than in technologically developed countries that already had working transportation systems before airmail routes and airlines came upon the scene. Brazilia, for example, would not have been feasible without the DC-3. Analogously, networks may have their most dramatic effects where there are few critical masses of knowledge and few self-reinforcing centers of intellectual activity. Networks may link the geographically separated subcritical foci of cognition in the developing world with the concentrated supercritical centers of the developed world, bringing the former deeply into the interaction patterns of the latter and making it much easier for the former to grow and advance. If networks do turn out to have such an effect, it may represent a new dimension of imperialism, or it may open up broad new avenues of technology, transfer, or, as seems most likely, it may look one way to some and the other way to others.

Although technology transfer to the developing world is often viewed as a matter of delivering journals, reports, and books and of transmitting data to third-world countries, truly effective transfer is a transfer not of data or of information but of knowledge and it flows through human interaction. Perhaps the most effective involves graduate study by promising young people from developing countries in leading graduate schools in developed countries—followed by work experience in the developed countries and then return to form foci of advanced technology in the developing countries. But the difficulties with even that pattern are well known: reluctance to return in the "brain drain," or isolation after return resulting in unhappiness and ineffectuality. What is needed is a way to return without breaking the ties of interaction with teachers, fellow students, managers, and coworkers in the centers of technology that, thanks to their highly developed intellectual and motivational supports, made possible the transfer of knowledge in the first place. Networks can fill that need. The ARPANET has provided several instances, albeit just in the U.S., in which students have remained functionally and motivationally in the research groups in which they worked for their degrees until they could build up self-sustaining foci of their own in their postdoctoral locations. It seems very likely that the postdoctoral locations could as well be in foreign countries, even countries with little technology, if only they had network connections and sufficient funding to keep local terminals in reliable operation and to pay for computing time and storage somewhere on the net. Indeed, it seems likely that technology transfer to developing countries could become real and effective through no more than informal extension of patterns of interaction that have become well established in the ARPANET community. But there is no reason to keep the patterns entirely informal. Formal associations between universities, not-for-profit organizations, and business firms in the have-technology and the have-not-technology countries would surely increase the productiveness of the technology-transfer enterprise[39].

We do not want to try to take sides with respect to, and we cannot hope to offer a solution to the problem suggested by the juxtaposition of "electronic imperialism" and "technology transfer." We do suggest, however, that the prospect of networking between the developed and the developing worlds deserves very serious study. From the point of view of the imperialist, it may well be that packet networks will be to the not-far-distant future what clipper ships (or were they packet ships?) and clipper aircraft were to the not-far-distant past. From the point of view of countries that need and want technology transfer, packet links to technologically developed countries may be by far the best way to get it if they can figure out how to keep the electronic colonialism from coming with it.

Unity, Federation, or Fragmentation

If we could look in on the future at, say, the year 2000, would we see a unity, a federation, or a fragmentation? That is: would we see a single multipurpose network encompassing all applications and serving everyone? Or a more or less coherent system of intercommunicating networks? Or an incoherent assortment of isolated noncommunicating networks, most of them dedicated to single functions or serving single organizations? The first alternative—the strongly unified network—seems improbable: of almost zero probability if the scope is taken as worldwide and still of very low probability even if the scope is taken to the U.S., which seems to engender pluralistic solutions to most problems. The third alternative—many separate noninterconnecting networks—is what would be reached by proceeding with a plan and, therefore, may be judged rather probable, but it would be very disappointing to all those who hope that the whole will be much greater than the sum of the parts, i.e., that many of the projected applications will facilitate and contribute to one another to such an extent that the overall value will grow combinatorially. The middle alternative—the more or less coherent network of networks—appears to have a fair probability and also to be desirable, but it brings with it the problem of how to achieve enough coherence to support fast and facile intercommunications among the subnetworks when required, and that may be a difficult problem. Let us consider first why coherence is desirable and then turn to the difficulty of achieving it.

"Coherence" characterizes a system in which all the parts articulate well and function in synergy and in which the subsystems are compatible and cooperative. In an information network, coherence is desirable partly for the same reason it is desirable in a telephone system: the value to a typical user increases as the number of other accessible users increases. This is true no matter whether the users are people or computers. The importance of coherence is amplified, however, by the fact that some of the applications of computer networks will involve several functions operating on common information. Planning a trip, for example, will require interaction among: your calendar program and the calendar programs of people you will visit, the reservations programs of airlines, car rental services, and hotels, the funds transfer systems of your bank and several other banks, your company's travel office, and perhaps data bases pertinent to business to be transacted on the trip. The computers could not be of much help in the planning if each function or service had its own separate network or if their networks were physically interconnectable but incompatible at various levels of protocol. Indeed, coherent interconnection of diverse functions will be essential if networks are to live up to expectations in electronic message services, computerized commerce, delivery of social services that involve both federal and state governments, and many military and intelligence applications.

Coherence, however, is a condition that has to be planned and striven for. It does not arise in a short time through evolution—at least not through evolution that conforms to the spontaneous-variation-plus-natural-selection model. Will the forces that are operating in the present network situation foster a sufficient degree of coherence for networks to fulfill their promise? On the positive side is the fact that a standard packet-switching interface protocol X.25[40] was formulated and agreed upon in an unusually short time and that the interconnection of such dissimilar networks as Telenet[3] and the Canadian data network has already occurred. Also on the positive side is the possibility that one or two commercial networks, such as the one being developed by AT&T and the one being developed by Satellite Business Systems, will dominate the network market and thereby create the kind of coherence that IBM has created in a large part of the computer software field.

However, the other factors seem to work against coherent interconnection. First, the network situation is evolving without any national policy. (Within the U.S. government, it is evolving without any federal policy). Several countries are building networks independently. Many companies are building networks independently. Second, although the lines may be leased from the same telephone company that leases lines to everyone else, there is some security in having

Planning Ahead for Your Data Communications Applications

one's own dedicated network. The need for security in such areas as EFT may be stronger than the need for interconnection. Third, wherever personal information is concerned, and especially in the federal government, privacy has become a major issue, and a simplistic interpretation of the privacy problem sets interconnection into opposition with privacy. Fourth, research and development in the area of network security and privacy assurance have not been and are not being supported at a high enough level to create—soon enough—a technology that will let one say: "You can have both interconnection and security, both interconnection and privacy; you can have your cake and eat it, too." Actually, the technology of communications security is rather well developed[41-44]–except for uncertainties arising from the "56-bit controversy"[45]—as a result of many years of work in the military intelligence area, but the technology of computer security is less well developed[46], and nontechnological aspects—plant, personnel, and operational aspects—of network security are not in good condition at all.

The conclusion with respect to unification, federation, or fragmentation must be: we should strive for the kind of federation of networks that will provide coherent interconnection where needed and justified and, at the same time, provide informational privacy and security. That will require planning at national and international levels. It will require intensified research and development in network security and in internetting. And it will require an elevation of the ongoing discourse about privacy—to a level on which legislative and administrative policies can be defined clearly and networks can be designed responsively.

Privacy

It is now very widely understood that the collection of large amounts of personal information in computer processible data banks tends to jeopardize personal privacy. The main reason, of course, is that aggregation and computerization open up the possibility of invading privacy efficiently and on a massive scale. At the same time, they open up the possibility of protecting privacy tirelessly and algorithmically and of using the personal data effectively in the effort to accomplish the legitimate purposes for which the data were collected. Indeed, the stage is set for a battle between the forces of good and evil.

The stage is, however, not set in reasonable balance. The things required for the protection of privacy in information networks are policy and technology: legislative and administrative policy to define what is to be achieved and a technological basis for achieving it. In the U.S., the legislation is the Privacy Act of 1974 and the administrative policy is an OMB Circular[47]. Both are cast in terms of absolutes. The technological basis, as mentioned in the preceding section, is a combination of communications security and computer security. Because computer security is a relatively new and neglected subject, it is difficult to provide convincing assurance to an intelligent skeptic that any proposed interconnection of personal data, transmission channels, information processors, and interrogation-and-display facilities will not jeopardize privacy. Repeatedly, indeed, the advocates of proposed federal data networks have failed to present convincing analyses of the threats to privacy and of the trade-offs between privacy and mission effectiveness—and, repeatedly, their requests for permission to procure such networks (e.g., FEDNET, NCIC upgrade, IRS Tax Administrative System) have been sidetracked or denied with good reason.

At least in government circles, therefore, the issue of privacy is a very real and central network issue. Before it can be solved, three things have to be done. 1) The technology of information security has to be improved to the point at which reasonable analyses can be made and assurances can be given. 2) Network advocates have to develop plans and justifications that take privacy into account and provide strong assurances that it will be protected. And 3) the members of the oversight committees and their staffs have to face the fact that to protect privacy by precluding interconnection is not a very satisfactory solution for the long term. They should support and foster the accomplishment of steps 1) and 2) with the aim of receiving plans and proposals that they would not have to kill.

Other Issues

Space limitation precludes substantial discussion of other issues, but there are several that should be discussed. We shall discuss only a few of them, and those only very briefly.

Transborder Data Flow: Several European countries are beginning to restrict the flow of personal (or personnel) data across their bodies on the ground that they must protect the informational privacy of their citizens against threats implicit in data processing in countries with less stringent privacy laws. Some of the countries have laws or regulations that preclude the transmission of encrypted information through their public communication facilities. Many believe that such restrictions may be used to discriminate against foreign (e.g., American) data processing firms and against multinational corporations.

Technology Export and Import: International networks can be expected to facilitate greatly the transfer of scientific and technical information and know-how among the technologically developed nations. From a nationalistic point of view, one can see both advantages and disadvantages in such transfer to

Planning Ahead for Your Data Communications Applications

ideological, military, and economic competitors. The advantages are mainly humanistic and short-term economic. The disadvantages are mainly security-related and long-term economic. The interplay of advantages and disadvantages is giving rise to issues that will probably intensify.

Competition Versus Monopoly and Free Enterprise Versus Regulation: These are the issues of the "Bell Bill" (Consumer Communications Reform Act), Computer Inquiry II, and several recent decisions of the Federal Communications Commission that have favored competition in the telecommunications industry. How these issues are settled will to a large extent determine who operates the networks of the future in the U.S. and how such applications as electronic message service and "the office of the future" are implemented.

Nature of Office Work and Workforce: Some of the network applications we have discussed would tend to alter markedly the nature of white-collar work and the knowledge and skills required of members of the workforce. That fact will give rise to issues involving reeducation and retraining, pay commensurate with responsibility, and displacement of labor by automation.

Impact on Productivity: Many people are expecting that applications such as electronic mail and office automation will significantly increase productivity, but there are as yet few if any definitive experiences with such applications or quantitative models of them that will convince skeptics. Impact-on-productivity may become a major issue in and of itself.

Educational Applications of Networks: Packet-switched and satellite networks, together with the great advances being made in computers, appear to open the door to revolutionary improvements in education, but much more than mere access and mere hardware will be required to achieve truly significant results. The issue that is arising is whether the society values education enough to support the long and difficult effort that will be required to develop effective computer- and network-based methods—or whether there will be another wave of premature exploitation followed by disappointment as there was in the computer-assisted-instruction "revolution" of the early 1960's.

Networks Versus Stand-Alone Systems: Why do we need time sharing when everyone can have his or her own microcomputer? What good is a network when one can have a whole library on a video disk? Those questions have answers, of course, in such applications as electronic message systems, distributed but cooperating "offices of the future," and computerized commerce, but the questions will nevertheless constitute a major issue. Microcomputers and inexpensive digital storage devices have significantly changed the network concept. Less than a decade ago, a computer network was something that provided access to a time-sharing system. Now it is a facility to support communication among spatially distributed people and computers and to supply people and computers with common information bases and supplementary storage and processing capabilities.

World Leadership: An important latent issue is implicit in the fact that different people have quite different perceptions of the importance of networking. A significant fraction of the people who have had experience as developers or intensive users of a packet-switching network believe they have been in on the beginning of a new era and that descendants of the ARPANET will constitute the nervous system of the world. On the other hand, most of the people who now determine the kind of national policy that earlier fostered the merchant marine, the railroads, the airlines, and the interstates seem not to be aware that any significant new potential exists or that there may be any reason to move rapidly to take advantage of it. And, of course, if it is meaningful at all to the man in the street, the term "information network" still suggests the telephone system, the radio, or a television network. In that situation, it is difficult to project as an issue the importance of networks to world economic leadership. We believe, nevertheless, that it is such an issue, and we hope that it will soon be recognized as such an issue.

Totalitarian Control: If almost all the telecommunications in an area were based on computerized networks controlled by an organization—say by a government—then, in the absence of effective safeguards, that organization could map the life space of every individual and record the business transactions of every company. The notion of telecommunications in the hands of a "big brother with computers" goes beyond the bounds of what is usually called "invasion of privacy" into the realm of totalitarian control. One can detect at least a trace of the "big brother" issue in the coldness of certain members of Congress toward, and the rejection by the OMB of, the plans (mentioned earlier) of the IRS to develop a computer- and network-based Tax Administration System. The mere possibility of subversion was enough to kill the system. It is of the utmost importance, of course, to develop truly effective safeguards against misuse of networks for purposes of social control. But such safeguards will be more difficult to devise than safeguards against ordinary invasion of privacy or against fraud and theft. Networks will have to be designed in such a way that representatives of diverse interests can satisfy themselves that there is no subversion and that the audit trails are not dossiers. And the arrangements will have to be dictator-proof. We think that that is a very great task and that it is being neglected.

Planning Ahead for Your Data Communications Applications

Conclusions

Shakespeare could have been foreseeing the present situation in information networking when he said," . . . What's past is prologue; what to come, in yours and my discharge".[48] Most of the applications that will shape the future of networking are now in the stage of conceptualization or in the stage of early development. But it seems possible that a "network of networks" will, even in this century, become the nervous system of the world and that its applications will significantly change the way we live and work. The degree to which the potentials of networking will be realized will depend upon how we resolve some of the issues that have been discussed.

The value of information networks will depend critically upon their connectivity and their ability to connect any one of many sources to any one or more of many destinations. High connectivity will be precluded if conditions force the development of many separate, independent, incompatible networks. One condition that would force such an incoherent development is the combination of 1) a need for security against loss of "electronic funds" and (other) proprietary information and 2) the lack of a technology capable of providing security in an interconnected network or network of networks. That would lead to what we have called "fragmentation." Another such condition is based in a similar way on a combination of need for informational privacy and lack of the technology necessary to protect it except by isolating the privacy-sensitive data. The rapid and intensive development of computer and network security technology is vital to many network applications.

Many forces are fostering the development of networks to interconnect organization or branches of organizations and the development of applications to serve organizations, but there are few forces that foster networks to interconnect individuals or network applications to serve individuals. Perhaps the main hope for the provision of network services to individuals—especially network services to individuals at home—is that the Bell System will move (as obviously it would like to do) into the processing, storage, and information-commodity parts of the overall information business. But the telephone companies will be very slow to provide high-information-rate services because they have such large investments in narrow-band facilities. To get inexpensive wide-band channels into homes at an early date, we need a new departure in cable (or fiberoptics) communication, taking off from cable television, or something truly revolutionary like a nation-wide network of aerostationary platforms: microwave platforms at 70,000 ft, supported by helium plus helicopter vanes, and relaying signals from housetop "dishes" a meter in diameter.

Examination of 30 actual and potential applications of networks suggests that the following network characteristics or capabilities are especially important: bidirectionality, freedom from undetected errors, efficiency despite "burstiness" in the transmission patterns, inherently low cost, high connectivity, high information transmission rate, security, privacy, authentication, and reliability. Mobility and broadcast capability turned out to be of the lowest priority in our analysis. Packet-switching and time-division-multiple-access networks, especially such networks with satellite relays, were suggested by the analysis to have the patterns of characteristics required to serve best the full range of applications. The analysis suggests an approach to the selection of the best network to serve any specified application or set of applications.

REFERENCES

[1]T.H. Myer and C.D. Mooers, *HERMES User's Guide.* Cambridge, MA: Bolt Beranek and Newman, Inc., June 1976.

[2]A. Vezza and M.S. Broos, "An electronic message system: Where does it fit?" in *Proc. IEEE Conf: Trends and Applications 1976: Computer Networks,* New York, Nov. 1976.

[3]A. Vezza, "A model for an electronic postal system," in *Telecommunications Policy Research Conf. Proc.,* Bruce M. Owen, Ed. Palo Alto, CA: Aspen Institute Program on Communications and Society, Apr. 1975, pp. 146-153.

[4]*SIGMA Message Service: Reference Manual,* Information Sciences Inst., Univ. Southern California, Marina Del Ray, CA, ISI/TM-78-11.

[5]J.B. Rule, "Value choices in electronic funds transfer policy," Office of Telecommunications Policy, Executive Office of the President, Washington, DC, Rep. Domestic Council Committee on the Right of Privacy, Oct. 1975.

[6]T.M. Dunn and J.H. Mirrissey, "Remote computing—An experimental system. Part 1: External specifications," in *Proc. Spring Joint Computer Conf.,* 1974.

[7]J.M. Keller, E.C. Strum, and G.H. Yang, "Remote computing—An experimental system. Part 2: Internal design," in *Proc., Spring Joint Computer Conf.,* 1974.

[8]J.E. Sammet, *Programming Languages: History and Fundamentals.* Englewood Cliffs, NJ: Prentice-Hall, 1969, pp. 226-229.

[9]T. Marill and D. Stern, "The datacomputer—A network data utility," in *AFIPS Conf. Proc.,* Montvale, NJ: AFIPS Press, May 1975.

[10]L.G. Roberts and B.D. Wessler, "Computer network development to achieve resource sharing," *AFIPS Conf. Proc.,* May 1970.

[11]L.G. Roberts, "ARPA network implications," ARPA Network Information Center Doc., 12 982, 1972.

[12]R.M. Metcalfe, "Packet communication," M.I.T. Lab. for Computer Science, MAC/TR-114, Dec. 1973.

[13]"SNDMSG," Tenex User's Guide, Bolt Beranek and Newman, Inc., Cambridge, MA, Oct. 1977.

[14]J. Vallee, "The forum project: Network conferencing and its future applications," *Computer Networks (Netherlands),* vol. 1, no. 1, pp. 39-52, June 1976.

[15]M. Turoff, "Party-line and discussion: Computerized conference systems," Office of Emergency Preparedness, Washington, DC, Rep., Apr. 1972.

[16]— "Delphi conferencing. Computer-based conferencing with anonymity," in *Technological Forcasting and Social Change.* Washington, DC: National Resource Analysis Center, 1972, vol. 3, no. 2, pp. 159-204.

[17]M. Turoff and J. Scher, "Computerized conferencing and its impact on engineering management," in *Proc. 23rd Annu. Engineering Management Conf. on Effective Management of Engineering Resources,* Washington, DC, Oct. 1975.

Planning Ahead for Your Data Communications Applications

[18]D.C. Engelbart, "Coordinated information services for a discipline- or mission-oriented community," in *Computer Communication Networks,* R.L. Grimsdale and F.F. Kuo, Eds. The Netherlands: Noordhoff International, 1975, pp. 89-99.
[19]D.C. Englebart, R.W. Watson, and J.C. Norton, "The augmented knowledge workshop," in *AFIPS Conf. Proc.,* vol. 42, New York, June 1973.
[20]—, *Advanced Intellect-Augmentation Techniques.* Menlo Park, CA: Stanford Research Institute, Feb. 1972.
[21]—, "Human intellect augmentation techniques," Stanford Research Institute, Menlo Park, CA, Final Rep., 1969.
[22]R. Flesch, *The Art of Plain Talk.* New York: Harper & Brothers, 1946.
[23]J.O. Mason, Jr., "Management-information-systems-the auditor's role," *The Internal Auditor,* vol. 32, no. 5, pp. 40-48, Sept./Oct. 1975.
[24]S.J. Pokemper, "Management information-systems—A pragmatic survey," *Conf. Board Rec.,* vol. 10, no. 5, pp. 49-54, May 1973.
[25]R.G. Schroeder, "A survey of management science in university operations," *Manage. Sci:—Appl.,* vol. 19, no. 8, pp. 895-906, Apr. 1973.
[26]P.A. Strassman, "Managing the costs of information," *Harvard Bus. Rev.,* vol. 54, no. 5, pp. 133-142, Sept.-Oct. 1976.
[27]G.H. Barnes, R.M. Brown, M. Kato, D.J. Kuck, D.L. Slotnick, and R.A. Stoker, "The Illiac IV computer," *IEEE Trans. Comput.,* vol. C-17, pp. 746-757, 1968.
[28]W.J. Bouknight, S. A. Denenberg, D. E. McIntyre, J. M. Randall, A. H. Sameh, and D. L. Slotnick, "The Illiac IV system," *Proc. IEEE,* vol. 60, pp. 369-388, 1972.
[29]"Self-teaching design and analysis tools offered on timeshared basis," *Comput. Des.,* vol. 17, no. 2, pp. 18-20, Feb. 1978.
[30]I. E. Sutherland, *Computerized Commerce,* Santa Monica, CA: The Rand Paper Series, Rand Corp., Sept. 1975.
[31]National Research Council, Telecommunication for metropolitan areas: Near-term needs and opportunities. Washington, DC: Nat. Acad. Sci., 1977.
[32]L. L. Weed, "Technology is a link, not a barrier, for doctor and patient," *Modern Hospital,* vol. 114, pp. 80-83, Feb. 1970.
[33]—, *Medical Records, Medical Education, and Patient Care. The Problem-Oriented record as a Basic Tool.* Chicago, IL: Year Book Medical Publishers, Inc., 1970.
[34]R. Davis, B. Buchanan, and E. Shortliffe, "Production rules as a representation for a knowledge-based consultation program," Stanford Univ. Artificial Intelligence Lab., Stanford, CA Memo AIM-266, Oct. 1975.
[35]H. Silverman, "A digitalis therapy advisor," M.I.T. Lab for Computer Science, MAC/TR-143, Jan. 1975.
[36]"GECOS: General Comprehensive Operating Supervisor," Honeywell Information Systems, Waltham, MA, Rep. DD19.
[37]A. Stathopoulos and H. F. Caley, "The Autodin II network," in *EASCON-77 Rec. Proc. IEEE Electronics and Aerospace Systems Convention* (Arlington, VA), Sept. 1977.
[38]Social Security Administration, Office of Advanced Systems, *Recommended Design Concept for the Future SSA Process,* Washington, DC, OAS Pub. No. 005, Apr. 1977.
[39]V. Slamenka, private communication, May 1978.
[40]CCITT, "Recommendation X.25: Interface between data terminal equipment (DTE) and data circuit-terminating equipment (DCE) for terminals operating in the packet mode on public data networks," in *Public Data Networks, Orange Book* (vol. VII.2). Geneva, Switzerland: Sixth Plenary Assembly, International Telecommunications Union, 1977, pp. 70-108.
[41]W. Diffie and m. Hellman, "New directions in cryptography," *IEEE Trans. Inform. Theory,* pp. 644-654, Nov. 1976.
[42]R. Rivest, A. Shamir, and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems," M.I.T. Lab. for Computer Science, LCS/TM-82, Apr. 1977.
[43]*Federal Register,* vol. 40, no. 52, Mar. 17, 1975.
[44]*Federal Register,* vol. 40, no. 149, Aug. 1, 1975.
[45]W. Diffie and M. Hellman, "Exhaustive cryptanalysis of the NBS data encryption standard," *Computer,* vol. 10, no. 6, pp. 74-84, June 1977.
[46]J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems," *Proc. IEEE,* vol. 63, pp. 1278-1308, Sept. 1975.
[47]Office of Management and the Budget, OMB Circular A-108, Washington, DC, July 1975.
[48]William Shakespeare, *The Tempest,* act II, scene I.□

A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs

Problem:

If we were asked what is the most difficult part of planning, designing, and operating a data communications system, our microsecond-quick response would be Tariffs. Tariffs, in turn, are supported by the huge, unwieldy, creaky, rusting machinery of Regulations. Unfortunately, the Regulation/Tariff problem couplet can't be ignored because you could wind up in jail, be heavily fined, or have your transmission service cut off if you decide to disregard the regulations or not pay your bill, and the problem won't be washed away, as many other communications problems are, by the cleansing flow of Technology. More unfortunately, transmission line costs are becoming a major component of system planning, not because they are increasing but because most other costs (particularly hardware) are shrinking rapidly and significantly relative to transmission costs.

The current regulatory burden is a complex heritage of limited 19th Century perspectives, a schizophrenic attitude toward monopolies, and a well-intended but shortsighted attempt to legislate for the public good. The equally complex Tariff tangle is the result of well meaning efforts by the common carriers to distribute costs fairly and within the spirit of the Regulations while assuring themselves a profit. The complexity of the Tariffs mirrors the thousands upon thousands of service options that have evolved out of what seemed to have been a definitive solution to a simple problem back in 1934.

This report is historical in the sense that it explains where we are now, but "where we are now" is probably adequate for at least a five-year plan because it will take the Government easily that long to change the Regulations, for the common carriers to interpret the Regulations, and for the changes to filter down to the users in the form of different Tariffs. Hopefully, you can plan for substantially lower Tariffs by the mid-80's.

Solution:

Over the intervening almost half century from 1934 to now, common carrier regulations and communication tariffs have become quite complex, requiring much interpretation by those who would begin to understand them. This is certainly the impression one often has after discussing some special need, innovative idea, or critical communication requirement with a communications common carrier. The presentation here is intended to shed some light on

From the *Automatic Data Processing Handbook* edited by The Diebold Group. © 1977 by McGraw-Hill Inc. Reprinted by permission of McGraw-Hill Book Company.

A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs

common carrier regulation and several related tariffs in order to give communications managers and system designers additional insight into the alternatives open to them in the design and operation of a data communications system.

A thorough understanding of the regulatory and legal aspects of communication is equally as important to the operation of a large telecommunication network as is a thorough understanding of communication technology, engineering, operations, administration, and maintenance. In this report we will look primarily at the underlying regulatory structure of common carrier tariffs and their provisions in an effort to gain an understanding of the services offered to the data communications customer and the techniques available to the customer for minimizing costs.

CARRIER REGULATION

Regulation of the supply of communication services, as we know it today, was instituted by the Communications Act of 1934, which created the Federal Communication Commission and defined its authority for the regulation of interstate communication. The supply of common carrier communication services serving locations within a given state (intrastate) is regulated by each state's public utility or public services commission. Under the 1934 act, communications common carriers within a given area are granted monopolistic power and are required to file before the appropriate regulatory bodies Certificates of Necessity and Convenience as well as tariffs for services. The purpose of these tariff documents is to record the products and policies of the communications vendor (e.g., the telephone company). In effect, any standard service that a communications vendor offer is described and priced in its tariff.

Federal vs. State Regulation

As a general rule, the state regulatory commissions follow the precepts of their Federal counterparts (or, perhaps more exactly, the Bell operating companies within each state follow the lead of AT&T). There are, however, several fine distinctions worth noting in this relationship:

1. As a general rule, changes in intrastate services or the introduction of new services within a state follows the interstate service filing, but lags from several months to several years behind it.

2. Prices (rates) are generally higher for intrastate than for similar interstate services.

3. As a matter of philosophy, a common carrier service is generally considered intrastate (subject to state regulation) if it can be used for communications wholly within the state. For example, the telephone handset and attached line at your home or office are considered intrastate services even though you may call locations in other states. Only the rates for interstate calls are subject to Federal regulation. This fact has interesting ramifications when one studies the interconnection of various devices to the telephone network.

The distinction between intrastate and interstate communication regulation may seem inconsequential. However, consider for a moment its effect on a specific communication requirement (voice-grade, lease-line, and series 3000) between San Francisco, Los Angeles, and Phoenix. The situation is demonstrated graphically in Figure 1.

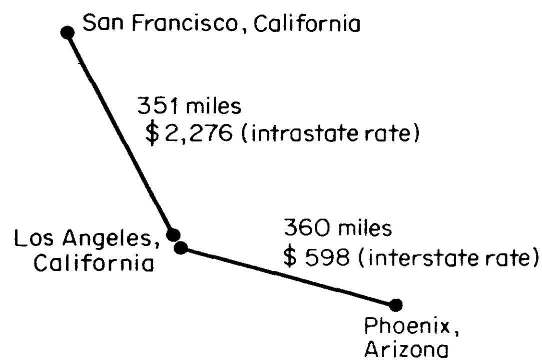


Figure 1. Intrastate and interstate legs of a network compared

The surprising point is that in this case, the intrastate leg of the network is over 3.7 times as expensive as the interstate leg of approximately the same length. In fact, for the price of the intrastate line from San Francisco to Los Angeles, you can get an interstate line from Los Angeles to New York. Even more interesting is the fact that if you request the carrier to install a switch linking the lines in Los Angeles, the San Francisco to Los Angeles link becomes interstate under the tariff, and its cost drops to $589 per month. The overall system savings is $1,627 per month, or over 73 per cent. In order to install this switch, you must have a legitimate requirement for through service between San Francisco and Phoenix.

TARIFFS

Several common interstate tariffs filed with the FCC which specify communication services are:

Federal tariff	Subject
254	Western Union Private Line Services
255	AT&T Administrative Rate Centers and Central Offices
259	Wide Area Telecommunication Service (WATS)
260	AT&T Interstate Private Line Services
263	Long-Distance Message Telecommunications Service

A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs

Because these tariffs are representative of many common principles of communications regulation, and since state tariffs tend to follow Federal tariffs, it is instructive for our purpose here to study specific clauses of a single tariff. Therefore, let us look at AT&T's Interstate Private Line Tariff, FCC 260.

FCC Tariff 260

Usage Provision. One of the most significant provisions in this tariff specifies the allowable uses of a leased, point-to-point, voice-grade telephone line.[1] Basically, this provision says that as a telephone service customer, you are permitted to use this service for transmission of communications to or from yourself, where such communication is directly related to your business. The provision also allows transmission relating directly to the business of a wholly-owned or controlled subsidiary, or simultaneous transmission relating directly to matters of common-interest parties who are in the same general line of business and who are connected to the communications facility. In addition, the facility must be terminated (at least on one end) on your property.

The important point to note here is that the tariff states that you may use your communications facilities for yourself and may not provide a communication service to others. The only exception to the above rules are for joint use (discussed later), for governmental use, for members of an electrical power pool, and for certain use by a licensed aeronautical communications company.

Reliability and Responsibility. For damages arising out of mistakes, omissions, interruptions, delays, or errors associated with the purchase of communications service, the tariff specifies that the common carrier assumes no liability greater than the proportionate charge for the service period in which the defect occurred. The courts, of course, have held that these tariff provisions do not protect the common carrier in the event of its negligence.

It is interesting to note that tariffs say nothing about the quality of service the carrier must offer its users. There are no minimum provisions for reliability, error rates, noise levels, etc. Tariff 260 does specify the bandwidth and delay distortion characteristics for the various levels of conditioning of a voice-grade line. However, this is a far cry from actual specification of the communications channel characteristics that determine the reliability of transmission.

Service from a common carrier is generally provided on a 1-month-minimum-period basis, with pro rata charges for fractional portions of a month. Generally, credits for service interruptions are given on the pro rata basis shown in Table 1, assuming a 30-day month.

Interruption	Credit
Less than 30 minutes	None
30 minutes to 3 hours	1/10th day
3 hours to 6 hours	1/5th day
6 hours to 9 hours	2/5th day
9 hours to 12 hours	3/5th day
12 hours to 15 hours	4/5th day
15 hours to 24 hours	One day

Table 1. Prorated credits for service interruptions

It is the customer's responsibility under the tariff to provide space and power for common carrier facilities installed on premises, and to provide access to common carrier employees for the installation and maintenance of the equipment. In addition, of course, the customer is responsible for all payments and, in some instances, the common carrier is allowed to require a 2-month advance deposit as a guarantee for payments. The common carrier, however, must pay interest on the deposit.

The common carrier, by written notice to a customer, may immediately discontinue providing communications service for nonpayment or for violation of various conditions governing the furnishing of service. Disputes leading to service termination over the past few years have centered around violations of interconnection provisions or usage provisions of the tariff.

Joint Use. The "Joint Use" provision of the tariff allows a communications customer to share his or her common facilities with other "shared users" where each such user has a legitimate communications requirement as defined under the Usage provision discussed above. The important criteria here are (1) the primary user must have a legitimate requirement for the facilities (i.e., communications to or from this user and directly relating to his business), as must each joint user, and (2) the joint user must have a station or service terminal on his premises and a "through" connection to all portions of the network being shared. Each joint user is billed by the common carrier a pro rata charge for his share of the common facilities, as specified by the primary user. In general, there is also a charge for each individual joint user connected to a "shared user network."

The general intent of the Joint Use tariff provision is to allow increased utilization of the telecommunication resources of the country while precluding the resale for profit of communication services by entities other than common carriers. In effect, the primary user is allowed to defray communication costs by letting joint users share unused facilities without allowing

[1]Generally used for dedicated data transmission up to 9,600 bps.

the primary user to profit by the supply of a communication service. There are some services where joint use is precluded. For example, joint use is specifically not allowed on WATS, foreign exchange lines, and certain wideband facilities. In considering any specific joint-use arrangement, it is recommended that approval from both the common carrier and the appropriate regulatory bodies be procured.

SUPPLY OF COMPUTER/COMMUNICATIONS SERVICE

In March, 1972, the FCC released its Final Decision and Order with respect to regulations that control the independent relationship between computer and communication service. This action was reviewed by the Second Circuit Court of the U.S. Court of Appeals which, by its decision of February 1, 1973, affirmed the Commission's ruling with certain minor modifications.

As a direct result of the Final Decision, the Commission concluded that no public interest would be served by the regulation of data processing services, whether or not the sale of those services involves the use of communication facilities to provide access by the customer to the computer of the data processing supplier. On the contrary, the Commission decided that the existing market for the sale of data processing services was inherently competitive and that this market would develop and flourish best in a competitive environment. With respect to communications common carriers, the Commission found no reason to require the carriers to withdraw from the data processing market. In fact, the Commission felt that the competitive environment might well benefit from carrier participation under appropriate conditions.

Thus, the Commission invoked the doctrine of "maximum separation" to ensure that the regulated activities of the carrier are in no way commingled with any of its non-regulated activities, including data processing. The rules adopted by the Commission and affirmed by the court provide that a common carrier must set up a separate corporate entity to furnish data processing to others. That separate entity must maintain its own books of account, have separate officers, and utilize separate operating personnel and separate equipment and facilities. Further, the Commission's regulations bar a carrier from selling, leasing, or otherwise making available to any other entity any capacity of a computer used by the carrier in any way for the provision of its common carrier communications services.

Essentially, the degree of separation required by the Commission was based on the following regulatory premises:

1. That the sale of data processing services by carriers should not adversely affect the provision of efficient and economic common carrier services.

2. That the costs related to the furnishing of such data processing services should not be passed on directly or indirectly to the users of the common carrier services.

3. That revenues derived from common carrier services should not be used to subsidize any data processing services.

4. That the furnishing of such data processing services by carriers should not inhibit free and fair competition between communication common carriers and data processing companies, or otherwise involve practices contrary to the policies and prohibitions of the antitrust laws.

To maximize separation, the Commission also ruled that no carrier may engage in the sale or promotion of data processing activities on behalf of its data processing affiliate.

However, the versatility of computers readily enables their use both for data processing and for switching of messages among terminals connected by communication channels to the computer. Message switching including the storage and forwarding of correspondence, is regarded by the Commission as inherently a communication operation. When performed as a service for hire the services are therefore subject to regulation as common carrier service. Thus, the intriguing issue confronting the Commission is whether or not to regulate a specific computer entity offering a mix of data processing and message switching services.

In effect, the Commission's decision ruled that hybrid services, namely those combining into a single integrated service both data processing and message switching procedures, would be subject to regulation if their primary purpose was communication. On the other hand, to the extent that message switching is merely incidental to the sale of data processing, the hybrid service would not be subject to regulation Clearly the regulatory treatment of hybrid services does not lend itself to generalized formulas. The possible combinations and permutations of different services capable of being amalgamated in a common service offering by a computer entity are too many and too diversified for categorical definitions. It is for this reason that, subject to the very generalized guidelines mentioned above, determination by the Commission concerning hybrid services will be made on the basis of review and evaluation of particular factual situations as they develop and is the essence of the problem it now faces with AT&T's filing for an Advanced Communications Service (ACS).

A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs

A CHOICE OF ALTERNATIVES

We have looked briefly into several specific current tariff provisions, as well as some considerations of the overall regulatory environment that currently prevails. The keyword of this environment is one of competition and an increase in the alternatives open to the computer communications system designer and manager.

As a result of the FCC's interconnect decision in the Carterfone case, users have a much wider range of equipment options. This has created a new interconnect industry, which, in the best tradition of a market economy, has introduced new equipment and lower prices. The new interconnect policy has also galvanized Bell and other common carriers into competitive marketing efforts. Bell has revised its prices on existing modems and has introduced a new series of modems. It has marketed a new family of PBX equipment, key telephone systems, and CRT terminals. It has also responded with a re-evaluation and revision of its tariffs dealing with leased private lines, PBX hardware, and station equipment.

In the specialized common carrier field, new carriers are mainly engaged in the construction of their systems and the initial offering of services. Here again, Bell has responded on a broad front by restructuring the existing tariffs and by installing a new end-to-end all-digital service. With the introduction of these competitive carriers has come the dissolution of the age-old concept of "nationwide rate averaging." Further, the FCC has approved open entry for domestic satellite systems which will undoubtedly result in the establishment of several competitive long-haul satellite transmission facilities by the early 1980's.

The provision for competition in what has traditionally been a regulated monopolistic industry has mobilized great concern for assuring a fair competitive environment, where cross-subsidy from a carrier's noncompetitive products to its competitive products is closely monitored. Both the carriers and the FCC are taking broad steps to prepare for this increasingly competitive marketplace. For example, common carriers are showing significant new interest in cost-related pricing, and all carriers must have greater knowledge of their costs in order to meet the outside competition. This is promoting greater operating efficiency within the carriers and an increase in the concomitant benefits to the user.

THE INTERNATIONAL REGULATORY SERVICE

A user in Western Europe wishing to have a data communication link joining two locations on separate sites has a multiple choice as to how this will be achieved, but basically there are two choices. The user can either install a private link (subject to any regulations that apply), which could be a line or a radio link, or hire the facilities from the established communication authority—usually the telephone company. The former choice is usually not possible on the ground of expense except for short-line links such as exist on one site (a university campus, for example). The latter choice is therefore almost the universal one. Technically, it may not be the best choice, but financially it is usually the only choice. Thus, with few exceptions, data communication links and networks are based on channels that were originally designed and intended for use as voice channels, i.e., telephone channels, and consequently the regulations that have evolved are specifically related to their use as such. Regulations are designed to prevent interference with the satisfactory operation of the voice channel and are discussed in some detail later.[2]

Types of Lines

A line is dedicated or leased when it is permanently connected between two or more locations, or the link may be provided by the public switched network (PSN) on dial-up operation. In general, different regulations apply, those for dial-up operation being more stringent than those for leased-line operation. Interference from a leased line can be caused only by crosstalk, whereas on the PSN there is direct connection to the telephone system.

REGULATORY AUTHORITIES: PTT

Throughout Western Europe, telephone networks are owned and administered by a government department responsibile for Posts, Telegraphs, and Telephones. Although the actual title of the department varies from country to country, they are known collectively as the PTTs, and reference to "the PTT" in any country is understood to mean the department concerned. In general, one PTT covers a complete country, but in Finland and Yugoslavia, for example, where there are a number of regional authorities, this is not so.

Each PTT is completely autonomous and has absolute right to determine what equipment it shall permit to be connected to its telephone system. In practice, PTTs do not exercise their rights in an autocratic manner. Each publishes its requirements, and, provided that these requirements are met, permission to connect is not withheld.

PTT Approval or Homologation

Before any equipment is connected to equipment owned by the PTT, it must be approved by the PTT. Stiff penalties may be applied to unauthorized connection. Note that it is equipment directly connected to the telephone line which is primarily subject to approval. In data communications, it is the modem (or data set) that must be approved. Other equipment making connection to the line only through the modem may not require approval depending on its nature and on the PTT concerned.

Each PTT publishes regulations and requirements, which must be followed. The technical requirements are concerned with three aspects:

1. Safety of plant and personnel.

2. Interference with other equipment.

3. Correct performance.

The safety aspect is concerned with preventing dangerous voltage levels being applied to PTT equipment. In particular, it must not be possible for the power supply voltage, which in Europe is generally in the range of from 210 to 250 volts at 50 Hz, to be applied to the telephone line either under normal operating conditions, or under fault conditions. The PTT inspects the specifications and construction of the power transformer and of the line transformer to satisfy itself that they are suitable for the purpose.

To prevent interference with other equipment, the PTT specifies the frequency bands and maximum level of the signal transmitted to line. This ensures that common equipment is not overloaded and that a continuous tone could not be confused with a control tone used by the PTT and thus cause spurious operation.

Procedure for Obtaining Approval

It is usual for approval to be requested by the manufacturer or supplier rather than by the intending user. The PTT normally requires this, as the user may not be in possession of sufficiently detailed technical information relating to the equipment, and would therefore require a statement of compliance from the manufacturer. Thus, responsibility for ensuring that equipment complies with the technical requirements is more easily placed on the manufacturer or on the manufacturer's appointed representative. From the manufacturer's point of view, it is much more satisfactory to deal directly with the PTT. This also has the advantage that the PTT has a smaller number of people to deal with, and offers the manufacturer the opportunity of reaching a close understanding of the requirements of the PTT and of establishing a beneficial relationship with the PTT based on mutual trust.

While business users of data communications are not likely to find themselves in the situation of having to make their own PTT approval application, they should understand the fundamentals and place their reliance on the manufacturers. In most cases the manufacturer will already have obtained approval as a necessary preliminary to marketing operations.

The first action is for the manufacturer to prepare a detailed technical document that describes the equipment for which approval is required. Manufacturers with previous experience are best fitted to do this, as they know the type of informtion and form of presentation preferred. The document must give sufficient detail of the equipment's operation, function, and performance, to enable the PTT to assess its effect when connected to line. Thus the technical description includes block schematic diagrams, operating instructions, power output levels and frequencies, specifications of transformers, and so on. Such a document does not normally form part of a manufacturer's publications, as it is neither a sales data sheet nor a technical manual for operation and maintenance. Nor is it yet a design manual. Nevertheless, it is a very important document, and a well-prepared one can be of great assistance in obtaining approval. The same information can, of course, be used for all PTTs, so that the investment made in it can be recouped.

Some PTTs, particularly those of the more nationalistic nations, require the approval document to be presented in the language of the country concerned.

The approval document is formally presented to the PTT with a request that the equipment to which it refers be considered for approval. Having inspected the specifications and found them to be satisfactory, the PTT then requests the loan of equipment for evaluation. Such a loan enables the PTT to inspect the construction of the equipment and to carry out any tests it wishes as a check that its performance complies with the previously submitted specification. If all is well, the PTT then issues a letter or other document to the applicant, formally stating that the equipment is approved, and it sets forth the conditions that apply to its use. Usually the approval is given a reference number. The PTT makes a charge for the evaulation the charge is payable by the applicant—who, as a manufacturer engaged in commercial operation, would not normally charge a customer.

[2]Author's note: In this report, basic concepts and philosophies of communications regulations are discussed, and specific detail is given only by way of illustration and example. There are two reasons for this: First, the number of individual regulatory authorities concerned is large, generally one per country; and second, the situation is by no means static. It could be misleading to make statements purporting to be factual which may well be outdated by the time the report is published. All comments made are believed to be true at the time of writing. Intending users should check either with their equipment supplier or the local regulatory authority before making connections to a nonprivate line.

A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs

If, during the initial study of the approval document, the PTT finds the equipment has a feature or characteristic not acceptable to it, the applicant is then informed and allowed the opportunity of modifying the equipment accordingly and resubmitting the application. If, during the evaluation loan, the PTT decides that the equipment is not acceptable, it may either reject it or grant an approval subject to certain changes being made in the equipment. For example, the equipment may permit adjusting the transmitted output level, and thus the maximum output level may exceed that permitted by the PTT. Although it is possible for the level to be adjusted to be less than the maximum permitted level, the PTT may insist on the equipment's being modified so that the output level is fixed at less than the permitted level, or so that when adjusted to give its maximum output, the latter is less than the permitted level. Normally, approval would not be withheld for such a reason, but the applicant would not be allowed to connect equipment not modified as required. Some PTTs require a label to be fixed to the equipment, stating that it has been so modified.

The time taken to process an approval application varies with the type of equipment, the PTT concerned, difficulties encountered, and the workload at the time. An average time is 2 to 3 months. In the planning stages of a new network it is essential that sufficient time be allowed for approvals to be obtained for equipments requiring them.

It is not permissible to connect approved equipments without authority of the PTT. Normally this is a formality, a courtesy even, but it is the means PTTs adopt for keeping themselves informed of equipments connected to line. It is sufficient to give approval reference numbers where they exist, and to give details of the line to which connection is to be made.

Permissible Operations

In the above discussion, it has been assumed that approval has been required for a specific modem, but we must now return to the general situation and examine that more fully. It was stated earlier that user's have a choice of either installing their own lines or renting from the PTT. This is generally true, but the situation is somewhat more complex than that simple statement implies. It must first be established whether the intended mode of operation is one that is permittted by the PTT. If it is, the next point is whether or not the PTT operates a monopoly, and if not, whether the equipment is approved by the PTT. Interwoven with the above is the question of whether the line is to be leased, or whether operation will be by the PSN.

Most PTTs are monopolistic on PSN operations—that is, modems for connection to the PSN must be obtained from the PTT and modems from other sources are not permitted. Certain PTTs do not permit access to a multiplexer over the PSN. Thus it is important to establish at an early stage, preferably directly with the PTT or PTTs concerned, whether the intended mode of operation is permitted in principle.

For operation over leased lines, there is less tendency to operate monopolies, but the trend is for PTTs to extend their monopolies, presumably for commercial reasons. In Italy there is a complete monopoly at all speeds; whereas in the Netherlands there is no monopoly. However, even where there is no monopoly, the PTT always requires a modem to be approved before it is connected to line.

A time division multiplexer falls into an interesting category. Because it is not connected directly to line but via a modem, some PTTs take the view that approval is not required (although it must be checked that multiplexing is a permitted mode of operation). A few PTTs require the multiplexer to be approved.

International Links

The above discussion has been in reference to links wholly contained within a specific country. There are many European links that connect locations in two or more countries, some of which connect to locations in the United States. Of particular concern are those links that have a location within a country where the PTT exercises a monopoly at the intended operating speed. This implies that the modem must be obtained from the PTT, but this modem may not be on-line compatible with the modem it is desired to use in the other countries. In certain countries (Sweden is an example), the PTT permits the use of an approved modem for international links, even though it operates a monopoly for internal links.

SUMMARY OF INTERNATIONAL PROCEDURES

Intending users of data communication within Europe should first check with the PTT or PTTs concerned regarding the intended mode of operation to ensure that it is permitted. They must also check whether the PTT exercises a monopoly at the intended speed of operation and, if not, they will then check with the modem supplier of their choice as to whether the modems are approved or if approval can be obtained. It is the responsibility of suppliers to obtain approval; it is the responsibility of users to notify the PTT of the intended connections.

The regulations governing the connection of modems to telephone lines are numerous and complex. Fortunately users do not need to be familiar with them in detail, for they will rely upon the PTTs in monopoly situations or upon suppliers otherwise.

A Short Guide to the Arcana of National and International Data Communications Regulations and Tariffs

Austria

Bundesministerium für Verkehr und
Verstaalichete Unternehmungen
Generaldirektion für die Post und
Telegraphenverwaltung
1011 Wien
Postgasse 8

Belgium

Regie van Telegrafie en Telefonie
1030 Brussel
Paleizenstraat 42

Denmark

Ministeriet for Offentlige Arbejder
Generaldirektoratet for Post og Telegrafvaesenet
Favergade 17
1007 Kobenhavn K

Eire

Department of Posts and Telegraphs
Marlborough Street
Dublin 1

France

Ministere des Postes et Telecommunications
Direction Generale des Telecommunications
20, Avenue de Segur
Paris 7

Germany

Deutsche Bundespost
Fernmeldetechnisches Zentralamt
61 Darmstadt
Postfach 800

Italy

Ministero delle Poste e delle Telecomunicazioni
Ispettorato Generale delle Telecomunicazioni
Direzione Centrale Telegrafi
00100 Roma

Luxembourg

Administration des Postes et Telecommunications
Division Technqiue
17 rue de Hollerich
Case Postale 2061
Luxembourg

Netherlands

Staatsbedrijf der Posterijen
Telegrafie en Telefonie
Centrale Directie
Kortenaerkade 12
4S—Gravenhage
Postgirorekening 45100

Norway

Teledirektoratet
Universitetsgata 2
Oslo 1

Spain

Compania Telefonica Nacional de Espana
Avda de Jose Antonio 28
Madrid 13

Sweden

Televerkets Centralforvaltning
Projekteringsavdelningen
Marbackagatan 11
123 86 FARSTA
Stockholm

Switzerland

Schweizerische Post, Telephon und Telegraphenbetriebe
Generaldirektion
Viktoria strasse 21
3000 Bern 33

U.K.

Post Office Telecommunications Headquarters
Tenter House
45 Moorfields
London EC2Y 9TH

Table 2. Summary listing of major European PTT addresses

Table 2 lists the addresses of the major European PTT's. Contact them directly if you are planning for international operations.

RELATED READINGS

Greenberger, Martin, *Computers, Communications, and the Public Interest,* Johns Hopkins, Baltimoare, 1971.

Jones, Malcolm M., Chap. 26, "The FCC Inquiry and the Data Communications User," in The Diebold Group, *Rethinking the Practice of Management*, Praeger, New York, 1973.

Trebing, Harry M., *Essays on Public Utility Pricing and Regulation,* Michigan State University Press, East Lansing, 1971.□

A Short Checklist of Terminal Planning Factors

Problem:

We cannot overstress the importance of the man-machine contact points—the terminals—in the planning, design, and even the operational stages of a system's life. Every communications system must ultimately report to its users through at least one of these points for a very simple man-machine linkage and very likely through many ports in more complex man-machine-man relationships. True, the terminal planning/design problems do not rest solely on the communications system designer's shoulders because terminal features are also important considerations for other departments in the company, but this lack of total responsibility actually aggravates the problems because all of the departments must work together closely to find the best compromises between cost and general utility. However, the communications planner/designer is probably in the best overall position to coordinate the terminal planning/design job since his is the only view that accounts for the needs of all the using departments.

This report is offered as a checklist to remind you of certain terminal planning and design considerations that extend beyond the basic mechanical considerations of cabling and protocols. Some of them, like terminal-user dialogue formats, may not be your direct responsibilities, but you must be aware of them to fulfill your total responsibilities for terminal planning.

Solution:

When the terminal is manned, human operating factors must always be taken into consideration. In some systems they are not a dominating consideration; they would not be, for example, with a terminal designed for the batch transmission of punched cards. In some systems, however, the human factors become all important because success or failure depends on how well men can communicate with the system and it with them.

From *Systems Analysis for Data Transmission* by James Martin, Chpt. 6. pp. 55-59 and Chpt. 13, pp. 151-179. © 1972 by Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Where the system is primarily oriented toward communicating with human beings, the design must have begun with the planning of the man-machine interface, and this planning start will have determined the types of terminal selected, the communication line speeds, the probability of obtaining the desired service at the first attempt, and the response times to be achieved. These factors will also have predicted the choice of the data transmission network structure.

In this report we will review the most important terminal-user considerations to perhaps uncover any lingering flaws in the man-machine interface.

A Short Checklist of Terminal Planning Factors

DEDICATED OR CASUAL OPERATOR?

On some real-time systems, the operator spends the whole of her working day sitting in front of the terminal. This operator can be specially trained for the job, perhaps with a lengthy training program. She will have plenty of time to practice her interaction with the machine, to learn its language, and to become accustomed to its idiosyncrasies.

On other systems, the terminal is used only occasionally by someone, like a salesman or a manager, who spends most of his day doing entirely different work. This person is not highly trained in terminal usage. He will be easily confused by unclear terminal responses and easily frustrated by lengthy response times. It is probable that an increasing proportion of terminal operators will fall into this category in the years ahead. For them, the man-machine interface must appear as natural as possible; otherwise bewilderment will quickly turn into annoyance, criticism, or behavior that amounts to rejection of the system.

As we shall see in the next report the man-machine conversation is often designed very differently for these two categories, and the difference in design is important to the organization of the communication line network.

OPERATOR WITH PROGRAMMING SKILLS?

On some terminals, the operator uses a programming language. It may be one of the standard programming languages and a set of program statements devised for the application. On the majority of terminals for commercial use, however, programming is not used.

Operator skills might be categorized as in Table 1, and for each division of the table the conversation structure would be different.

	Dedicated Operator	Casual Operator
Operator with programming skills		
Nonprogrammer operator with high IQ and detailed training		
Nonprogrammer operator without high IQ but with detailed training		
Operator with little training		
Totally untrained operator		

Table 1. Operator skills

MEDIA USED

The most common means for input at a terminal for human interaction are a keyboard, switches, knobs, a badge reader, or a light pen. The most common means for output are printing, display of characters on a screen, display of lines or curves of a screen, display of photographic images—for example, from slides or film frames at the terminal—voice answerback, and combinations of these devices.

The systems analyst must select the most suitable media for his particular application, taking into consideration terminal cost, communication network cost, and user psychology.

LANGUAGE AND RESPONSE STRUCTURE

Having selected the media, the analyst then plans the language and response structure. What exactly will the man say to the machine and in what form, exactly, will it respond?

There are endless possible structures for terminal conversations. The choice will depend on the nature of the application, the terminal, and the categories of operator. In the planning of terminal applications, other than the simplest, it has generally taken many months to define the input and response structure. The more successfully this is done, the greater the chance of the system being widely accepted and efficiently employed by its users.

If we narrow our observations to a system on which the input and the responses are alphanumeric, we can categorize a variety of techniques for facilitating man-machine conversation. It will usually be true that techniques designed to make the terminal as easy as possible to use will result in many more characters being transmitted. On a far-flung system, this situation can result in a much more expensive communication network.

SPEED OF TERMINAL

The speed of the terminal has a major effect on its possible interactive users. Suppose that a response of 500 characters must be sent. This might be a well-formatted table, so that many of the characters are blanks. With a teleprinter operating at 7.5 characters per second, the time taken to print this response would be more than 1 minute and 6 seconds. With a visual-display unit operating on a 4800 bit-per-second line, the time could be less than a second. A conversation involving many lengthy responses would be prohibitively slow with the teleprinter. Even with faster printing devices operating at 15 or 30 characters per second, it would be very frustrating and would often inhibit creative thought at the terminal. The high speed of the visual-display screen is needed for applications with multiple lengthy responses.

A Short Checklist of Terminal Planning Factors

RESPONSE TIME

Certain applications are very sensitive to the response time at the terminal. For some terminal uses, a very fast response time is necessary; for others it is not. For the applications needing it, fast response time is not simply a luxury. Research by behavioral psychologists has shown that it is essential for certain terminal actions, including problem solving, creative processes, complex interrogation, and, in many cases, plain business uses of terminals.

AVAILABILITY

This category refers to the probability that the terminal is usable at the time the operator wishes to use it. With some forms of system design, he may obtain a busy signal occasionally, just as he may on the telephone. Such a signal could be caused by full occupancy of all the lines or all the "ports" into the computer, or all the paths through or registers in an exchange. It might also be caused by overload in a computer or concentrator.

The probability of not obtaining service when it is requested is sometimes referred to as "the grade of service." Some systems are designed so that the terminals will always receive service when they request it, except in a period of equipment failure. This, however, may be unnecessarily expensive, especially when the terminals have low utilization. A figure for "grade of service" should be decided on by the systems analyst and the data transmission facilities should be planned so as to achieve that figure. Too low a system availability will demoralize its users and sometimes cause them to avoid using the system. In some installations, utilization of the system has grown more rapidly than anticipated and the grade of service has dropped severely. If this condition can be anticipated, the designer can take steps to protect the system from degradation.

CONTROL OF ERRORS

When files of data are being built up on a real-time system, the information often comes from large numbers of terminal operators. In many cases, the terminal operator is less trained and less accurate than the keypunch operator on a batch system. Unless measures are taken for controlling their accuracy, substantial erroneous data could be entered into the system.

Again, the likelihood of error will depend to a large extent on human factors. A well-designed conversation structure minimizes errors; and an on-line system has the great advantage that many errors can be caught when they are made and the operator notified immediately. Error-detecting formats can be devised, and the information being entered can be checked against existing files. On-line data collection systems have been installed in factories to give a fraction of the errors that equivalent off-line systems had. Besides notifying the operator of suspected errors, the system can inform a control center or supervisor if desirable.

PRIVACY AND SECURITY

When data on files can be read or changed by persons at terminals, it is necessary to prevent unauthorized persons from reading information that does not concern them and from modifying data or creating new records. The unauthorized reading of records can consitute an invasion of privacy on some systems—a situation that, understandably, has caused concern in the press and Congress. Unauthorized modification of records provides opportunities for crime or for tampering with the system.

Steps should be taken to prevent this form of access to a system. Improper access could originate from terminals, from the computer room, by wire tapping, by means of programs, or by entry into tape or disk stores. Access by wire-tapping or terminal misuse can be controlled by appropriate design of terminal procedures. Security might add 5 percent to the overall system cost, or more if very tight controls are needed. Some users have been willing to spend this money to prevent the possibility of embezzlement. Terminal access by a wide variety of people can offer temptations for tampering with the records. Yet, some users have been unwilling to spend money to secure privacy of personal information. This attitude may have to change as increasing personal information is stored in computer data banks.

It is important that the public, the press, and political authorities understand that computer data banks and data transmission can be made secure. Data can be locked up in computer systems just as it can in a bank vault, but the system becomes more expensive. In devising electronic locks and procedures, computer technology ultimately works on the side of security rather than on the side of the invader.

SUMMARY CHECKLIST

The following pages summarize the terminal features to consider when selecting a terminal.

Manual Input Facilities

—Typewriterlike keyboard
—Keyboard with letters in alphabetic sequence
—Keyboard like a Touchtone telephone
—Keyboard like a calculating machine

A Short Checklist of Terminal Planning Factors

—Matrix keyboard
—Keys with special labels
—Keyboard with interchangeable key labels
—Keyboard with overlays or templates
—Lever set
—Rotary switches
—Pushbuttons
—Light pen with display tube
—Coupled stylus on "desk pad"
—Pen-following mechanism
—Badge reader
—Punched-card reader
—Matrix card holder

- Can the data be keyed into a buffer and modified before transmission?
- Is paper tape or any other serial medium used for buffering?
- Does the keyboard give any help in formatting messages?
- Does it have the necessary keys for the dialogue in question?
- Does it have special facilities for when the computer fails?
- Can it be operated with one hand?
- Can the keys be changed?
- Does a bell ring at the end of a line?
- Are there good cursor controls?
- Are there facilities for easy modification of computer data?
- Are there skip and tab keys?
- Are there page or scroll keys?
- Are there YES/NO or other keys for high-speed scanning?
- Is the manual correction of errors easy?
- Can the numeric part of the keyboard be operated by one hand (3 x 4 matrix)?
- Does the keyboard have a good "feel" to a fast touch-typist?
- Are HELP or INTERRUPT keys desirable in the man-machine dialogue?
- Does the input means have appropriate security features?

Document Input Facilities

—Paper-tape reader/punch
—Card reader/punch
—Magnetic-tape cassette
—Disk
—Magnetic card reader

- Does the machine in the foregoing cases have or need the facility to write or punch a document as it is being keyed in?
- Can one storage media be shared by many keyboards?

—Badge and credit card (identity card) reader
—Optical document reader
—Magnetic-ink document reader
—Mark-sense card reader
—Matrix plate or card reader

- Can various devices be attached to one control unit?

Output Facilities

—Typewriterlike printer
—Printer that operates faster than a typewriter
—Inexpensive numeric-only printer, like a calculating machine

- Should the printer print on a special document, such as a bank passbook?
- Should the printer have special characters—for example, for mathematics, text editing, chemical formulas, or producing diagrams?
- Should the character set be interchangeable as with an IBM Selectric "golf-ball"?
- Is hard copy essential or would it be possible to do without it?
- Could the hard-copy facility be at the computer or concentrator rather than at the terminal?
- Could a camera be used rather than an expensive copying machine?
- If a printer or plotter is needed, could one such device serve many terminals?

Visual-Display Screen

—Picturephone
—Interface to standard television set
—Light panel
—Graph plotter
—Strip recorder

A Short Checklist of Terminal Planning Factors

—Telephone voice answerback
—Dials
—Facsimile machine
—Projector for slides, microfilm, microfiche, EVR frames, etc.

- How rapidly must the data be printed or displayed? This factor relates to the delivery time for bulk transmission and to man-machine interaction processes in a dialogue system?
- Is there a means of alerting the operator's attention?
- Is there an audible alarm?
- Can certain fields be highlighted—for example, with color?
- Does it have appropriate tabbing, skipping, and page-change features?

Features of Display Screens

- Can it display enough characters?
- Are the displayed characters large enough?
- Are the characters easy to read?
- Is it flicker-free?
- Is the image bright enough? Some displays have caused operator headaches.
- Is the image suitably protected from external glare?
- Is the display rate fast enough for the man-machine dialogue?
- Can it handle vectors or other graphic features?
- Destructive or nondestructive cursor?
- Can the cursor be made either destructive or nondestructive under program control?
- What cursor movements are possible?
- Are the keys for cursor movement straighforward?
- What character insert and delete capabilities are available?
- Does it have suitable special characters?
- Is the character set large enough?
- Should the character set be interchangeable: for example using microprogramming?
- Does it have a scroll feature (text roll up and roll down)?
- Does it have selectable horizontal tabs, reverse tabs, field skips, or other formatting features?
- What editing capabilities are available?
- Can individual fields be highlighted in some way—for example, by blinking, color, different brightness, or reverse field (either black characters on white or white characters on black)?
- Can masks be stored for editing of screen contents?
- Can data fields be protected (made unchangeable by the operator)?
- Can the data fields be program-defined?
- Are the features changeable—for example, microprogrammed?
- Are spaces to the right of an "end-of-line" character transmitted?
- What data compaction occurs on transmission?
- Does it have line addressing so that part of a display can be changed without the rest?
- Can a protected field be used for selective data entry?
- Are upper- and lower-case characters needed (e.g., in text editing)?
- Should images be displayed that are not composed digitally—for example, documents, signatures, photographs, diagrams? These may be stored at the terminal on film, microfilm, slides, EVR cartridges, etc.
- Can such images be half-tone (like a photograph)?
- Can such images be in color?
- Can "panels" be stored at the terminal control unit for display or for editing purposes?
- If locally stored images are displayed, should they be combined with transmitted data?
- Does it have a light pen?
- Is the detection of light-pen positioning fine enough?
- Does it have pen tracking capability and is it fast enough?

A Short Checklist of Terminal Planning Factors

- Can light-pen field selection be allowed and disallowed under program control?
- Do selectable fields brighten as the pen approaches to indicate to the operator that he may select them?
- Does the light pen have a pressure-sensitive switch in its tip?
- Can the images be printed?
- Can the images be automatically copied onto a monitor screen?
- Is a group display needed?
- Is an extra-high-capacity screen needed (10,000 characters or higher)?
- Can the screen image be printed?
- If so is the printer at the terminal, elsewhere at the terminal location serving several users, or at a concentrator or central computer location?
- What is the quality of the printing?

Features for Security

—Unique terminal identification by the computer

- Do dial-up terminals automatically transmit terminal ID?

—Lockable keyboard
—Nonprinting feature for when keying in security code or password.
—Automatic print/display suppression of security code or password field.

- Print/display suppression of other fields possible?
- Identification card reader?

—Cryptography feature
—Physical lock and key
—Feature for prevention of copying on other terminal or printer.
—Feature for erasing buffer.
—Feature to prevent terminal cable connections being switched.

- Protection from effects of control unit failure?

Features of the Communication Line Interface

- Is a standardized code used (e.g., ASCII)?
- Is a compact code needed to maximize transmission efficiency?
- Can different codes be used, for flexibility?
- Can any characters be used (e.g., with an escape character mechanism)?

Features for Control of Errors

- Erase, or backspace, key.
- Cancel transaction key.
- Automatic error detection. The code used for automatic error detection can range from relatively insecure parity checks to virtually errorproof polynomial codes of a high order.
- Automatic transmisssion when an error is detected (which implies some form of buffer at the terminal).
- Forward-error correction.
- Transaction logging in the terminal.
- Accumulators in the terminal for keeping totals.
- Logging facilities and/or accumulators that record details of transactions entered when the computer is inoperative.
- A recording mechanism (e.g., tape cassette) for recording transactions when the computer is inoperative, which can later be transmitted to it.
- This transmission may or may not be automatic.
- Synchronous or start-stop operation?
- Is full-duplex or half-duplex transmission used?
- If it is full-duplex, is it designed so that full advantage can be taken of simultaneous transmission in both directions?
- Is there any form of interrupt mechanism?
- Does the terminal have dial-up capabilities?
- Can it dial a remote machine automatically?
- Will it automatically redial if it first obtains a busy signal?
- Is a buffer used so that transmission can be at maximum-line speed (important, as we shall see on high-performance multidrop lines)?
- Can operator editing of input be done before transmission (especially with a video display)?
- Is the transmission rate suitably high?

A Short Checklist of Terminal Planning Factors

- Is the transmission rate changeable for bad line conditions?
- Is a modem or acoustical coupler built in?
- If not, does it have a standard EIA RS232B (or other) interface with the modem?
- Does it have multiplexing facilities—for example, a modulation device selecting a set portion of the available bandwidth?
- Can several terminals be connected to one buffer or control unit?
- Does the control unit dynamically assign buffer space?
- Can there be terminal-to-terminal communication?
- Does it contain logic for multidrop operation (many terminals on one line) such as polling?
- If polling is used, is it roll call or hub?
- If one device on a multidrop line fails, will this affect the line functioning?
- What communication line turnaround time is associated with the terminal?
- How does the turnaround time affect the network organization and response time?
- If it is normally attached to a leased line, can it have a alternate dial-up connection, possibly at lower speed?
- Will it automatically establish a dial-up connection if a leased line fails?
- Is a modem built into the terminal or separate?
- Is the terminal monitored for running out of paper or other holdups?
- Can unsolicited messages be sent to an idling terminal?
- Can an idling terminal be dialed?

General

- Is it portable?
- Need it be battery operated (e.g., when a terminal is taken in a car and used in a public call box)?
- Is it silent?
- Is it compact enough?
- Is it attractive?
- Is it robust?
- Is it reliable?
- Is it designed so that it can be easily serviced?
- Is the maintenance contract good enough?
- Is the maintenance organization good enough?
- Are there replaceable blocks of components for quick repair?
- Can these blocks be changed by trained local staff so that a visit from a repairman is unnecessary?
- Can the remote computer be used in terminal fault diagnosis or checkout?
- Is the construction modular so that separate keyboards or other devices can be used?
- Can the device be used off-line for a needed function—for example, typewriter or desk calculation.□

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Problem:

Every communications system design must begin with some kind of man-machine contact point, which is most frequently a keyboard-oriented terminal that can range from a simple, dumb teleprinter to an extremely intelligent minicomputer. The combination of very intelligent terminals and easy communications opens up a wholly different set of problems related to the evolving techniques of distributed processing. We treat the problems of planning for distributed processing in the Applications segment of this section. This report deals with the ordinary garden varieties of teleprinters and alphanumeric display terminals and offers some excellent guidelines for calculating how many terminals you will need to handle the existing traffic, loading, customers, etc., all of which you will have very accurately recorded after having evaluated your basic communications needs. Refer to Section CS15 if you need more information about the kinds of terminal hardware currently available.

Parts of this report will take you beyond ordinary mathematics and into the realm of calculus because some of the headier aspects of queuing theory cannot be explained adequately by any other method; but the excursions are brief and will not seriously interfere with your understanding of this report.

Solution:

Before doing the calculations necessary in planning the communication-line network, we need to decide how many terminals are required at each location. That is the subject of this report.

In some cases, we must begin one step further back and decide which locations need to have terminals. This decision may be clear-cut if the terminals are used for computing. Certain people need to use the terminal, and it must be available close to their locality, just as a desk calculator was in earlier days. On the other hand, where the terminal is not actually handled by the person requiring or giving information, he does not mind too much where it is located, providing that the service he receives is efficient and convenient. When you telephone an airline, for example, to make a reservation, the girl you speak to may be in your own town or in a distant city. You do not necessarily know where she is, and you do not care providing you have no trouble or expense in contacting her. Therefore, when designing such a system, we need not have a terminal in every town but can group them in certain strategic locations and run foreign exchange lines to the towns without terminals.

Similarly, in management information systems, a terminal is not available in every manager's office.

From *Systems Analysis for Data Transmission* by James Martin, Chpt. 32, pp 481-504, pp 840-42, 848-50, 851-53, and pp 857-60. © 1972 Prentice-Hall Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

How to Calculate the Number of Terminals You Will Need in Your Communications Network

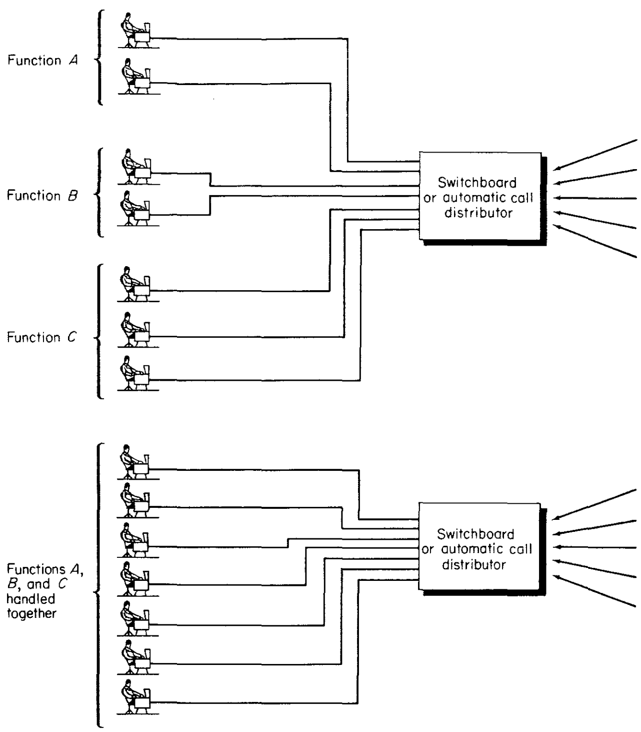


Figure 1. Longer waiting and more calls result from the top arrangement than from the bottom

Some managers use the telephone, and some, in a more expensive scheme, have a closed-circuit television link to an information center.

The grouping together of terminals and their operators lowers the total number needed. This cost saving must be balanced against the increased cost of telephone, television facsimile, and other to bring the service to its users and, in some systems, against the inconvenience of making the users come to a central location. As we shall see from the equations in this report, some arrangements of terminals and services are better than others and give a lower waiting time or better availability of services for the same cost.

QUEUING CALCULATIONS

In determining how many terminals and operators we need on a system, we must calculate which queues form for the use of the terminal and which form for the service of the operator. If the queue forms for one single operator or one single terminal, then single-server queuing theory can be used. If the persons queuing are free to choose between several operators or terminals, then we will use multiserver queuing theory.

If the traffic volume requires more than one server, then it is better to arrange for a free choice of servers rather than have a group of single-server queues. The arrangement in the bottom half of Figure 1 is better than that in the top half. In fact, the more servers we can have grouped together in one place to handle the traffic, the better. Suppose that the number of servers are chosen so that the facility utilization p = 0.8 and the mean service time is 10 minutes. Imagine a barber shop with 6 barbers. They take 10 minutes per customer on the average, and this time is exponentially distributed. The barbers are 80 percent utilized during the period of our observations. If a customer can select the first barber who becomes free, we have a six-server queue. The mean time the customer will have to wait before being served is 4.31 minutes.

$$\frac{E(t_q)}{E(t_s)}$$

from Table 1 is 1.431. Thus

$$E(t_q) = 1.431 \times 10 = 14.31$$

$$E(t_w) = E(t_q) - E(t_s) = 14.31 - 10 = 4.31 \text{ minutes}$$

If, on the other hand, each customer goes to his own particular barber, we have six single-server queues. The facility utilization p is the same, (0.8), but the mean time the customer waits before being served is 40 minutes. This is clearly no way to run a barber shop.

Suppose that 2 of the barbers can only do haircuts, 2 can only do rinses, and 2 can only do shaves. Suppose, also, for the sake of making the arithmetic easier, that the times for a shave, rinse, and haircut are the same and that one-third of the customers want shaves, one-third want rinses, and one-third want haircuts. We then have 3 two-server queues, with p = 0.8 again. The customers now have to wait, on the average, 17.78 minutes before their hair is cut. This is more than four times the six-server figure of 4.31 minutes. When the shop is busier and p = 0.9, the differences in the preceding times are much greater. The management would have given better customer service if it had trained all the barbers to do shaves, rinses, and haircuts.

The same applies to the use of terminals. When designing a system in which terminal operators handle a variety of functions for clients who telephone, it is tempting to have different operators handle different types of work. Better service could be given, however, if every terminal operator could handle all the functions that a telephone caller might need. This process would complicate the job of the operator and may necessitate additional training. It may be facilitated by designing the terminal language

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Mean waiting time before being served, $E(t_w) = E(t_q) - E(t_s)$

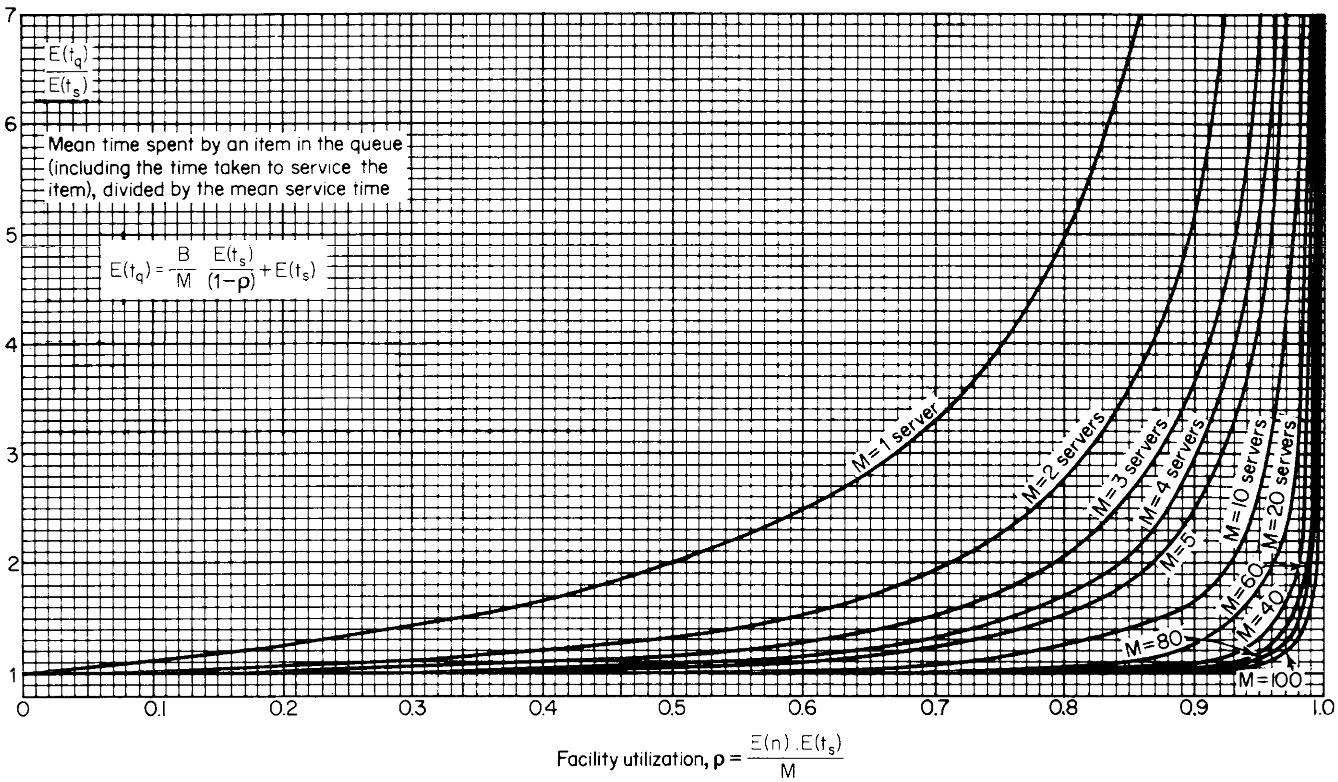


Table 1. Mean queuing times for multiserver queues, with exponential interarrival times, exponential service times, all servers equally loaded, and first-in, first-out dispatching

to give the maximum help to the operator. The systems analyst must calculate the benefits and determine what he considers to be the best trade-off.

One savings bank in the author's experience had teller windows labeled by alphabetic grouping A—G, H—M, and so on. The customers went to the appropriate window, depending on their name. Before automation, this practice was a good one because each teller had a more manageable file of customer records. Then a real-time system was installed and the tellers had terminals. Each terminal could access any customer's record. Nevertheless, the alphabetic grouping remained. The time the customers had to wait was substantially longer than if they had been able to go to any teller. The busier the bank, the larger were the values of p, and the worse was this difference in waiting time. After more than a year of operating this way, the bank finally removed the alphabetic grouping signs.

On some systems, the lines are a particularly sensitive issue because excessive waiting might result in lost business. This factor may be true for an airline. If you telephone an airline to make an enquiry or a reservation and are kept waiting beyond the limits of your patience, you may ring off and call a different airline. An airline would like as many customers as possible to have zero waiting time when they telephone.

Example 1. Probability That a Caller Will Be Kept Waiting. *In a small city, 20 customers per hour, on the average, call an airline during its period of peak activity to enquire about flights or make reservations. Real-time terminals are being installed, and once the terminals are installed, it will take an average of 6 minutes to deal with a call. This time is approximately exponentially distributed. Each agent handling the calls is equipped with a terminal. It is considered desirable that not more than one-tenth of the callers should be kept waiting during the peak period. How many terminals will be needed to meet this criterion? If a caller is kept waiting, how long will the wait be?*

The airline has both foreign and domestic flights. Of the 20 calls, 7 are for foreign flights and 13 are for domestic flights, on the average. Separate agents have always handled the foreign flights. What will be the effect of keeping them separate?

If a separate group of agents handles the foreign calls, the facility utilization (p) for these agents will be

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Values of $\frac{E(t_q)}{E(t_s)}$

Facility Utilization ρ	Number of Servers, M: $M=1$	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$	$M=7$	$M=8$	$M=9$	$M=10$	$M=11$
0.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.02	1.020	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.04	1.042	1.002	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.064	1.004	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.08	1.087	1.006	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.10	1.111	1.010	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	1.136	1.015	1.002	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.14	1.163	1.020	1.004	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.16	1.190	1.026	1.005	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.18	1.220	1.033	1.008	1.002	1.001	1.000	1.000	1.000	1.000	1.000	1.000
0.20	1.250	1.042	1.010	1.003	1.001	1.000	1.000	1.000	1.000	1.000	1.000
0.22	1.282	1.051	1.014	1.004	1.001	1.001	1.000	1.000	1.000	1.000	1.000
0.24	1.316	1.061	1.017	1.006	1.002	1.001	1.000	1.000	1.000	1.000	1.000
0.26	1.351	1.073	1.022	1.008	1.003	1.001	1.001	1.000	1.000	1.000	1.000
0.28	1.389	1.085	1.027	1.010	1.004	1.002	1.001	1.000	1.000	1.000	1.000
0.30	1.429	1.099	1.033	1.013	1.006	1.003	1.001	1.001	1.000	1.000	1.000
0.32	1.471	1.114	1.040	1.017	1.008	1.004	1.002	1.001	1.001	1.000	1.000
0.34	1.515	1.131	1.048	1.021	1.010	1.005	1.003	1.001	1.001	1.000	1.000
0.36	1.562	1.149	1.057	1.026	1.013	1.007	1.004	1.002	1.001	1.001	1.000
0.38	1.613	1.169	1.067	1.031	1.016	1.009	1.005	1.003	1.002	1.001	1.001
0.40	1.667	1.190	1.078	1.038	1.020	1.011	1.006	1.004	1.002	1.001	1.001
0.42	1.724	1.214	1.091	1.045	1.025	1.014	1.008	1.005	1.003	1.002	1.001
0.44	1.786	1.240	1.105	1.054	1.030	1.018	1.011	1.007	1.004	1.003	1.002
0.46	1.852	1.268	1.121	1.063	1.036	1.022	1.014	1.009	1.006	1.004	1.003
0.48	1.923	1.299	1.138	1.074	1.044	1.027	1.017	1.012	1.008	1.005	1.004
0.50	2.000	1.333	1.158	1.087	1.052	1.033	1.022	1.015	1.010	1.007	1.005
0.52	2.083	1.371	1.180	1.101	1.062	1.040	1.027	1.019	1.013	1.009	1.007
0.54	2.174	1.412	1.204	1.117	1.073	1.048	1.033	1.023	1.017	1.012	1.009
0.56	2.273	1.457	1.231	1.135	1.086	1.058	1.040	1.029	1.021	1.016	1.012
0.58	2.381	1.507	1.262	1.156	1.101	1.069	1.049	1.036	1.027	1.020	1.015
0.60	2.500	1.562	1.296	1.179	1.118	1.082	1.059	1.044	1.033	1.025	1.020
0.62	2.632	1.624	1.334	1.206	1.138	1.097	1.071	1.053	1.041	1.032	1.025
0.64	2.778	1.694	1.377	1.236	1.160	1.114	1.085	1.064	1.050	1.039	1.032
0.66	2.941	1.772	1.427	1.271	1.186	1.135	1.101	1.078	1.061	1.049	1.040
0.68	3.125	1.860	1.483	1.311	1.217	1.159	1.120	1.094	1.075	1.060	1.049
0.70	3.333	1.961	1.547	1.357	1.252	1.187	1.143	1.113	1.091	1.074	1.061
0.72	3.571	2.076	1.621	1.411	1.293	1.220	1.170	1.135	1.110	1.091	1.076
0.74	3.846	2.210	1.708	1.474	1.342	1.259	1.203	1.163	1.133	1.111	1.093
0.76	4.167	2.367	1.810	1.548	1.400	1.306	1.242	1.196	1.162	1.136	1.115
0.78	4.545	2.554	1.932	1.637	1.469	1.362	1.289	1.236	1.197	1.166	1.142
0.80	5.000	2.778	2.079	1.746	1.554	1.431	1.347	1.286	1.240	1.205	1.176
0.82	5.556	3.053	2.259	1.879	1.659	1.518	1.420	1.349	1.295	1.253	1.220
0.84	6.250	3.397	2.486	2.047	1.792	1.627	1.512	1.428	1.365	1.315	1.275
0.86	7.143	3.840	2.780	2.265	1.965	1.770	1.633	1.533	1.457	1.397	1.349
0.88	8.333	4.433	3.172	2.558	2.197	1.962	1.797	1.675	1.582	1.509	1.450
0.90	10.000	5.263	3.724	2.969	2.525	2.234	2.029	1.877	1.761	1.669	1.594
0.92	12.500	6.510	4.553	3.589	3.019	2.644	2.379	2.183	2.031	1.912	1.815
0.94	16.667	8.591	5.938	4.626	3.847	3.333	2.968	2.697	2.488	2.321	2.186
0.96	25.000	12.755	8.711	6.705	5.508	4.716	4.152	3.732	3.407	3.148	2.937
0.98	50.000	25.253	17.041	12.950	10.503	8.877	7.718	6.851	6.178	5.641	5.202

Table 1. (Continued)

$$\rho = \frac{E(n)E(t_s)}{M} = \frac{\frac{7}{60} \times 6}{M} = \frac{0.7}{M}$$

where M is the number of agents.

Let us see if 2 agents will be sufficient to handle the foreign calls. We want to find B, the probability of all agents being busy, which is tabulated in Table 2. B must be no greater than 0.1 if the customer-waiting criterion is to be satisfied.

If M = 2, then p = 0.35 and from Table 2 B = 0.18. The criterion is not met. Let us try 3 agents.

If M = 3, then p = 0.233 and from Table 2 B = 0.035. We thus need 3 agents for foreign calls.

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Values of $\frac{E(t_q)}{E(t_s)}$

Number of Servers, M

Facility Utilization ρ	$M=12$	$M=13$	$M=14$	$M=15$	$M=16$	$M=17$	$M=18$	$M=19$	$M=20$	$M=25$	$M=30$
0.40	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.42	1.001	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.44	1.001	1.001	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.46	1.002	1.001	1.001	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.48	1.003	1.002	1.001	1.001	1.001	1.001	1.000	1.000	1.000	1.000	1.000
0.50	1.004	1.003	1.002	1.002	1.001	1.001	1.001	1.000	1.000	1.000	1.000
0.52	1.005	1.004	1.003	1.002	1.002	1.001	1.001	1.001	1.001	1.000	1.000
0.54	1.007	1.005	1.004	1.003	1.002	1.002	1.001	1.001	1.001	1.000	1.000
0.56	1.009	1.007	1.005	1.004	1.003	1.003	1.003	1.002	1.001	1.000	1.000
0.58	1.012	1.009	1.007	1.006	1.005	1.004	1.003	1.003	1.002	1.001	1.000
0.60	1.016	1.012	1.010	1.008	1.007	1.005	1.004	1.004	1.003	1.001	1.001
0.62	1.020	1.016	1.013	1.011	1.009	1.007	1.006	1.005	1.004	1.002	1.001
0.64	1.026	1.021	1.017	1.014	1.012	1.010	1.008	1.007	1.006	1.003	1.001
0.66	1.032	1.027	1.022	1.019	1.016	1.013	1.011	1.010	1.008	1.004	1.002
0.68	1.041	1.034	1.029	1.024	1.021	1.018	1.015	1.013	1.012	1.006	1.003
0.70	1.051	1.043	1.037	1.031	1.027	1.023	1.020	1.018	1.016	1.008	1.005
0.72	1.064	1.054	1.047	1.040	1.035	1.031	1.027	1.024	1.021	1.012	1.007
0.74	1.079	1.068	1.059	1.051	1.045	1.040	1.035	1.031	1.028	1.016	1.010
0.76	1.099	1.086	1.075	1.066	1.058	1.051	1.046	1.041	1.037	1.023	1.015
0.78	1.123	1.107	1.094	1.084	1.074	1.066	1.060	1.054	1.049	1.031	1.021
0.80	1.154	1.135	1.119	1.106	1.095	1.086	1.077	1.070	1.064	1.042	1.029
0.82	1.193	1.170	1.152	1.136	1.122	1.111	1.101	1.092	1.084	1.057	1.040
0.84	1.243	1.216	1.194	1.175	1.158	1.144	1.132	1.121	1.111	1.077	1.056
0.86	1.310	1.277	1.250	1.227	1.206	1.189	1.174	1.160	1.148	1.105	1.078
0.88	1.402	1.362	1.327	1.298	1.273	1.251	1.232	1.215	1.200	1.145	1.110
0.90	1.533	1.482	1.439	1.402	1.370	1.342	1.317	1.295	1.275	1.203	1.157
0.92	1.734	1.667	1.610	1.561	1.518	1.481	1.448	1.419	1.392	1.295	1.232
0.94	2.074	1.980	1.900	1.831	1.771	1.718	1.671	1.630	1.593	1.453	1.363
0.96	2.761	2.614	2.488	2.379	2.284	2.200	2.126	2.061	2.001	1.779	1.632
0.98	4.838	4.529	4.266	4.038	3.839	3.663	3.507	3.368	3.243	2.770	2.457

ρ	$M=35$	$M=40$	$M=45$	$M=50$	$M=55$	$M=60$	$M=65$	$M=70$	$M=80$	$M=90$	$M=100$
0.64	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.66	1.001	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.68	1.002	1.001	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.70	1.003	1.002	1.001	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.72	1.004	1.003	1.002	1.001	1.001	1.001	1.000	1.000	1.000	1.000	1.000
0.74	1.007	1.004	1.003	1.002	1.002	1.001	1.001	1.001	1.000	1.000	1.000
0.76	1.010	1.007	1.005	1.003	1.003	1.002	1.001	1.001	1.001	1.000	1.000
0.78	1.014	1.010	1.007	1.006	1.004	1.003	1.002	1.002	1.001	1.001	1.000
0.80	1.021	1.015	1.011	1.009	1.007	1.005	1.004	1.003	1.002	1.001	1.001
0.82	1.029	1.022	1.017	1.013	1.011	1.009	1.007	1.006	1.004	1.003	1.002
0.84	1.042	1.032	1.026	1.021	1.017	1.014	1.011	1.009	1.007	1.005	1.004
0.86	1.060	1.047	1.038	1.031	1.026	1.022	1.018	1.016	1.012	1.009	1.007
0.88	1.086	1.069	1.057	1.047	1.040	1.034	1.029	1.025	1.020	1.015	1.012
0.90	1.126	1.103	1.086	1.073	1.062	1.054	1.047	1.042	1.033	1.026	1.022
0.92	1.189	1.157	1.133	1.115	1.100	1.088	1.078	1.069	1.056	1.046	1.039
0.94	1.299	1.253	1.217	1.189	1.167	1.149	1.133	1.120	1.100	1.084	1.072
0.96	1.529	1.452	1.394	1.347	1.310	1.279	1.253	1.230	1.195	1.168	1.147
0.98	2.234	2.069	1.940	1.838	1.755	1.686	1.628	1.578	1.498	1.437	1.388

Table 1. (Continued)

Now the domestic calls:

$$\rho = \frac{E(n)E(t_s)}{M} = \frac{\frac{13}{60} \times 6}{M} = \frac{1.3}{M}$$

Try M = 3: p = 0.433 and B = 0.17.
Try M = 4: p = 0.325 and B = 0.05.

We need 4 agents for domestic calls. Thus we have a total of 7 agents (and terminals).

The 5 percent of callers who are kept waiting will wait

pick up

$$E(t_d) = \frac{E(t_s)}{M(1-\rho)} = \frac{6}{4 \times (1 - 0.325)} = 2.22 \text{ minutes}$$

How to Calculate the Number of Terminals You Will Need in Your Communications Network

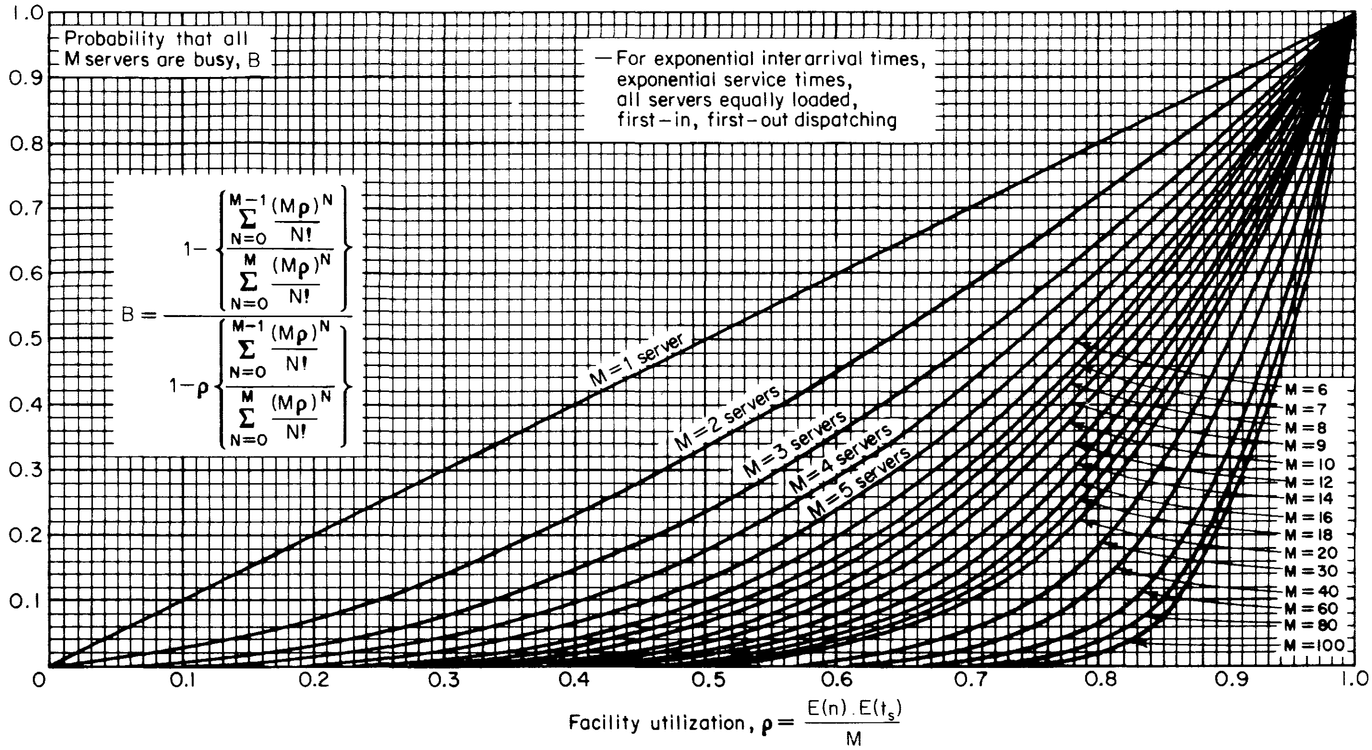


Table 2. Probability that all servers are busy in a multiserver queue, with exponential interarrival times, exponential service times, all servers equally loaded, and first-in, first-out dispatching

The callers for foreign flights will wait

$$E(t_d) = \frac{6}{3 \times (1 - 0.233)} = 2.61 \text{ minutes}$$

Now suppose that the same agents handle either foreign or domestic calls. For this group, we have

pick up $$\rho = \frac{\frac{20}{60} \times 6}{M} = \frac{2}{M}$$

Try M = 4: p = 0.5 and B = 0.174.
Try M = 5: p = 0.4 and B = 0.06.

With this arrangement, we need 5 agents instead of 7. This number, when all the cities with agents are considered, is probably a sufficient saving to justify training the agent to handle both foreign and domestic calls.

The callers who are kept waiting will now wait, on the average,

$$E(t_d) = \frac{6}{5 \times (1 - 0.4)} = 2 \text{ minutes}$$

somewhat less than in the foregoing case.

GROUPING CITIES TOGETHER

Suppose that there were five small cities not too far apart, each having this same traffic volume. If all their calls were routed to one point, there would be 100 calls per hour in the peak period. Thus

$$\rho = \frac{\frac{100}{60} \times 6}{M} = \frac{10}{M}$$

With 15 agents, we than have p = 0.666, and from Table 2 B = 0.1.

In this case, 15 agents would be sufficient instead of the preceding 25 (or 35 with foreign calls handled separately).

Customers kept waiting would have to wait only

$$\frac{6}{15 \times (1 - 0.666)} = 1.2 \text{ minutes}$$

on the average.

Here 20 such cities could be handled by 50 agents, and customers who had to wait would wait only 0.6 minute

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Facility Utilization ρ	Number of Servers, M: M=1	M=2	M=3	M=4	M=5	M=6	M=7	M=8	M=9	M=10	M=11
0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.02	0.020	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.04	0.040	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.06	0.060	0.007	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.08	0.080	0.012	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.10	0.100	0.018	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.12	0.120	0.026	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.14	0.140	0.034	0.009	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.16	0.160	0.044	0.014	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.18	0.180	0.055	0.019	0.007	0.002	0.001	0.000	0.000	0.000	0.000	0.000
0.20	0.200	0.067	0.025	0.010	0.004	0.002	0.001	0.000	0.000	0.000	0.000
0.22	0.220	0.079	0.032	0.013	0.006	0.003	0.001	0.001	0.000	0.000	0.000
0.24	0.240	0.093	0.040	0.018	0.008	0.004	0.002	0.001	0.000	0.000	0.000
0.26	0.260	0.107	0.049	0.023	0.011	0.006	0.003	0.001	0.001	0.000	0.000
0.28	0.280	0.122	0.059	0.030	0.015	0.008	0.004	0.002	0.001	0.001	0.000
0.30	0.300	0.138	0.070	0.037	0.020	0.011	0.006	0.004	0.002	0.001	0.001
0.32	0.320	0.155	0.082	0.046	0.026	0.015	0.009	0.005	0.003	0.002	0.001
0.34	0.340	0.173	0.095	0.055	0.033	0.020	0.012	0.007	0.005	0.003	0.002
0.36	0.360	0.191	0.110	0.066	0.040	0.025	0.016	0.010	0.007	0.004	0.003
0.38	0.380	0.209	0.125	0.078	0.049	0.032	0.021	0.014	0.009	0.006	0.004
0.40	0.400	0.229	0.141	0.091	0.060	0.040	0.027	0.018	0.013	0.009	0.006
0.42	0.420	0.248	0.158	0.105	0.071	0.049	0.034	0.024	0.017	0.012	0.009
0.44	0.440	0.269	0.177	0.120	0.084	0.059	0.043	0.031	0.022	0.016	0.012
0.46	0.460	0.290	0.196	0.137	0.098	0.071	0.052	0.039	0.029	0.022	0.016
0.48	0.480	0.311	0.216	0.155	0.114	0.084	0.064	0.048	0.037	0.028	0.022
0.50	0.500	0.333	0.237	0.174	0.130	0.099	0.076	0.059	0.046	0.036	0.028
0.52	0.520	0.356	0.259	0.194	0.149	0.115	0.090	0.072	0.057	0.046	0.037
0.54	0.540	0.379	0.281	0.216	0.168	0.133	0.106	0.086	0.069	0.057	0.046
0.56	0.560	0.402	0.305	0.238	0.190	0.153	0.124	0.102	0.084	0.069	0.058
0.58	0.580	0.426	0.330	0.262	0.212	0.174	0.144	0.120	0.100	0.084	0.071
0.60	0.600	0.450	0.355	0.287	0.236	0.197	0.165	0.140	0.119	0.101	0.087
0.62	0.620	0.475	0.381	0.313	0.262	0.221	0.188	0.161	0.139	0.120	0.105
0.64	0.640	0.500	0.408	0.340	0.289	0.247	0.213	0.185	0.162	0.142	0.125
0.66	0.660	0.525	0.435	0.369	0.317	0.275	0.241	0.212	0.187	0.166	0.148
0.68	0.680	0.550	0.463	0.398	0.347	0.305	0.270	0.240	0.215	0.193	0.173
0.70	0.700	0.576	0.492	0.429	0.378	0.336	0.301	0.271	0.245	0.222	0.202
0.72	0.720	0.603	0.522	0.460	0.410	0.369	0.334	0.303	0.277	0.254	0.233
0.74	0.740	0.629	0.552	0.493	0.444	0.404	0.369	0.339	0.312	0.288	0.267
0.76	0.760	0.656	0.583	0.526	0.480	0.440	0.406	0.376	0.349	0.326	0.304
0.78	0.780	0.684	0.615	0.561	0.516	0.478	0.445	0.416	0.390	0.366	0.345
0.80	0.800	0.711	0.647	0.596	0.554	0.518	0.486	0.458	0.432	0.409	0.388
0.82	0.820	0.738	0.680	0.633	0.593	0.559	0.529	0.502	0.478	0.455	0.435
0.84	0.840	0.767	0.713	0.670	0.634	0.602	0.574	0.548	0.525	0.504	0.485
0.86	0.860	0.795	0.747	0.709	0.675	0.646	0.621	0.597	0.576	0.556	0.538
0.88	0.880	0.824	0.782	0.748	0.718	0.693	0.669	0.648	0.629	0.611	0.594
0.90	0.900	0.853	0.817	0.788	0.762	0.740	0.720	0.702	0.687	0.669	0.654
0.92	0.920	0.882	0.853	0.829	0.808	0.789	0.772	0.757	0.743	0.729	0.717
0.94	0.940	0.911	0.889	0.870	0.854	0.840	0.827	0.815	0.803	0.793	0.783
0.96	0.960	0.940	0.925	0.913	0.902	0.892	0.883	0.874	0.866	0.859	0.852
0.98	0.980	0.970	0.962	0.956	0.950	0.945	0.940	0.936	0.932	0.928	0.924

Table 2. (Continued)

on the average. The saving in agents and terminals must be balanced against the cost of lines to route calls to the centralized location.

With terminals grouped together in this way, the system is also better protected from the effect of terminal failures or sick operators. In more than one way, there is safety in numbers.

PHYSICAL QUEUES OF PEOPLE

An automatic call distributor is likely to be used with the above system to route calls to the agents. If no agent is free when a customer calls, a recorded voice will ask him to wait until one becomes free. On some systems for other applications this function is done by a manual switchboard. Either way, it is a multiserver

How to Calculate the Number of Terminals You Will Need in Your Communications Network

ρ	M=12	M=13	M=14	M=15	M=16	M=17	M=18	M=19	M=20	M=25	M=30
0.30	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.32	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.34	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.36	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.38	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.40	0.004	0.003	0.002	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000
0.42	0.006	0.005	0.003	0.002	0.002	0.001	0.001	0.001	0.000	0.000	0.000
0.44	0.009	0.007	0.005	0.004	0.003	0.002	0.002	0.001	0.001	0.000	0.000
0.46	0.012	0.009	0.007	0.005	0.004	0.003	0.002	0.002	0.001	0.000	0.000
0.48	0.017	0.013	0.010	0.008	0.006	0.005	0.004	0.003	0.002	0.001	0.000
0.50	0.022	0.018	0.014	0.011	0.009	0.007	0.006	0.005	0.004	0.001	0.000
0.52	0.029	0.024	0.019	0.016	0.013	0.010	0.009	0.007	0.006	0.002	0.001
0.54	0.038	0.031	0.026	0.021	0.018	0.015	0.012	0.010	0.008	0.003	0.001
0.56	0.048	0.040	0.034	0.028	0.024	0.020	0.017	0.015	0.012	0.005	0.002
0.58	0.060	0.051	0.044	0.037	0.032	0.027	0.024	0.020	0.017	0.008	0.004
0.60	0.075	0.064	0.056	0.048	0.042	0.036	0.032	0.028	0.024	0.012	0.007
0.62	0.091	0.080	0.070	0.061	0.054	0.048	0.042	0.037	0.033	0.018	0.010
0.64	0.110	0.098	0.087	0.077	0.068	0.061	0.055	0.049	0.044	0.025	0.015
0.66	0.112	0.118	0.106	0.095	0.086	0.077	0.070	0.063	0.057	0.035	0.022
0.68	0.156	0.142	0.128	0.117	0.106	0.097	0.088	0.081	0.074	0.048	0.032
0.70	0.184	0.168	0.154	0.141	0.130	0.119	0.110	0.101	0.094	0.064	0.044
0.72	0.214	0.198	0.183	0.169	0.157	0.146	0.135	0.126	0.117	0.083	0.060
0.74	0.248	0.231	0.215	0.201	0.188	0.176	0.165	0.154	0.145	0.107	0.080
0.76	0.285	0.267	0.251	0.236	0.223	0.210	0.198	0.187	0.177	0.136	0.105
0.78	0.325	0.307	0.291	0.276	0.262	0.248	0.236	0.225	0.214	0.169	0.136
0.80	0.369	0.351	0.335	0.319	0.305	0.292	0.279	0.267	0.256	0.209	0.173
0.82	0.416	0.399	0.382	0.367	0.353	0.339	0.327	0.315	0.303	0.255	0.217
0.84	0.467	0.450	0.434	0.419	0.405	0.392	0.380	0.368	0.356	0.307	0.268
0.86	0.521	0.505	0.490	0.476	0.462	0.450	0.438	0.426	0.415	0.367	0.327
0.88	0.579	0.564	0.550	0.537	0.524	0.513	0.501	0.490	0.480	0.434	0.395
0.90	0.640	0.627	0.614	0.603	0.591	0.581	0.570	0.560	0.551	0.508	0.471
0.92	0.705	0.694	0.683	0.673	0.663	0.654	0.645	0.636	0.628	0.590	0.557
0.94	0.773	0.765	0.756	0.748	0.740	0.732	0.725	0.718	0.711	0.680	0.653
0.96	0.845	0.839	0.833	0.827	0.822	0.816	0.811	0.806	0.801	0.779	0.759
0.98	0.921	0.918	0.914	0.911	0.908	0.905	0.903	0.900	0.897	0.885	0.874

ρ	M=35	M=40	M=45	M=50	M=55	M=60	M=65	M=70	M=80	M=90	M=100
0.52	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.54	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.56	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.58	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.60	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.62	0.006	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.64	0.009	0.006	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000
0.66	0.014	0.009	0.006	0.004	0.002	0.002	0.001	0.001	0.000	0.000	0.000
0.68	0.021	0.014	0.010	0.007	0.005	0.003	0.002	0.002	0.001	0.000	0.000
0.70	0.031	0.022	0.015	0.011	0.008	0.006	0.004	0.003	0.002	0.001	0.000
0.72	0.044	0.032	0.024	0.018	0.013	0.010	0.008	0.006	0.003	0.002	0.001
0.74	0.061	0.046	0.036	0.028	0.021	0.017	0.013	0.010	0.006	0.004	0.003
0.76	0.083	0.065	0.052	0.042	0.034	0.027	0.022	0.018	0.012	0.008	0.005
0.78	0.110	0.090	0.074	0.061	0.051	0.042	0.035	0.029	0.021	0.015	0.011
0.80	0.144	0.121	0.102	0.087	0.074	0.063	0.054	0.047	0.035	0.026	0.020
0.82	0.186	0.160	0.139	0.121	0.106	0.093	0.081	0.072	0.056	0.044	0.035
0.84	0.235	0.208	0.184	0.164	0.147	0.131	0.118	0.106	0.087	0.071	0.059
0.86	0.293	0.265	0.240	0.218	0.199	0.182	0.167	0.153	0.130	0.110	0.094
0.88	0.361	0.332	0.307	0.284	0.264	0.246	0.229	0.214	0.187	0.165	0.146
0.90	0.440	0.412	0.386	0.364	0.343	0.325	0.307	0.291	0.263	0.238	0.217
0.92	0.529	0.503	0.479	0.458	0.438	0.420	0.404	0.388	0.359	0.334	0.312
0.94	0.629	0.607	0.587	0.568	0.551	0.535	0.519	0.505	0.479	0.455	0.434
0.96	0.740	0.724	0.709	0.694	0.681	0.669	0.657	0.645	0.624	0.605	0.587
0.98	0.864	0.855	0.846	0.838	0.831	0.823	0.816	0.810	0.797	0.786	0.775

Table 2. (Continued)

How to Calculate the Number of Terminals You Will Need in Your Communications Network

queuing situation in which the calls are routed to the first free agent. Where physical queues of people form in front of counters, the situation is not quite so clear. If customers are free to select any of a group of servers, all doing identical jobs, as in a large bank, they will normally form separate queues for each server, but will switch queues if a nearby queue becomes shorter than their own, or if a nearby server becomes available. A new customer will freely select the shortest queue. It therefore seems reasonable to expect that the queue sizes will approximate those given by multiserver queuing theory, and calculations based on this assumption are used in practice for calculating the number of servers needed.

Example 2. Number of Terminal Operators Needed. *A real-time system is being designed for a savings bank. In a big city branch there have been large queues of customers at lunch time, and bank management fears that it is losing customers because of this. It has set a criterion on the design of the new system that the mean queue size for each teller should not exceed 2 (including the customer being served).*

The design team has made studies of customer volumes and has estimated that 20 customers in a 5-minute period can be expected during the lunch time peak. Studies of the tellers' work have indicated that when they each have a terminal their mean time to handle one customer will be 2 minutes, and that the distribution of this time will be approximately exponential.

Given this information, how many tellers (and terminals) will the bank require?

$$E(n) = \frac{20}{5} = 4 \text{ customers per minute.}$$

$$E(t_s) = 2 \text{ minutes}$$

$$\rho = \frac{E(n)E(t_s)}{M}$$

where M is the number of tellers. Therefore

$$\rho = \frac{8}{M}$$

Whe M = 8, p = 1, which is too high

Let us try M = 9. Using Table 1 this gives approximately:

$$\frac{E(t_q)}{E(t_s)} = 1.7$$

$$\therefore E(t_q) = 1.7 \times 2 = 3.4$$

$$E(q) = E(n) \times E(t_q)$$

$$\therefore E(q) = 4 \times 3.4 = 13.6$$

Therefore the mean number of customers queuing per teller is

$$\frac{13.6}{9} = 1.5$$

The design criterion is therefore satisfied by using 9 teller windows.

The time each teller takes to operate the terminal is estimated to be as follows:

Insert pass-book		5 sec
Key in transaction		4 sec
Response time		2 sec
Printing		2 sec
Remove pass-book		4 sec
	Total:	17 sec

This time was included in the estimate of 2 minutes as the mean time to handle each customer. It was suggested that one terminal should be shared by 3 tellers. If it was shared by 4 or more, then some of them would have to walk a distance to the terminal. If the teller is working non-stop, handling one customer every 2 minutes, then the utilization of the terminal will be

$$\rho = E(n) \times E(t_s) = \frac{1}{2 \times 60} \times 3 \times 17 = .425$$

The time taken to use the terminal is going to be nearly constant. The standard deviation is unlikely to be higher than, say, 5 seconds. Using the $(\sigma_{t_s}/E(t_s))^2 = 0.1$ column in Table 3 we have for the terminal (single-server):

$$\frac{E(t_q)}{E(t_s)} = 1.4$$

$$\therefore \frac{E(t_w)}{E(t_s)} = 0.4$$

The mean time the teller waits for the terminal is

0.4 X 17 = 6.8 seconds

In practice, the time will probably be less because the assumptions behind Table 3 are worse than the conditions met in this situation. There is not an infinite population of users, and the arrival rate may be smoother than a Poisson distribution.

Using this calculation, it was not felt necessary to revise upward the estimate of 2 minutes to handle one customer. The bank therefore needs 9 teller windows with 3 terminals. It was felt advisable to install a fourth terminal in case one is out of action during the peak period.

How to Calculate the Number of Terminals You Will Need in Your Communications Network

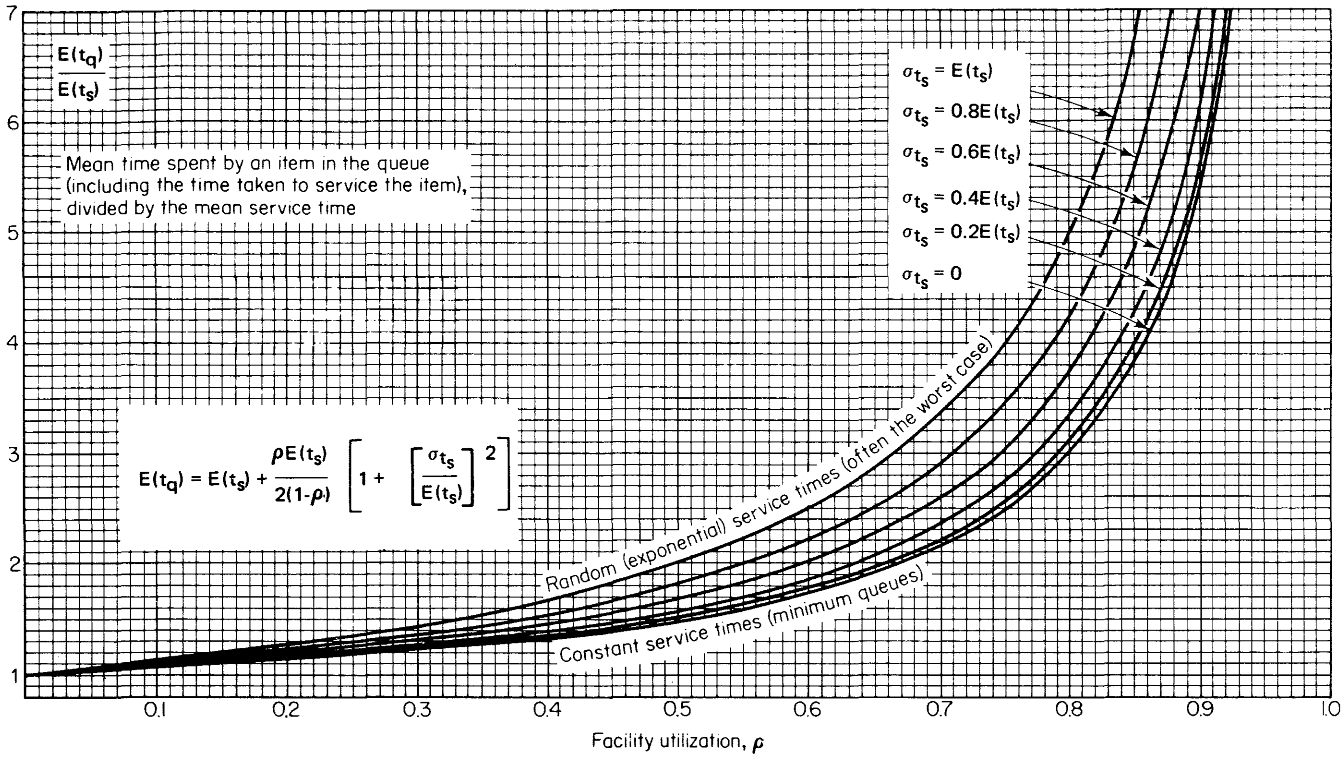


Table 3. Mean queuing time for single-server queues with poisson arrival pattern.

How would this situation work out in practice? Is there any practical experience that justifies the queuing model we have used? Interestingly enough, there is—and it suggests that our calculation here is not quite adequate. We shall discuss this point later in the report and do the preceding calculation again in the light of experience.

NUMBER OF TERMINALS PER OPERATOR

In savings banks, having one terminal between three tellers has worked well in practice. Some banks have no more than two tellers per terminal. In other applications, however, the "conversation" between the operator and the terminal is more complex. It may last for a longer period of time than the time the operator spends in talking to the client. There may be no client, and the operator may need to have the terminal available all the time. The counter of an airline office looks, perhaps, not dissimilar to the counter of a bank and the lines of people are similar, but the airline is likely to have one terminal per agent because the agent carries on a lengthier conversation with the machine.

The correspondence between operators and terminals must clearly be an early decision in the calculation of numbers of terminals needed and must be based on a study of the overall job of the person using the terminal.

GROWTH IN TRAFFIC VOLUMES

If the waiting time for this single-server queue was equal to the service time—that is,

$$\frac{E(t_q)}{E(t_s)} = 2$$

this situation should not be worrisome. An increase in traffic volume of a few percent would cause an increase in waiting time of a few percent—nothing catastrophic.

On the other hand, if we had a 10-server queue and the waiting time was equal to the service time

$$\frac{E(t_q)}{E(t_s)} = 2$$

this situation would be highly dangerous. An increase in traffic volume of a few percent would send the waiting time hurtling upward.

It is important on any data transmission system or real-time computer to make measurements of the traffic volumes at the terminals, on the lines, in the processing unit, in the computer channels, and so on and to relate these measurements to the shape of the queuing curves.

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Values of $\frac{E(t_q)}{E(t_s)}$

Facility Utilization ρ	Coefficient of Variation, Squared, for Service Time, $\left[\frac{\sigma_{t_s}}{E(t_s)}\right]^2$: 0.0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.02	1.010	1.011	1.011	1.012	1.012	1.013	1.013	1.014	1.014	1.015	1.015
0.04	1.021	1.022	1.023	1.024	1.025	1.026	1.027	1.028	1.029	1.030	1.031
0.06	1.032	1.034	1.035	1.037	1.038	1.040	1.041	1.043	1.045	1.046	1.048
0.08	1.043	1.046	1.048	1.050	1.052	1.054	1.057	1.059	1.061	1.063	1.065
0.10	1.056	1.058	1.061	1.064	1.067	1.069	1.072	1.075	1.078	1.081	1.083
0.12	1.068	1.072	1.075	1.078	1.082	1.085	1.089	1.092	1.095	1.099	1.102
0.14	1.081	1.085	1.090	1.094	1.098	1.102	1.106	1.110	1.114	1.118	1.122
0.16	1.095	1.100	1.105	1.110	1.114	1.119	1.124	1.129	1.133	1.138	1.143
0.18	1.110	1.115	1.121	1.126	1.132	1.137	1.143	1.148	1.154	1.159	1.165
0.20	1.125	1.131	1.137	1.144	1.150	1.156	1.162	1.169	1.175	1.181	1.187
0.22	1.141	1.148	1.155	1.162	1.169	1.176	1.183	1.190	1.197	1.204	1.212
0.24	1.158	1.166	1.174	1.182	1.189	1.197	1.205	1.212	1.221	1.229	1.237
0.26	1.176	1.184	1.193	1.202	1.211	1.220	1.228	1.237	1.246	1.255	1.264
0.28	1.194	1.204	1.214	1.224	1.233	1.243	1.253	1.262	1.272	1.282	1.292
0.30	1.214	1.225	1.236	1.246	1.257	1.268	1.279	1.289	1.300	1.311	1.321
0.32	1.235	1.247	1.259	1.271	1.282	1.294	1.306	1.318	1.329	1.341	1.353
0.34	1.258	1.270	1.283	1.296	1.309	1.322	1.335	1.348	1.361	1.373	1.386
0.36	1.281	1.295	1.309	1.323	1.337	1.352	1.366	1.380	1.394	1.408	1.422
0.38	1.306	1.322	1.337	1.352	1.368	1.383	1.398	1.414	1.429	1.444	1.460
0.40	1.333	1.350	1.367	1.383	1.400	1.417	1.433	1.450	1.467	1.483	1.500
0.42	1.362	1.380	1.398	1.416	1.434	1.453	1.471	1.489	1.507	1.525	1.543
0.44	1.393	1.412	1.432	1.452	1.471	1.491	1.511	1.530	1.550	1.570	1.589
0.46	1.426	1.447	1.469	1.490	1.511	1.532	1.554	1.575	1.596	1.618	1.639
0.48	1.462	1.485	1.508	1.531	1.554	1.577	1.600	1.623	1.646	1.669	1.692
0.50	1.500	1.525	1.550	1.575	1.600	1.625	1.650	1.675	1.700	1.725	1.750
0.52	1.542	1.569	1.596	1.623	1.650	1.677	1.704	1.731	1.758	1.785	1.812
0.54	1.587	1.616	1.646	1.675	1.704	1.734	1.763	1.792	1.822	1.851	1.880
0.56	1.636	1.668	1.700	1.732	1.764	1.795	1.827	1.859	1.891	1.923	1.955
0.58	1.690	1.725	1.760	1.794	1.829	1.863	1.898	1.932	1.967	2.001	2.036
0.60	1.750	1.787	1.825	1.862	1.900	1.937	1.975	2.012	2.050	2.087	2.125
0.62	1.816	1.857	1.897	1.938	1.979	2.020	2.061	2.101	2.142	2.183	2.224
0.64	1.889	1.933	1.978	2.022	2.067	2.111	2.156	2.200	2.244	2.289	2.333
0.66	1.971	2.019	2.068	2.116	2.165	2.213	2.262	2.310	2.359	2.407	2.456
0.68	2.062	2.116	2.169	2.222	2.275	2.328	2.381	2.434	2.487	2.541	2.594
0.70	2.167	2.225	2.283	2.342	2.400	2.458	2.517	2.575	2.633	2.692	2.750
0.72	2.286	2.350	2.414	2.479	2.543	2.607	2.671	2.736	2.800	2.864	2.929
0.74	2.423	2.494	2.565	2.637	2.708	2.779	2.850	2.921	2.992	3.063	3.135
0.76	2.583	2.662	2.742	2.821	2.900	2.979	3.058	3.137	3.217	3.296	3.375
0.78	2.773	2.861	2.950	3.039	3.127	3.216	3.305	3.393	3.482	3.570	3.659
0.80	3.000	3.100	3.200	3.300	3.400	3.500	3.600	3.700	3.800	3.900	4.000
0.82	3.278	3.392	3.506	3.619	3.733	3.847	3.961	4.075	4.189	4.303	4.417
0.84	3.625	3.756	3.887	4.019	4.150	4.281	4.412	4.544	4.675	4.806	4.937
0.86	4.071	4.225	4.379	4.532	4.686	4.839	4.993	5.146	5.300	5.454	5.607
0.88	4.667	4.850	5.033	5.217	5.400	5.583	5.677	5.950	6.133	6.317	6.500
0.90	5.500	5.725	5.950	6.175	6.400	6.625	6.850	7.075	7.300	7.525	7.750
0.92	6.750	7.037	7.325	7.612	7.900	8.187	8.475	8.762	9.050	9.337	9.625
0.94	8.833	9.225	9.617	10.008	10.400	10.792	11.183	11.575	11.967	12.358	12.750
0.96	13.000	13.600	14.200	14.800	15.400	16.000	16.600	17.200	17.800	18.400	19.000
0.98	25.500	26.725	27.950	29.175	30.400	31.625	32.850	34.075	35.300	36.525	37.750

Table 3. (Continued)

Example 3. The Effect of Traffic Increase. *A public utility has many enquiries from its customers and staff—querying bills and customer accounts. These questions are answered by a group of five operators using terminals with screens. The system has recently been installed and is regarded as working successfully. During a peak period, an average of 174 calls per hour are handled. It takes 1.5 minutes, on the average, to deal with one enquiry. During this peak time, a caller is kept waiting an average 1.4 minutes before talking to a terminal operator, but it is not thought worthwhile to spend money to reduce this waiting period.*

The company anticipates that the number of enquiries will be 10 percent higher during the winter

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Values of $\frac{E(t_q)}{E(t_s)}$

Facility Utiliza-tion ρ	Coefficient of Variation, Squared, for Service Time, $\left[\frac{\sigma_{t_s}}{E(t_s)}\right]^2$ 0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
0.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.02	1.016	1.016	1.017	1.017	1.018	1.108	1.019	1.019	1.020	1.020
0.04	1.032	1.033	1.034	1.035	1.036	1.037	1.039	1.040	1.041	1.042
0.06	1.049	1.051	1.053	1.054	1.056	1.057	1.059	1.061	1.062	1.064
0.08	1.067	1.070	1.072	1.074	1.076	1.078	1.080	1.083	1.085	1.087
0.10	1.086	1.089	1.092	1.094	1.097	1.100	1.103	1.106	1.108	1.111
0.12	1.106	1.109	1.112	1.116	1.119	1.123	1.126	1.130	1.133	1.136
0.14	1.126	1.130	1.134	1.138	1.142	1.147	1.151	1.155	1.159	1.163
0.16	1.148	1.152	1.157	1.162	1.167	1.171	1.176	1.181	1.186	1.190
0.18	1.170	1.176	1.181	1.187	1.192	1.198	1.203	1.209	1.214	1.220
0.20	1.194	1.200	1.206	1.212	1.219	1.225	1.231	1.237	1.244	1.250
0.22	1.219	1.226	1.233	1.240	1.247	1.254	1.261	1.268	1.275	1.282
0.24	1.245	1.253	1.261	1.268	1.276	1.284	1.292	1.300	1.308	1.316
0.26	1.272	1.281	1.290	1.307	1.307	1.316	1.325	1.334	1.343	1.351
0.28	1.301	1.311	1.321	1.331	1.340	1.350	1.360	1.369	1.379	1.389
0.30	1.332	1.343	1.354	1.364	1.375	1.386	1.396	1.407	1.418	1.429
0.32	1.365	1.376	1.388	1.400	1.412	1.424	1.435	1.447	1.459	1.471
0.34	1.399	1.412	1.425	1.438	1.451	1.464	1.477	1.489	1.502	1.515
0.36	1.436	1.450	1.464	1.478	1.492	1.506	1.520	1.534	1.548	1.562
0.38	1.475	1.490	1.506	1.521	1.536	1.552	1.567	1.582	1.598	1.613
0.40	1.517	1.533	1.550	1.567	1.583	1.600	1.617	1.633	1.650	1.667
0.42	1.561	1.579	1.597	1.616	1.634	1.652	1.670	1.688	1.706	1.724
0.33	1.609	1.629	1.648	1.668	1.687	1.707	1.727	1.746	1.766	1.786
0.46	1.660	1.681	1.703	1.724	1.745	1.767	1.788	1.809	1.831	1.852
0.48	1.715	1.738	1.762	1.785	1.808	1.831	1.854	1.877	1.900	1.923
0.50	1.775	1.800	1.825	1.850	1.875	1.900	1.925	1.950	1.975	2.000
0.52	1.840	1.876	1.894	1.921	1.948	1.975	2.002	2.029	2.056	2.083
0.54	1.910	1.939	1.968	1.998	2.027	2.057	2.086	2.115	2.145	2.174
0.56	1.986	2.018	2.050	2.082	2.114	2.145	2.177	2.209	2.241	2.273
0.58	2.070	2.105	2.139	2.174	2.208	2.243	2.277	2.312	2.346	2.381
0.60	2.162	2.200	2.237	2.275	2.312	2.350	2.387	2.425	2.462	2.500
0.62	2.264	2.305	2.346	2.387	2.428	2.468	2.509	2.550	2.591	2.632
0.64	2.378	2.422	2.467	2.511	2.556	2.600	2.644	2.689	2.733	2.778
0.66	2.504	2.553	2.601	2.650	2.699	2.747	2.796	2.844	2.893	2.941
0.68	2.647	2.700	2.753	2.806	2.859	2.912	2.966	3.019	3.072	3.125
0.70	2.808	2.867	2.925	2.983	3.042	3.100	3.158	3.217	3.275	3.333
0.72	2.993	3.057	3.121	3.186	3.250	3.314	3.379	3.443	3.507	3.571
0.74	3.206	3.277	3.348	3.419	3.490	3.562	3.633	3.704	3.775	3.846
0.76	3.454	3.533	3.612	3.692	3.771	3.850	3.929	4.008	4.087	4.167
0.78	3.748	3.836	3.925	4.014	4.102	4.191	4.280	4.368	4.457	4.545
0.80	4.100	4.200	4.300	4.400	4.500	4.600	4.700	4.800	4.900	5.000
0.82	4.531	4.644	4.758	4.872	4.986	5.100	5.214	5.328	5.442	5.556
0.84	5.069	5.200	5.331	5.462	5.594	5.725	5.856	5.987	6.119	6.250
0.86	5.761	5.914	6.068	6.221	6.375	6.529	6.682	6.836	6.989	7.143
0.88	6.683	6.867	7.050	7.233	7.417	7.600	7.783	7.967	8.150	8.333
0.90	7.975	8.200	8.425	8.650	8.875	9.100	9.325	9.550	9.775	10.000
0.92	9.912	10.200	10.487	10.775	11.062	11.350	11.637	11.925	12.212	12.500
0.94	13.142	13.533	13.925	14.317	14.708	15.100	15.492	15.883	16.275	16.667
0.96	19.600	20.200	20.800	21.400	22.000	22.600	23.200	23.800	24.400	25.000
0.98	38.975	40.200	41.425	42.650	43.875	45.100	46.325	47.550	48.775	50.000

Table 3. (Continued)

period. Will the present number of operators be able to handle that growth?

$$\rho = \frac{E(n)E(t_s)}{M} = \frac{\frac{174}{60} \times 1.5}{5} = 0.87$$

Using Table 1, we would expect that

$$\frac{E(t_q)}{E(t_s)} = 2.1$$

Thus $E(t_q) = 2.1 \times 1.5 = 3.15$ minutes

Therefore $E(t_w) = E(t_q) - E(t_s) = 3.15 - 1.5 = 1.65$ minutes

Multiserver queuing theory gives a mean waiting time of 1.65 minutes. The measured waiting time in the peak hour is 1.4 minutes. This waiting time may be due to the fact that the service times are not exponentially distributed, or it may be due to random

How to Calculate the Number of Terminals You Will Need in Your Communications Network

fluctuations of the values measured. In any case, the theory is reasonably close to the measured result.

When the number of enquiries increases by 10 percent, we will have

p = 0.87 + 0.087 = 0.957

This situation is far from healthy.

Table 1 gives

$$\frac{E(t_q)}{E(t_s)} = 5.5$$

Hence

$$E(t_q) = 8.25 \text{ minutes}$$

which would cause considerable impatience and annoyance to most customers. The data processing manager would be well advised to install at least one more terminal quickly.

With 6 terminals and a 10 percent increase in usage,

$$\rho = \frac{\frac{191}{60} \times 1.5}{6} = 0.80$$

drops to 2.15 minutes, which gives a mean waiting time of 0.65 minute.

If, however, the winter peak is 20 percent higher than now (rather than 10 percent), p would once again be uncomfortably high. It would be wise to have 7 terminals and operators in readiness, in addition to the spare one needed to cover illness or failure.

NONEXPONENTIAL SERVICE TIMES

If measurements of the service time in the foregoing calculations had indicated that it was far from exponential, then the approximation given in the following equations 1 and 2 might have been used in calculating the queue sizes and waiting times:

$$E(w) = \frac{E(w_{\text{exp}})}{2}\left\{1 + \left[\frac{\sigma_{t_s}}{E(t_s)}\right]^2\right\} \qquad \text{eq. 1}$$

and

$$E(t_w) = \frac{E(t_{w\text{exp}})}{2}\left\{1 + \left[\frac{\sigma_{t_s}}{E(t_s)}\right]^2\right\} \qquad \text{eq. 2}$$

where $E(w_{\text{exp}})$ and $E(t_{w\text{exp}})$ are the number of items waiting and the waiting time for the case in question if its service time distribution were exponential.

In the preceding example we were told that the mean time to handle one call was 1.5 minutes, and we assumed that the time was exponentially distributed. We might have been given the additional information that the standard deviation of the time to handle a call was 0.75 minute. We would then have

$$\left[\frac{\sigma_{t_s}}{E(t_s)}\right]^2 = \left(\frac{0.75}{1.5}\right)^2 = \frac{1}{4}$$

Using this assumption,

$$E(t_w) = \frac{E(t_{w\text{exp}})}{2}\left[1 + \frac{1}{4}\right] = \frac{5}{8}E(t_{w\text{exp}}).$$

For p = 0.87, we found from Table 1 that

$E(t_{w\text{exp}})$ = 1.65 minutes

Therefore $E(t_w) = \frac{5}{8} \times 1.65 = 1.03$ minutes

When the number of inquiries rises by 10 percent, however, p still rises to 0.957 and $E(t_w)$ rises to 4.22 minutes. The system is still on a highly unstable part of the curve and the same conclusions apply. Even if the service times were constant as in the lowest curve in Figure 5, a utilization p of 0.957 would be highly undesirable.

The data processing manager would certainly not be justified in feeling content with his system because the customers only wait 1.03 minutes on the average.

NONQUEUING SITUATION

The preceding models have related to situations in which the persons being served queue for the terminal or its operator. In other types of systems, if a terminal is not free the users go away. No queue forms. Laboratory personnel, for example, may go to a terminal room to use the terminal of a time-sharing system. If no terminal is free, they return to other work and come back later.

In this situation we are interested in calculating the probability that a user will not find a free terminal.

The probability that no terminal is free is

$$P_B = \frac{(M\rho)^M}{M!} \bigg/ \sum_{N=0}^{M} \frac{(M\rho)^N}{N!}$$

where M is the number of terminals and p is the terminal utilization.

$$M\rho = E(n)E(t_s)$$

where n is the number of persons arriving to use the terminal in unit time and t_s is the time the user occupies the terminal.

How to Calculate the Number of Terminals You Will Need in Your Communications Network

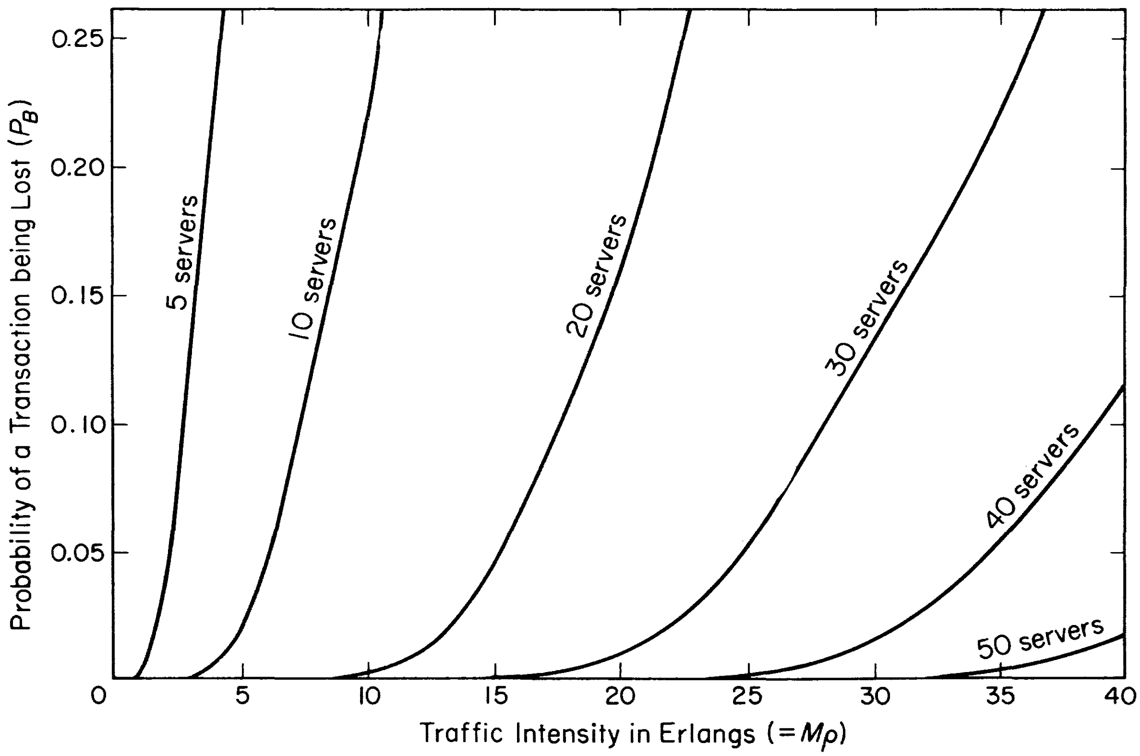


Table 4. Probability that a transaction will be lost in a multiserver, nonqueuing situation. Assumptions: poisson arrival pattern; all servers are equal; if an item cannot be served immediately, it is lost

The equation assumes that n follows a Poisson distribution and t_s an exponential distribution.

Example 4. Availibility of Terminals on a Time-Sharing System. *A laboratory has used an experimental time-sharing system for 2 years. The service has proven itself to be of great potential, yet the staff has continually felt frustrated by not being able to have access to a terminal at the moment one is needed. They complain that much of their time has been wasted, and consequently many have resorted to batch processing. A new system is now to be installed. A design criterion is that the probability of a staff member being unable to obtain a free terminal when he needs it should be no greater than 0.01.*

The duration of terminal occupancy is found to be approximately exponentially distributed with a mean of 40 minutes. It is estimated that with the new service 30 users will sign on per hour during the morning hours when activity has been observed to be highest. How many terminals are needed?

$$M\rho = E(n) \cdot E(t_s) = \frac{30}{60} \times 40 = 20$$

Using Table 4, when Mp = 20, thirty terminals will give P_B = 0.00846. Twenty-nine terminals will give P_B = 0.01279. Thirty terminals are therefore needed to fill the design criterion.

This assumes that any of the 30 terminals will be available to a new user. If some of the terminals are in persons' offices or in separate areas, this will not be so. Just as with the earlier calculations, if all the terminals are grouped together, it lowers the total number of terminals needed (or alternatively raises the standard of service given). Making users walk to the other end of the building, or go to a different floor, may provoke complaints ("What are telecommunications for?"). However, it may result in better service because after their walk they have a higher probability of finding a terminal free than if the terminals were dispersed throughout the building.

In my own place of work, the secretaries use terminals for producing documents, text editing and so on. A variety of scientific computing systems are used by different people. Some use terminals attached to a computer in the building and others dial distant machines. However, most of these activities make use of the same type of terminal (although there are also a small number of other more specialized terminals). In many types of organization it is an advantage to employ a common type of terminal, and to pool the terminals so that secretaries and scientists alike use the same group of machines (a popular arrangement).

Figure 2 shows the numbers of servers needed for different traffic rates, in order to achieve a probability-of-not-obtaining-a-server of 0.1, 0.01 and

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Traffic Rate in Erlangs (=Mρ)	Number of Servers, M: M=1	2	3	4	5	6	7	8	9	10
0.10	.09091	.00452	.00015	.00000	.00000	.00000	.00000	.00000	.00000	.00000
0.20	.16667	.01639	.00109	.00005	.00000	.00000	.00000	.00000	.00000	.00000
0.30	.23077	.03346	.00333	.00025	.00002	.00000	.00000	.00000	.00000	.00000
0.40	.28571	.05405	.00716	.00072	.00006	.00000	.00000	.00000	.00000	.00000
0.50	.33333	.07692	.01266	.00158	.00016	.00001	.00000	.00000	.00000	.00000
0.60	.37500	.10112	.01982	.00296	.00036	.00004	.00000	.00000	.00000	.00000
0.70	.41176	.12596	.02855	.00497	.00070	.00008	.00001	.00000	.00000	.00000
0.80	.44444	.15094	.03869	.00768	.00123	.00016	.00002	.00000	.00000	.00000
0.90	.47368	.17570	.05007	.01114	.00200	.00030	.00004	.00000	.00000	.00000
1.00	.50000	.20000	.06250	.01538	.00307	.00051	.00007	.00001	.00000	.00000
1.10	.52381	.22366	.07579	.02042	.00447	.00082	.00013	.00002	.00000	.00000
1.20	.54545	.24658	.08978	.02623	.00625	.00125	.00021	.00003	.00000	.00000
1.30	.56522	.26868	.10429	.03278	.00845	.00183	.00034	.00006	.00001	.00000
1.40	.58333	.28994	.11918	.04004	.01109	.00258	.00052	.00009	.00001	.00000
1.50	.60000	.31034	.13433	.04796	.01418	.00353	.00076	.00014	.00002	.00000
1.60	.61538	.32990	.14962	.05647	.01775	.00471	.00108	.00022	.00004	.00001
1.70	.62963	.34861	.16496	.06551	.02179	.00614	.00149	.00032	.00006	.00001
1.80	.64286	.36652	.18027	.07503	.02630	.00783	.00201	.00045	.00009	.00002
1.90	.65517	.38363	.19547	.08496	.03128	.00981	.00265	.00063	.00013	.00003
2.00	.66667	.40000	.21053	.09524	.03670	.01208	.00344	.00086	.00019	.00004
2.20	.68750	.43060	.23999	.11660	.04880	.01758	.00549	.00151	.00037	.00008
2.40	.70588	.45860	.26841	.13871	.06242	.02436	.00828	.00248	.00066	.00016
2.60	.72222	.48424	.29561	.16118	.07733	.03242	.01190	.00385	.00111	.00029
2.80	.73684	.50777	.32154	.18372	.09329	.04172	.01641	.00571	.00177	.00050
3.00	.75000	.52941	.34615	.20611	.11005	.05216	.02186	.00813	.00270	.00081
3.20	.76190	.54936	.36948	.22814	.12741	.06363	.02826	.01118	.00396	.00127
3.40	.77273	.56778	.39154	.24970	.14515	.07600	.03560	.01490	.00560	.00190
3.60	.78261	.58484	.41239	.27069	.16311	.08914	.04383	.01934	.00768	.00276
3.80	.79167	.60067	.43209	.29102	.18112	.10290	.05291	.02451	.01024	.00388
4.00	.80000	.61538	.45070	.31068	.19907	.11716	.06275	.03042	.01334	.00531
4.20	.80769	.62910	.46829	.32963	.21685	.13179	.07328	.03705	.01699	.00709
4.40	.81481	.64191	.48493	.34786	.23437	.14667	.08441	.04436	.02123	.00925
4.60	.82143	.65389	.50066	.36538	.25158	.16169	.09605	.05234	.02605	.01184
4.80	.82759	.66513	.51555	.38221	.26843	.17678	.10811	.06092	.03147	.01488
5.00	.83333	.67568	.52966	.39834	.28487	.19185	.12052	.07005	.03746	.01838
5.20	.83871	.68560	.54304	.41382	.30088	.20683	.13318	.07967	.04401	.02237
5.40	.84375	.69495	.55573	.42865	.31645	.22167	.14603	.08973	.05109	.02685
5.60	.84848	.70377	.56779	.44287	.33156	.23632	.15900	.10015	.05866	.03181
5.80	.85294	.71211	.57926	.45650	.34621	.25075	.17202	.11089	.06669	.03724
6.00	.85714	.72000	.59016	.46957	.36040	.26492	.18505	.12188	.07514	.04314
6.20	.86111	.72748	.60055	.48210	.37414	.27882	.19804	.13306	.08397	.04948
6.40	.86486	.73458	.61045	.49411	.38743	.29242	.21095	.14439	.09312	.05624
6.60	.86842	.74132	.61990	.50565	.40028	.30571	.22375	.15583	.10255	.06339
6.80	.87179	.74774	.62892	.51671	.41271	.31868	.23639	.16731	.11223	.07090
7.00	.87500	.75385	.63755	.52734	.42472	.33133	.24887	.17882	.12210	.07874
7.20	.87805	.75967	.64579	.53756	.43633	.34366	.26116	.19031	.13213	.08687
7.40	.88095	.76523	.65369	.54737	.44755	.35566	.27325	.20176	.14229	.09526
7.60	.88372	.77054	.66125	.55681	.45839	.36734	.28512	.21313	.15253	.10388
7.80	.88636	.77562	.66850	.56589	.46887	.37870	.29676	.22441	.16281	.11269
8.00	.88889	.78049	.67546	.57464	.47901	.38975	.30816	.23557	.17314	.12166
8.20	.89130	.78515	.68214	.58305	.48881	.40049	.31933	.24660	.18346	.13077
8.40	.89362	.78962	.68856	.59117	.49828	.41093	.33026	.25748	.19376	.13997
8.60	.89583	.79390	.69474	.59899	.50745	.42108	.34094	.26821	.20401	.14926
8.80	.89796	.79802	.70068	.60653	.51632	.43094	.35139	.27877	.21419	.15860
9.00	.90000	.80198	.70640	.61381	.52491	.44052	.36158	.28916	.22430	.16796
9.20	.90196	.80579	.71191	.62084	.53322	.44983	.37154	.29936	.23431	.17734
9.40	.90385	.80945	.71722	.62762	.54127	.45887	.38126	.30939	.24422	.18671
9.60	.90566	.81299	.72234	.63418	.54907	.46766	.39075	.31922	.25401	.19604
9.80	.90741	.81639	.72729	.64053	.55663	.47621	.40001	.32886	.26368	.20534
10.00	.90909	.81967	.73206	.64666	.56395	.48451	.40904	.33832	.27321	.21458

Table 4. (Continued) Probability of a transaction being lost, P

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Traffic Rate of Erlangs (=$M\rho$)	Number of Servers, M									
	$M=11$	$M=12$	$M=13$	$M=14$	$M=15$	$M=16$	$M=17$	$M=18$	$M=19$	$M=20$
1.50	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
2.00	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
2.50	.00005	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
3.00	.00022	.00006	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000
3.50	.00073	.00021	.00006	.00001	.00000	.00000	.00000	.00000	.00000	.00000
4.00	.00193	.00064	.00020	.00006	.00002	.00000	.00000	.00000	.00000	.00000
4.50	.00427	.00160	.00055	.00018	.00005	.00002	.00000	.00000	.00000	.00000
5.00	.00829	.00344	.00132	.00047	.00016	.00005	.00001	.00000	.00000	.00000
5.25	.01107	.00482	.00194	.00073	.00025	.00008	.00003	.00001	.00000	.00000
5.50	.01442	.00657	.00277	.00109	.00040	.00014	.00004	.00001	.00000	.00000
5.75	.01839	.00873	.00385	.00158	.00060	.00022	.00007	.00002	.00001	.00000
6.00	.02299	.01136	.00522	.00223	.00089	.00033	.00012	.00004	.00001	.00000
6.25	.02823	.01449	.00692	.00308	.00128	.00050	.00018	.00006	.00002	.00001
6.50	.03412	.01814	.00899	.00416	.00180	.00073	.00028	.00010	.00003	.00001
6.75	.04062	.02234	.01147	.00550	.00247	.00104	.00041	.00015	.00005	.00002
7.00	.04772	.02708	.01437	.00713	.00332	.00145	.00060	.00023	.00009	.00003
7.25	.05538	.02827	.01773	.00910	.00438	.00198	.00084	.00034	.00013	.00005
7.50	.06356	.03821	.02157	.01142	.00568	.00265	.00117	.00049	.00019	.00007
7.75	.07221	.04456	.02588	.01412	.00724	.00350	.00159	.00068	.00028	.00011
8.00	.08129	.05141	.03066	.01722	.00910	.00453	.00213	.00094	.00040	.00016
8.25	.09074	.05872	.03593	.02073	.01127	.00578	.00280	.00128	.00056	.00023
8.50	.10051	.06646	.04165	.02466	.01378	.00727	.00362	.00171	.00076	.00032
8.75	.11055	.07460	.04781	.02901	.01664	.00902	.00462	.00224	.00103	.00045
9.00	.12082	.08309	.05439	.03379	.01987	.01105	.00582	.00290	.00137	.00062
9.25	.13126	.09188	.06137	.03897	.02347	.01338	.00723	.00370	.00180	.00083
9.50	.14184	.10095	.06870	.04454	.02744	.01603	.00888	.00466	.00233	.00110
9.75	.15251	.11025	.07637	.05050	.03178	.01900	.01078	.11581	.00297	.00145
10.00	.16323	.11974	.08434	.05682	.03650	.02230	.01295	.00714	.00375	.00187
10.25	.17398	.12938	.09257	.06347	.04157	.02594	.01540	.00869	.00467	.00239
10.50	.18472	.13914	.10103	.07044	.04699	.02991	.01814	.01047	.00575	.00301
10.75	.19543	.14899	.10969	.07768	.05274	.03422	.02118	.01249	.00702	.00376
11.00	.20608	.15889	.11851	.08519	.05880	.03885	.02452	.01477	.00848	.00464
11.25	.21666	.16883	.12748	.09292	.06515	.04380	.02817	.01730	.01014	.00567
11.50	.22714	.17877	.13655	.10085	.07177	.04905	.03212	.02011	.01202	.00687
11.75	.23752	.18869	.14570	.10896	.07864	.05460	.03636	.02319	.01414	.00824
12.00	.24777	.19857	.15490	.11721	.08573	.06041	.04090	.02654	.01649	.00980
12.25	.25788	.20839	.16414	.12559	.09302	.06648	.04572	.03017	.01908	.01155
12.50	.26786	.21815	.17739	.13406	.10049	.07279	.05080	.03408	.02193	.01352
12.75	.27768	.22782	.18263	.14261	.10811	.07932	.05615	.03825	.02503	.01570
13.00	.28735	.23740	.19185	.15121	.11587	.08604	.06173	.04268	.02838	.01811
13.25	.29686	.24687	.20103	.15985	.12373	.09294	.06755	.04737	.03198	.02074
13.50	.30621	.25622	.21016	.16850	.13168	.10000	.07357	.05229	.03582	.02361
13.75	.31539	.26545	.21922	.17716	.13971	.10719	.07978	.05744	.03991	.02671
14.00	.32441	.27456	.22820	.18580	.14799	.11451	.08617	.06281	.04424	.03004
14.25	.33325	.28353	.23711	.19442	.15690	.12192	.09272	.06839	.04879	.03359
14.50	.34193	.29237	.24591	.20299	.16404	.12942	.09941	.07415	.05355	.03738
14.75	.35045	.30107	.25462	.21152	.17218	.13699	.10623	.08008	.05853	.04138
15.00	.35879	.30963	.26322	.21998	.18032	.14460	.11315	.08617	.06270	.04559
15.50	.37499	.32631	.28009	.23670	.19652	.15993	.12726	.09876	.07456	.05463
16.00	.39055	.34242	.29649	.25309	.21257	.17531	.14163	.11181	.03606	.06441
16.50	.40548	.35796	.31240	.26910	.22840	.19064	.15614	.12521	.09807	.07485
17.00	.41981	.37293	.32781	.28472	.24396	.20585	.17071	.13884	.11050	.08586
17.50	.43356	.38736	.34273	.29992	.25921	.22089	.18526	.15262	.12325	.09734
18.00	.44675	.40124	.35715	.31469	.27411	.23569	.19972	.16647	.13623	.10921
18.50	.45942	.41461	.37108	.32902	.28866	.25024	.21403	.18031	.14935	.12138
19.00	.47158	.42748	.38453	.34291	.30282	.26449	.22816	.19409	.16254	.13376
19.50	.48325	.43987	.39752	.35637	.31660	.27843	.24206	.20775	.17575	.14629
20.00	.49447	.45179	.41005	.36940	.33000	.29203	.25571	.22126	.18891	.15889

Table 4. (Continued) Probability of a call being lost, P

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Traffic Rate in Erlangs	Number of Servers, M									
$(=M\rho)$	$M=21$	$M=22$	$M=23$	$M=24$	$M=25$	$M=26$	$M=27$	$M=28$	$M=29$	$M=30$
6.00	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
6.50	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
7.00	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
7.50	.00003	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
8.00	.00006	.00002	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000
8.50	.00013	.00005	.00002	.00001	.00000	.00000	.00000	.00000	.00000	.00000
9.00	.00026	.00011	.00004	.00002	.00001	.00000	.00000	.00000	.00000	.00000
9.50	.00050	.00022	.00009	.00004	.00001	.00000	.00000	.00000	.00000	.00000
10.00	.00089	.00040	.00018	.00007	.00003	.00001	.00000	.00000	.00000	.00000
10.50	.00150	.00072	.00033	.00014	.00006	.00002	.00001	.00000	.00000	.00000
11.00	.00242	.00121	.00058	.00027	.00012	.00005	.00002	.00001	.00000	.00000
11.50	.00375	.00195	.00098	.00047	.00022	.00010	.00004	.00002	.00001	.00000
12.00	.00557	.00303	.00158	.00079	.00038	.00017	.00008	.00003	.00001	.00001
12.50	.00798	.00452	.00245	.00127	.00064	.00031	.00014	.00006	.00003	.00001
13.00	.01109	.00651	.00367	.00198	.00103	.00051	.00025	.00011	.00005	.00002
13.50	.01495	.00909	.00531	.00298	.00160	.00083	.00042	.00020	.00009	.00004
14.00	.01963	.01234	.00745	.00433	.00242	.00130	.00067	.00034	.00016	.00008
14.50	.02516	.01631	.01018	.00611	.00353	.00197	.00105	.00055	.00027	.00013
15.00	.03154	.02105	.01354	.00839	.00501	.00288	.00160	.00086	.00044	.00022
15.50	.03876	.02658	.01760	.01124	.00692	.00411	.00235	.00130	.00069	.00036
16.00	.04678	.03290	.02238	.01470	.00932	.00570	.00337	.00192	.00106	.00056
16.50	.05555	.03999	.02789	.01881	.01226	.00772	.00470	.00276	.00157	.00086
17.00	.06499	.04782	.03414	.02361	.01580	.01023	.00640	.00387	.00226	.00128
17.50	.07503	.05632	.04109	.02909	.01996	.01326	.00852	.00530	.00319	.00185
18.00	.08560	.06545	.04873	.03526	.02476	.01685	.01111	.00709	.00438	.00262
18.50	.09660	.07513	.05699	.04208	.03020	.02103	.01421	.00930	.00590	.00362
19.00	.10796	.08528	.06582	.04952	.03627	.02582	.01785	.01197	.00778	.00490
19.50	.11959	.09584	.07515	.05755	.04296	.03121	.02205	.01512	.01007	.00650
20.00	.13144	.10673	.08493	.06610	.05022	.03720	.02681	.01879	.01279	.00846
20.50	.14342	.11789	.09508	.17512	.05802	.04375	.03215	.02299	.01599	.01081
21.00	.15548	.12924	.10554	.08454	.06631	.05083	.03803	.02773	.01969	.01359
21.50	.16758	.14072	.11625	.09432	.07503	.05842	.04445	.03301	.02389	.01683
22.00	.17960	.15230	.12715	.10439	.08413	.06646	.05137	.03880	.02859	.02054
22.50	.19168	.16390	.13818	.11469	.09356	.07490	.05875	.04508	.03380	.02472
23.00	.20361	.17550	.14930	.12517	.10327	.08370	.06656	.05184	.03949	.02939
23.50	.21542	.18706	.16046	.13578	.11319	.09281	.07474	.05903	.04565	.03452
24.00	.22709	.19855	.17162	.14648	.12329	.10217	.08326	.06661	.05225	.04012
24.50	.23860	.20993	.18275	.15723	.13351	.11175	.09207	.07455	.05925	.04616
25.00	.24993	.22119	.19382	.16798	.14382	.12149	.10112	.08281	.06663	.05260
25.50	.26107	.23231	.20481	.17872	.15418	.13136	.11037	.09133	.07434	.05943
26.00	.27201	.24326	.21568	.18940	.16456	.14131	.11978	.10009	.08235	.06661
26.50	.28274	.25405	.22643	.20001	.17493	.15131	.12931	.10904	.09061	.07411
27.00	.29327	.26466	.23704	.21053	.18525	.16134	.13893	.11814	.09909	.08188
27.50	.30358	.27509	.24750	.22094	.19552	.17136	.14860	.12736	.10776	.08990
28.00	.31367	.28532	.25780	.23122	.20570	.18135	.15830	.13666	.11657	.09812
28.50	.32355	.29535	.26792	.24137	.21578	.19129	.16799	.14602	.12550	.10652
29.00	.33322	.30519	.27788	.25137	.22576	.20115	.17767	.15542	.13451	.11507
29.50	.34267	.31483	.28765	.26121	.23561	.21094	.18730	.16481	.14358	.12372
30.00	.35190	.32426	.29724	.27090	.24533	.22062	.19687	.17419	.15268	.13246
30.50	.36093	.33350	.30664	.28041	.25490	.23019	.20637	.18354	.16180	.14126
31.00	.36975	.34255	.31586	.28977	.26433	.23964	.21577	.19283	.17090	.15009
31.50	.37838	.35139	.32490	.29895	.27361	.24896	.22508	.20225	.17997	.15894
32.00	.38680	.36005	.33375	.30796	.28274	.25815	.23428	.21120	.18900	.16778
32.50	.39503	.36851	.34242	.31680	.29170	.26720	.24336	.22025	.19797	.17659
33.00	.40307	.37679	.35091	.32546	.30051	.27611	.25232	.22921	.20687	.18537
33.50	.41092	.38489	.35922	.33396	.30916	.28486	.26114	.23806	.21569	.19410
34.00	.41860	.39281	.36736	.34229	.31764	.29348	.26984	.24680	.22441	.20277
34.50	.42610	.40055	.37532	.35045	.32597	.30194	.27840	.25542	.23304	.21136
35.00	.43342	.40812	.38312	.35845	.33414	.31025	.28682	.26391	.24157	.21987

Table 4. (Continued) Probability of a call being lost, P

How to Calculate the Number of Terminals You Will Need in Your Communications Network

Traffic Rate in Erlangs	Number of Servers, M:									
($=M\rho$)	35	40	45	50	55	60	70	80	90	100
15.00	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
16.00	.00002	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
17.00	.00005	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
18.00	.00013	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
19.00	.00031	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
20.00	.00069	.00003	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
21.00	.00139	.00007	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
22.00	.00262	.00017	.00001	.00000	.00000	.00000	.00000	.00000	.00000	.00000
23.00	.00458	.00037	.00002	.00000	.00000	.00000	.00000	.00000	.00000	.00000
24.00	.00751	.00075	.00004	.00000	.00000	.00000	.00000	.00000	.00000	.00000
25.00	.01165	.00141	.00009	.00000	.00000	.00000	.00000	.00000	.00000	.00000
26.00	.01715	.00250	.00020	.00001	.00000	.00000	.00000	.00000	.00000	.00000
27.00	.02413	.00418	.00041	.00002	.00000	.00000	.00000	.00000	.00000	.00000
28.00	.03261	.00662	.00077	.00005	.00000	.00000	.00000	.00000	.00000	.00000
29.00	.04253	.00999	.00137	.00011	.00001	.00000	.00000	.00000	.00000	.00000
30.00	.05377	.01444	.00232	.00022	.00001	.00000	.00000	.00000	.00000	.00000
31.00	.06617	.02006	.00375	.00042	.00003	.00000	.00000	.00000	.00000	.00000
32.00	.07954	.02689	.00579	.00076	.00006	.00000	.00000	.00000	.00000	.00000
33.00	.09367	.03493	.00856	.00130	.00012	.00001	.00000	.00000	.00000	.00000
34.00	.10838	.04412	.01219	.00212	.00023	.00002	.00000	.00000	.00000	.00000
35.00	.12348	.05435	.01677	.00334	.00042	.00003	.00000	.00000	.00000	.00000
36.00	.13883	.06550	.02234	.00504	.00072	.00007	.00000	.00000	.00000	.00000
37.00	.15428	.07742	.02893	.00734	.00120	.00013	.00000	.00000	.00000	.00000
38.00	.16972	.08996	.03650	.01034	.00192	.00023	.00000	.00000	.00000	.00000
39.00	.18506	.10299	.04501	.01411	.00295	.00040	.00000	.00000	.00000	.00000
40.00	.20023	.11637	.05436	.01870	.00439	.00068	.00000	.00000	.00000	.00000
41.00	.21517	.12998	.06445	.02416	.00631	.00110	.00001	.00000	.00000	.00000
42.00	.22984	.14372	.07519	.03047	.00880	.00172	.00002	.00000	.00000	.00000
43.00	.24420	.15749	.08644	.03760	.01193	.00260	.00004	.00000	.00000	.00000
44.00	.25823	.17123	.09811	.04551	.01576	.00382	.00007	.00000	.00000	.00000
45.00	.27192	.18487	.11009	.05412	.02032	.00543	.00013	.00000	.00000	.00000
46.00	.28526	.19837	.12229	.06336	.02561	.00752	.00022	.00000	.00000	.00000
47.00	.29824	.21167	.13463	.07313	.03164	.01014	.00036	.00000	.00000	.00000
48.00	.31086	.22475	.14703	.08335	.03837	.01335	.00058	.00001	.00000	.00000
49.00	.32313	.23760	.15943	.09393	.04576	.01719	.00090	.00001	.00000	.00000
50.00	.33505	.25018	.17178	.10479	.05374	.02166	.00137	.00002	.00000	.00000
51.00	.34662	.26250	.18404	.11586	.06226	.02679	.00202	.00004	.00000	.00000
52.00	.35785	.27453	.19618	.12707	.07124	.03255	.00289	.00007	.00000	.00000
53.00	.36876	.28628	.20815	.13836	.08061	.03891	.00405	.00012	.00000	.00000
54.00	.37934	.29775	.21995	.14967	.09030	.04584	.00554	.00019	.00000	.00000
55.00	.38962	.30893	.23155	.16097	.10024	.05328	.00741	.00031	.00000	.00000
56.00	.39960	.31983	.24295	.17222	.11037	.06119	.00971	.00048	.00001	.00000
57.00	.40928	.33045	.25412	.18338	.12065	.06950	.01247	.00073	.00001	.00000
58.00	.41869	.34079	.26507	.19442	.13101	.07815	.01572	.00107	.00002	.00000
59.00	.42782	.35087	.27579	.20534	.14141	.08708	.01947	.00155	.00004	.00000
60.00	.43669	.36068	.28628	.21611	.15182	.09625	.02374	.00220	.00006	.00000
61.00	.44531	.37024	.29653	.22671	.16220	.10560	.02851	.00304	.00010	.00000
62.00	.45368	.37955	.30656	.23714	.17253	.11508	.03377	.00412	.00016	.00000
63.00	.46183	.38861	.31635	.24739	.18278	.12466	.03950	.00548	.00026	.00000
64.00	.46974	.39744	.32592	.25745	.19292	.13428	.04566	.00716	.00039	.00001
65.00	.47744	.40604	.33527	.26732	.20296	.14393	.05222	.00917	.00058	.00001
66.00	.48493	.41441	.34439	.27700	.21287	.15357	.05914	.01157	.00084	.00002
67.00	.49222	.42257	.35330	.28648	.22263	.16318	.06637	.01436	.00120	.00003
68.00	.49931	.43053	.36201	.29578	.23226	.17273	.07388	.01755	.00167	.00006
69.00	.50621	.43828	.37050	.30487	.24172	.18221	.08162	.02117	.00230	.00009
70.00	.51294	.44584	.37880	.31378	.25104	.19160	.08956	.02520	.00309	.00014
71.00	.51949	.45320	.38690	.32250	.26019	.20089	.09765	.02963	.00409	.00021
72.00	.52587	.46039	.39481	.33104	.26917	.21007	.10587	.03446	.00532	.00031
73.00	.53209	.46740	.40254	.33939	.27800	.21913	.11418	.03966	.00682	.00046
74.00	.53815	.47423	.41009	.34756	.28666	.22806	.12256	.04521	.00860	.00065
75.00	.54407	.48091	.41746	.35556	.29515	.23686	.13097	.05107	.01069	.00092

Table 4. (Continued) Probability of a call being lost, P

How to Calculate the Number of Terminals You Will Need in Your Communications Network

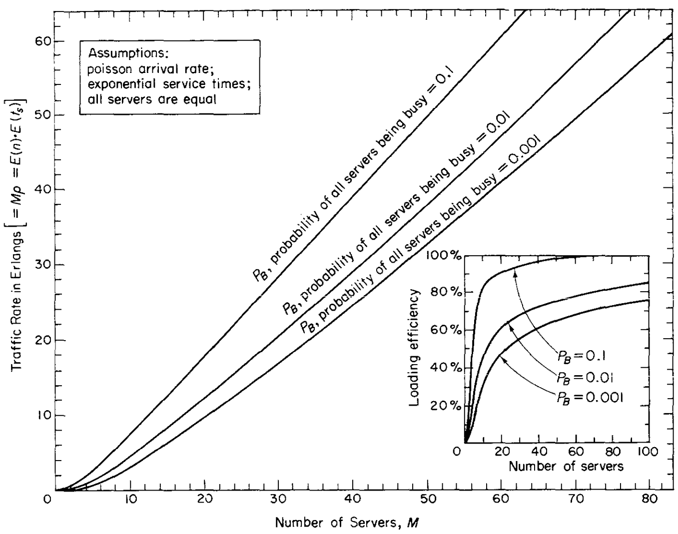


Figure 2. The numbers of servers required to give different levels of "grade of service" on a nonqueuing multiserver system

0.001. In the above example in which Mp = 20 we can read from these curves that to achieve a 0.01 chance of not finding a terminal we need 30 terminals. To achieve a 0.1 chance, we need 23 terminals; and to achieve a 0.001 chance, we need 35 terminals.

The inset shows how the loading efficiency increases with the number of servers for a given P criterion. Above about 40 terminals the improvement with an increased number of terminals is becoming less pronounced. For the criterion in the above example, P = 0.01, the slope of the efficiency curve is considerably steeper below 30 terminals. If there are less than 30 terminals it certainly pays to pool them and make all terminals equal.

PRACTICAL EXPERIENCE

How has the use of these models worked out in practice? We can compare our mathematical model against a simulation of the device or, eventually, against the performance of the device itself. We can make a precise statement about the accuracy of the model because the device behaves in a mechanically determined manner. In this report, however, we are talking mainly about people. The users and operators who come into contact with a system have a variety of idiosyncrasies and feel under no constraint to obey the assumptions of queuing theory.

Several decades of experience have been obtained in applying queuing theory to the design of telephone exchanges and trunks. On the other hand, experience in observing the efficacy of queuing theory as applied to terminals and operators is limited.

The holding times for telephone calls have been found in practice to be sufficiently close to an exponential distribution to make models based on this assumption give useful results.[1] Such models have been widely employed for many years. Simulation has also been used. Before the advent of computers, special-purpose machines were built to assist in simulating the behavior of exchanges.

The use of queuing equations in telephone exchange design has five advantages in its favor which may not all be present when terminals are involved. Let us discuss these.

Poissonian Input

The exchange serves a large population of users who are truly independent of one another. This means that the traffic volume follows a Poisson distribution very closely. Some computer terminals also serve a large population of independent users—airline terminals dealing with telephone reservations, for example—and here, as one might expect, measurements of the traffic volume indicate that it is indeed Poissonian.

On certain systems, however, the nature of the work might be such that clustering of traffic occurs, which would give rise to a higher waiting time than otherwise. In a factory when it is time for a work break, or in a university when a class ends, a cluster of transactions might arrive. Again, if the transactions are building up to some specific event, then the result may be clustering.

A. M. Lee describes a case in which the "transactions" were passengers arriving to check in for departing BEA flights at London Airport.[2] Before a flight, passengers arrived at times distributed as in Figure 3. There were flights with this arrival distribution every 5 to 10 minutes, and the sum of their passenger arrival rates was the input to the passenger queue. It was not clear how far such an arrival rate would deviate from the Poisson distribution, and so a count of arriving passengers was made. Figure 4 compares the results of this with a Poisson distribution having the same mean. It was decided in this case that it was sufficiently close to assume Poissonian input to the model.

In some systems there can be a specific cause of clustering, which the systems analyst should look out for. In others, and this is probably more common, the transactions go through some form of buffering

[1]Charles Clos and Roger I. Wilkinson, "Dialing Habits of Telephone Customers," *Bell System Tech. J.*, January 1952.
[2]A. M. Lee, Applied Queuing Theory (London: Macmillan, and New York: St. Martin's Press, 1966).

How to Calculate the Number of Terminals You Will Need in Your Communications Network

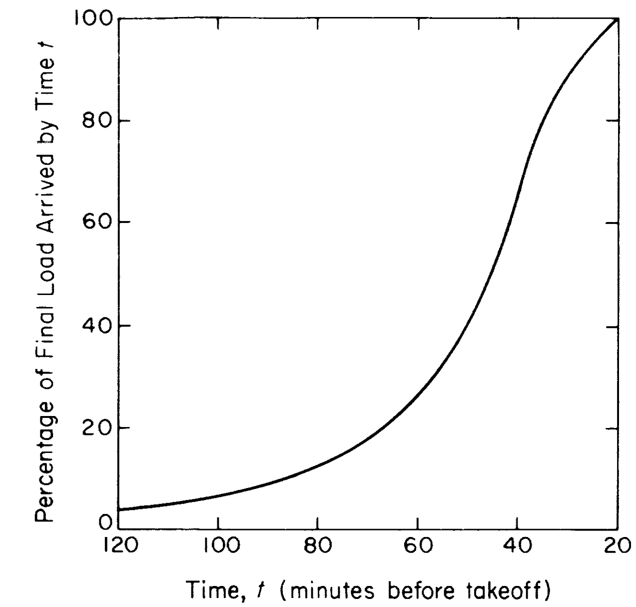


Figure 3. Passenger arrival pattern for a flight. Redrawn with permission from A. M. Lee, Applied Queuing Theory, MacMillan, London and St. Martin's Press, New York, 1966.

process before they reach the queue, thus evening out the interarrival times. In a registration scheme in which terminals are used, the clients have to fill in a form at a preliminary desk before reaching the clerk with the terminal. As a result, they reach the clerk somewhat more evenly than if they had walked straight in off the street. The systems analyst who does not want to use simulation may usually claim with reasonable justification that the Poissonian arrival rate assumption errs only slightly—and errs on the side of conservatism and safety.

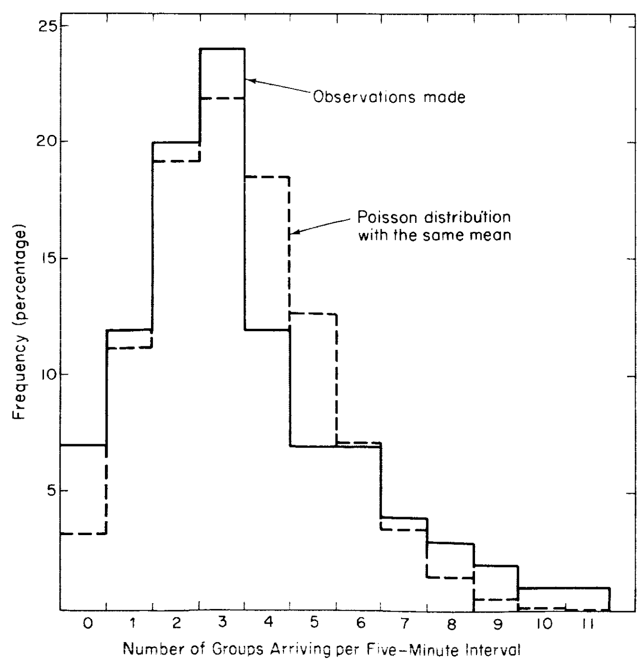


Figure 4. The analysts felt justified in using a Poisson input distribution.

Exceptionally Long Service Times

If the service times are found to be not exponential, the approximation in Eq. 2 might be used. Occasionally, however, when people are involved, the service time is going to be much longer than normal. For instance, the person standing front of us in a bank teller's line often seems to have some mysterious business that takes forever. Certain women on the telephone will defy any exponential distribution. The effect of a solitary long service time is going to be serious if there is only one server. It will cause delays if there are two or three servers. On the other hand, if there are a large number of servers, it will have very little effect. In a public telephone exchange, there are a large number of circuits that the calls can use, and thus one person who stays on the telephone all evening, will not degrade the service appreciably.

Using Table 1, if we have 20 servers and a facility utilization p of 0.8, we have a mean waiting time of 0.064 times the mean service time. If one server is effectively taken out of action (by a long telephone call, for example) and the remainder then have a facility utilization of 0.8 X 20/19 = 0.84, the mean waiting time is then 0.121 times the service time. If, on the other hand, there had been 6 servers and one is taken out of action, the facility utilization rises from 0.8 to 0.8 X 6/5 = 0.96. The mean waiting time rises from 0.431 to 4.508 times the mean service time, and the system is dangerously close to the condition of p = 1, in which it cannot handle all the transactions.

Moral: If there are likely to be some users who tie up a server for long periods, evaluate the effect of this factor in terms of the number of servers needed.

Variation in Service Time

In the examples in this report we have treated service time as though it were constant. In most cases it is. One law, however, that is not found in the works of mathematical statistics is Parkinson's law. Some terminal operators, when confronted with a long queue of people, can be observed to speed up remarkably, and when there is no queue other than the person being served, they tend to slow down, indulge in conversation, and perhaps give extra help to their customer. The queuing theory version of Parkinson's law is "t varies to maintain p close to 1."

This human phenomenon is clearly useful in some ways, but it means that we must not take our equations too literally. If the servers are observed to

How to Calculate the Number of Terminals You Will Need in Your Communications Network

be fully occupied during a nonpeak period, there is no need to panic. On the other hand it would not be a good idea to design the terminal arrangement for the minimum t_s the operator was capable of.

Interestingly, the variation in operator service time does not seem to occur in the author's experience on some systems with telephone customers. The operator does not know how big the queue is, and therefore all customers receive equal treatment, except perhaps when there are long gaps between customers. Is this good? Not necessarily. It is often desirable to have the terminal operator speed up when there is a queue. A mechanism for flashing a red light at the terminals when telephone calls are being queued could be worthwhile.

Selection of Servers

An important assumption in multiserver queuing theory is that the transactions form one queue and each item in turn is served by the first server to become free. Probably the most important deviation from theory that occurs in practice is that lines of standing people do not behave this sensibly. If a bank has four tellers with identical functions, the best way to marshal the customers would be to make them form one line and then have each teller, on becoming free, take the next person standing in line. In reality, four lines form and customers tend to transfer from one line to another in order to maintain the shortest number of people in front of them. We have assumed that this freedom in queue switching gives approximately the same effect as having one queue. We could, if we wished, have used some rather more complex queuing theory which relates to queues with jockeying. In practice, however, many people standing in lines seem remarkably reluctant to switch queues. Consequently, the mean waiting time is longer than that predicted by the equations. The equations seem to have predicted queue sizes reasonably well when there are just two servers. The inaccuracy with larger numbers of servers seems greater the greater their number.

British European Airline (BEA), in their study of queuing for passenger check-in counters, recorded passenger queuing times at two advance registration counters at London Airport. The measurements made are shown in Figure 5. These measurements agree as closely as could be expected with the two-server queuing equations. The agreement between the theory and practice was so encouraging that BEA went on to use the same equations to design check-in arrangements with more than two servers. Lee has reported the interesting way in which the theory then broke down.

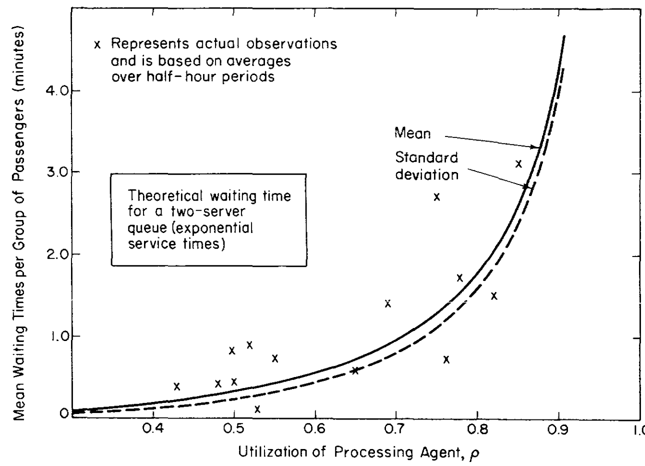


Figure 5. The agreement between theory and actuality. Redrawn with permission from A. M. Lee, Applied Queuing Theory, Macmillan, London, and St. Martin's Press, New York, 1966

Figure 6 shows the passenger check-in area. Counters 2 to 8 were commonly in use, and so a multiserve queuing model with seven servers might have been expected to predict the queue sizes. Such a model, however, assumes that the passengers distribute themselves evenly between the counters, which did not happen. Figure 7 shows how they did distribute themselves. The counters near the doors were more popular than the more distant ones. Furthermore, the passengers seemed remarkably reluctant to leave one queue and join another, shorter one. Sometimes there was a line at counter 7, for example, and none at counters 3 and 4.

The BEA operations research team observed that

"Jockeying for position in queues seemed to happen in epidemics, but even so it slowly began to dawn

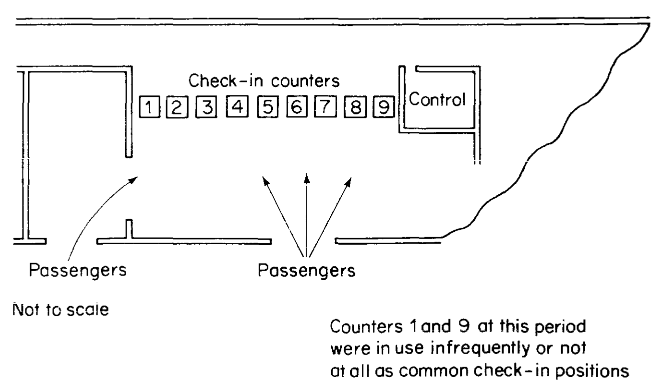


Figure 6. Passenger check-in counters at London Airport. Reproduced with permission from A.M. Lee, Applied Queuing Theory, Macmillan, London and St. Martin's Press, New York, 1966

How to Calculate the Number of Terminals You Will Need in Your Communications Network

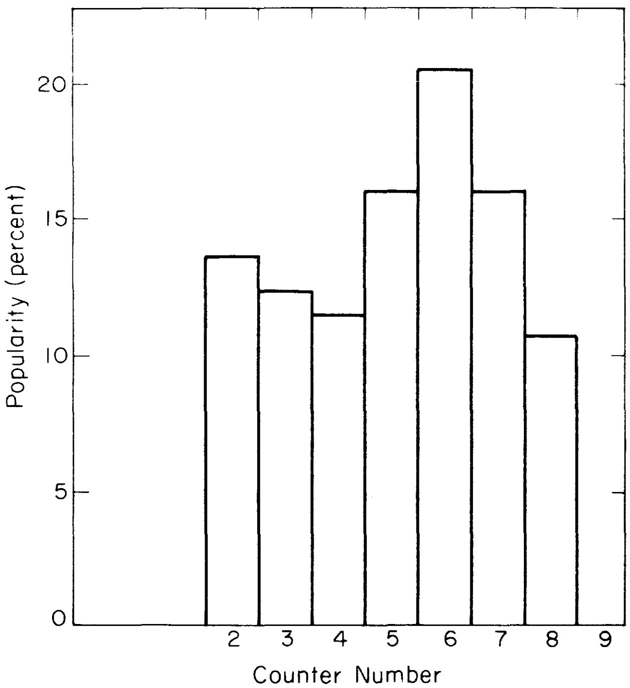


Figure 7. As shown here, the passengers did not distribute themselves evenly over the seven counters in use. This and their reluctance to switch queues, freely, gave a longer queuing time than was expected from multiserver queuing theory. Reproduced with permission from A. M. Lee, Applied Queuing Theory, Macmillan, London and St. Martin's Press, New York, 1966

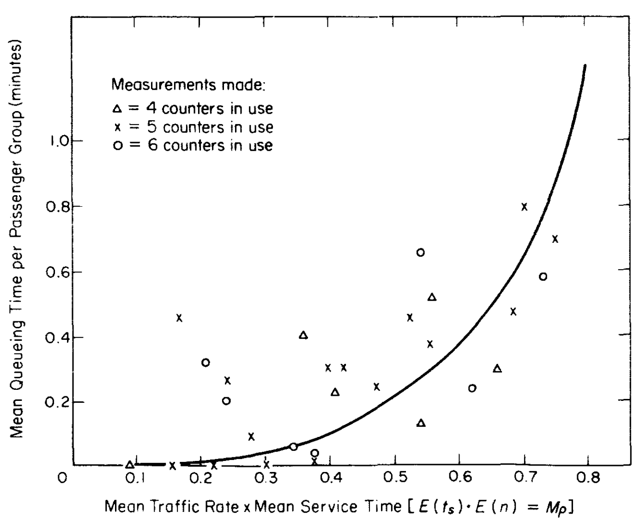


Figure 8. Theoretical waiting time with three servers in use. Reproduced with permission from A. M. Lee, Applied Queuing Theory, Macmillan, London and St. Martin's Press, New York, 1966

[3]A. M. Lee, *Applied Queuing Theory*, London: Macmillan and New York: St. Martin's Press, 1966.

upon the team that it was of a formal nature. If ready to jockey at all, passengers appeared willing to move no farther than one queue to either side of their original positions, and no passenger appeared ready to move more than once. If true this would imply that irrespective of the number of counters provided, the effect insofar as the waiting time of individual passengers was concerned would be no better than that given by groups of three counters with limited availability.[3]"

The latter conclusion seems to be supported by Figure 8. Here the mean waiting times over 15-minute periods were observed, and plotted, when 4, 5, and 6 counters were in operation. The multiserver queuing model for 3 servers would have given a better prediction of the queuing times than the models for 4, 5, or 6 servers.

The reluctance to switch lines may have been partly due to the fact that the passengers were carrying luggage. It may have also been due to the fact that the British seem to observe curiously disciplined queuing behavior. However, to some extent the same phenomenon can be observed elsewhere where physical lines of people queue for the services of operators with terminals. In one New York savings bank with a dozen or so tellers using real-time terminals, it is common to see some tellers idle while others have a queue.

It is interesting to note that, again, this phenomenon does not seem to occur where the terminal operators are handling telephone customers. Here the calls are automatically distributed to the next free operator and the multiserver queuing equations generally agree well with actual waiting times.

Behavior During Overloaded Periods

If the queues at an airline check-in desk became longer than anticipated, many passengers who arrive on time fail to reach the departure desk before the flight is closed. As a result, they complain, usually long and loud. While complaining, they sometimes block the counter, thereby increasing the delays. The overload here can be self-amplifying.

On other systems the opposite is true. If you keep a telephone caller waiting long enough, he gives up. The mean waiting time for the successful caller is then less than would be predicted, but some customers might be lost. In designing a system, it is generally desirable to work out what the effect of long waits will be.

FACE-TO-FACE SERVICE

In general, it seems that the worst deviations for the queuing models occur when the queues are

How to Calculate the Number of Terminals You Will Need in Your Communications Network

lines of people who come face-to-face with whoever serves them. In such situations, human idiosyncracies take over. Telephone queues are better disciplined and obey the queuing models better. Theory and practice seem to agree reasonably well with telephone queues except when the caller is kept waiting too long. With physical queues, it is perhaps unwise to assume an even distribution over more than four servers. Where there are many terminals, one might calculate the waiting time by using the four-server model.

Example 5. Number of Terminal Operators Needed (again). *In view of what has just been said, let us take another look at the way we did Example 2. We concluded that nine teller windows in use would meet the design criterion that the mean queue for each teller should not exceed two customers. In reaching this conclusion, however, we used the multiserver queuing model for nine servers, which means that we were assuming, in effect, that the customers would distribute themselves evenly among the nine tellers and switch queues freely. The foregoing BEA experience indicates that such would not be the case and that the queue sizes would not be better than those for a four-server model.*

Let us do the calculation again, using a four-server model. It is still true that

$$\rho = \frac{E(n)E(t_s)}{M} = \frac{8}{M}$$

where M is the number of tellers.

If we have 9 tellers,

$$\rho = \frac{8}{9} = 0.889$$

Using Table 1 again, but now taking the times for a four-server model with p = 0.889, we have approximately

$$\frac{E(t_q)}{E(t_s)} = 2.75$$

Therefore $\quad E(t_q) = 2.75 \times 2 = 5.5$

$$E(q) = E(n)E(t_q)$$

$$= 4 \times 5.5 = 22$$

We thus have 22 persons queuing for 9 tellers, which is more than 2 persons per teller, and so the design criterion is not met.

Let us try 10 tellers.

$$\rho = \frac{8}{M} = 0.8$$

From Table 1, taking the times for a four-server model with p = 0.8,

$$\frac{E(t_q)}{E(t_s)} = 1.746$$

$$E(t_q) = 2 \times 1.746 = 3.492$$

$$E(q) = E(n)E(t_q)$$

$$= 4 \times 3.492 = 14.0$$

Thus there are 1.4 persons per teller, on the average, and the design criterion is met. If we still permit no more than 3 tellers to one terminal, 10 tellers will need 4 terminals. Adding one more terminal to cover breakdowns, we obtain a total of 5 terminals on the system.

Doing the same calculation, using a three-server model, gives 1.66 persons per teller, and therefore it appears that 10 teller windows will be enough even if the customers queue as badly as London Airport passengers.□

How to Measure a Terminal's IQ

Problem:

Many years ago we worked with an older editor who perversely insisted that machines were not people. He relentlessly edited all anthropomorphisms (attributing human qualities to nonhuman things) out of every technical piece that crossed his desk and left an indelible blue-pencil imprint on our impressionable mind. The old editor partially won his battle. We do not permit computers to touch, see, or feel, but we do permit them to communicate, sometimes to talk, and (we allow this next one only with much trepidation) to think. Unfortunately, thought implies intelligence (we can hear a blue pencil snap somewhere), and intelligence can be graded—dumb, smart, very smart, genius. So we have dumb terminals, smart or intelligent terminals, and now even genius terminals, all perfectly acceptable descriptors in the new lexicography of machine intelligence. But, out of deference to our editorial ghosts, we must stress that machine intelligence is not the same thing as people intelligence, and we offer this excellent report to clarify precisely what factors make terminals more or less intelligent. You will also gain some valuable insights into how to assess the intellectual impacts of various terminal features and to further assess their cost/benefit tradeoffs when it comes time to go out and buy your terminals.

Solution:

Terminals vary from simple devices with no storage and no circuitry worth gracing with the word "logic," to highly complex machines with microprogramming or stored-program computing capability. A common terminal in the future may be "the mini-computer in the office."

The advantage of very simple terminals is that they are inexpensive. If there are many terminals and one computer on a system, it makes sense to keep the terminals cheap—and to put the logic in the computer. However, where the communication lines are expensive, a new factor enters the argument: can we lower the overall communication bill by building more logic functions into the terminals? Or, better, can we lower the overall system cost?

There are other arguments besides the question of cost. Can we create a better man-machine interface if we make the terminal more elaborate? Can we lessen the number of errors? Can we do something to lessen the inconvenience when the computer or communication line fails?

Logic capability in the terminal used to be expensive. Furthermore, it used to be somewhat unreliable, and the cost of maintenance engineers chasing round to fix the large quantity of terminals is much higher than

---

From *Systems Analysis for Data Trasmission* by James Martin. Chpt. 16, pp 215-226. © 1978 by Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

when they are merely servicing the computers. Consequently, many unbuffered, asynchronous, simple-minded terminals are still being sold. The cost of logic, however, is dropping fast, and its reliability, with the advent of large-scale integration circuitry, is becoming very high. Buffering and logic can reside on a single mass-produced silicon chip in the terminal. Minicomputers are falling in price, like other logic devices, much faster than the drop in price of communication lines. Therefore, the whole question of how much intelligence should reside in the terminal is being re-evaluated. The systems analyst will be confronted with a much wider diversity of choice.

Let us now review the types of operations that might be performed at the terminal and do so in order of increasing logic requirement.

GENERATION OF BITS TO BE SENT AND INTERPRETATION OF BITS RECEIVED

The terminal generates the bits to be sent, including START and STOP bits if asynchronous transmission is being used. It translates the bits received into appropriate mechanical action. In the simplest form of network, this would be the only function (other than data set functions) carried on away from the computer center. Synchronous transmission is more efficient than START-STOP, but it requires more terminal circuitry.

TERMINAL CLUSTERING

Several different input-output devices may be attached to the same control unit. The terminal may be designed so that only one such device can operate at once, or there may be several operators using different devices simultaneously. In the latter case, some means of interleaving the separate transmissions will be necessary. This may require buffer storage. The control unit must contain the logic for addressing the different devices.

FUNCTIONS ASSOCIATED WITH MULTI-DROP OPERATION

For reasons of economy, several terminals, often at different locations, are attached to one communication line. In a polled system, the terminal must recognize its address on messages sent to it and must ignore other messages. It must respond logically to polling, go-ahead, and end-of-transmission signals. The functions can be carried out by a control unit that serves one terminal device or several. The same is true for devices attached to a synchronously controlled loop.

Early polling schemes had minimal logic in the terminal, and many of these terminals are still being marketed. The terminal has no buffer, and roll-call polling is used. To achieve fast response times on manual terminals, a buffer is needed, and hub polling is better than roll call. Hub polling reduces the number of line turnarounds by making each termination relay the polling messages on to the next terminal if it has nothing to transmit. This process requires more logic in the terminal; in particular, it requires a terminal to change its action when its normal addressee is inoperative.

Again a continuous loop between the terminals using synchronous transmission increases the line efficiency but also increases the terminal logic requirements.

Such increases in logic requirement do not appear so disadvantageous in the era of large-scale integration.

ERROR CONTROL AND RETRANSMISSION

Certain elaborate codes can give a very high measure of protection from transmission errors. Virtually no error can slip by undetected. The most powerful error-detecting codes have not been used on the majority of terminals because of the amount of logic required. Automatic retransmission of items in error needs further logic and, in particular, requires each item to be stored until it is known that it has been received correctly. Some error-detecting terminals have not been equipped for automatic retransmission. Furthermore, during a period of noise on the line, it may be necessary for the terminal to try several times to send a message. It may vary the time between successive tries or even switch its modem to half-speed operation to improve the chances of success. How much of such logic is worth building into the terminal?

Again a combination of error-detecting codes may be used, the error-correcting code to minimize the amount of retransmission and the error-detecting code to ensure that uncorrected errors cannot slip through.

SECURITY

Terminal features are needed to enhance the security of the system. The terminal may respond with a unique identification number when interrogated. Otherwise most of the logic operations associated with security may be in the central computer. If cryptography is used, as it may be for the transmission of highly confidential data, greater levels of safety will require more complex logic operations.

FAILURE PROTECTION

When the computer becomes inoperative or inaccessible, the terminal may become useless, or it may carry on some functions of its own. Where the terminal is being used for data entry, it may be able to record data for later transmission to the computer. If it has

How to Measure a Terminal's IQ

some of the following features, the terminal may be able to carry on a limited dialogue with its user.

When a leased communication line fails, the terminal may automatically dial an alternate connection to the computer—switching its modem operation accordingly.

AUTOMATIC ANSWERING

An idling terminal may have the ability to respond with operator attention to a signal from the computer. In this way, the computer may send unsolicited data for printing or recording.

EDITING

A terminal operator may wish to edit his input at the terminal. He may, for example, make mistakes and want to backspace and correct them. He may want to insert words or figures into a format laid out on the terminal screen by the computer. On many systems the code for doing so—for example, two backspace characters followed by two replacement characters—travels all the way to the central computer for editing. Some terminal control units, however, have editing logic in them. For instance, the edited message may appear on a terminal screen and be checked by the operator before it is transmitted. Editing functions could also take place in a concentrator as part of the work of assembling a message for synchronous transmission.

Editing a message from an operator before transmission slightly reduces the number of characters to be transmitted. However, the reduction may not be enough to pay for the added terminal logic required. On the other hand, logic for formatting the messages from the computer may bring a very substantial saving. Suppose, for the purpose of illustration, that it is desired to display on the terminal screen tables or statements such as those shown in Figure 1. If the terminal or control unit has no formatting logic, then many blank characters must be transmitted in order to format the display in a readable manner. Figure 1 would need more blanks than other characters. If its data could be sent in a compressed form, with formatting characters but no blanks, then the number of characters transmitted would be almost halved. At the terminal or control unit (or possibly the concentrator), the formatting characters would be used to expand the data into a neat display format. Here again we have a trade-off between distributed logic hardware and communication line costs.

Formatting logic saves buffer storage as well as transmission time. The Sanders 620 Data Display System terminal, for example, has 2000 character positions on its screen but only 1000 characters of buffer storage.

A R JENKINS 397 E 34 ST, APT 19B, NEW YORK 10017
073–2–037948

DATE	PARTICULARS	PAYMENTS	RECEIPTS	BALANCES
9.24	OPENING BALANCE			2338.66
9.26			956.60	3295.26
9.30	INTEREST CHARGES	63.85		3231.68
10.5	INVESTMENTS BOUGHT	1218.00		2013.68
10.5	ALREADY ADVISED		993.87	3007.55
10.15		265.00		2742.55
10.16		44.00		2698.55
10.16		2600.00		98.55
10.22		100.00		1.45
10.23	CHARGES	2.10		3.55

Figure 1. The blanks between fields need not be transmitted

Formatting characters distributed among the stored characters cause the display to be tabulated neatly on the 2000 available screen positions.

COMPACTION

The use of such editing is, in a sense, a form of data compaction. There are many other ways of minimizing the numbers of bits that need to be transmitted. Numeric data can be converted into binary form. Alphabetic data and most punctuation could be transmitted as five-bit characters. In fact, there is much to be said for using the five-bit Baudot code for compact data transmission. It uses "letters shift" and "figures shift" characters to switch the meaning of the bit combinations between letters and figures (plus special characters), somewhat like the shift key on a typewriter. It is interesting to reflect that a five-bit code with three shifts could carry all the data that most users would be likely to want to transmit. A minor modification of the Baudot code could achieve this, as shown in Figure 2. The character 00010 has been given the meaning "transparency shift." When this character is transmitted, the characters following it have the meanings shown in the rightmost column until a "letters shift" or "figures shift" character is sent. Some of these are control characters. A variety of control characters is used on modern terminals. The remainder are four-bit binary characters, bits 1, 2, 3, and 5 being used, with the fourth (always a 0) being ignored. Any eight-bit binary character can thus be sent in two of the five-bit characters, and so programs could be transmitted in this code.

A somewhat tidier five-bit, three-shift code could be devised if the Baudot code were ignored. The Baudot code is, however, the most commonly used telegraph code outside North America. The character 00010 is "carriage return" in the Baudot code and "transparency shift" in Figure 2. In order to communicate with a Baudot-code machine, a machine using the code in Figure 2 need only inhibit its "transparency shift" feature and substitute "carriage return."

How to Measure a Terminal's IQ

Code	Letters Shift	Figures Shift	Transparency Shift
	CCITT standard international telegraph alphabet No. 2		Control characters or four-bit binary combinations (T)
1 1 0 0 0	A	–	T
1 0 0 1 1	B	?	Control
0 1 1 1 0	C	:	Control
1 0 0 1 0	D	Who are you?	Control
1 0 0 0 0	E	3	T
1 0 1 1 0	F	Note 1	Control
0 1 0 1 1	G	Note 1	Control
0 0 1 0 1	H	Note 1	T
0 1 1 0 0	I	8	T
1 1 0 1 0	J	Bell	Control
1 1 1 1 0	K	(	Control
0 1 0 0 1	L	)	T
0 0 1 1 1	M	.	Control
0 0 1 1 0	N	,	Control
0 0 0 1 1	O	9	Control
0 1 1 0 1	P	0	T
1 1 1 0 1	Q	1	T
0 1 0 1 0	R	4	Control
1 0 1 0 0	S	'	T
0 0 0 0 1	T	5	T
1 1 1 0 0	U	7	T
0 1 1 1 1	V	=	Control
1 1 0 0 1	W	2	T
1 0 1 1 1	X	/	Control
1 0 1 0 1	Y	6	T
1 0 0 0 1	Z	+	T
0 0 0 0 0	Blank		T
0 0 1 0 0	Space		T
0 1 0 0 0	Line feed		T
1 1 1 1 1	Letters shift		
1 1 0 1 1	Figures shift		
0 0 0 1 0	Transparency shift		

Note 1: Not allocated internationally by CCITT; available to each country for internal use.

Figure 2. A compact code. A suggested modification of Baudot code to make it transmit "transparent" binary code and a wide variety of control characters

If a large number of machines using a five-bit, three-shift code came into use, the translation circuitry could be on one LSI chip and would not be expensive. The throughput with alphanumeric data on a given data channel would be almost 40 percent higher than with the ASCII code and almost 60 percent higher than with the eight-bit codes in use.

Verbal data could be compacted more tightly because of the redundancy in English language, but the cost of the compacting and uncompacting would be much higher. Suppose, for example, that a communication link transmits eight-bit characters, as when transmitting programs. Suppose that each character contains combinations of bits that represent the 200 or so most commonly used words. Suppose, also, that 32 of the 256 possible combinations indicate that this character alone does not give the word in question but that the next character is also needed. This gives 32 x 256 = 8192 additional words that may be encoded. Another combination of bits in the first character indicates that the word is spelled out in BCD. The machine receiving this string of data converts it into verbal English with a table look-up operation.

The cost in storage at the terminals is high. The number of characters used could easily be 64,000, but a voice line could be made to transmit as many as 40,000 words per minute. At the present cost of storage, it would probably not be economical to use such a method. During the next decade, however, schemes of this type may appear increasingly attractive.

The vocabulary required for many specific applications is small, and it may be necessary to store no more than, say, 256 words. Alternatively, such a scheme may generate not words but messages or phrases.

GENERATION OF MESSAGES OR PHRASES

It is highly desirable that a terminal respond to its operator in easily understood English. On many applications, the dialogue that is designed contains a number of often-repeated phrases. Such responses have been generated in most systems by the central computer. A considerable amount of verbiage then has to be sent down the communication lines. An alternative that involves less transmission time is to send only a coded reference to the responses and have the phrases, table formats, and so on generated by the terminal.

A character or group of characters may cause a stock sentence or phrase to be generated. For many applications, 128 phrases would be plenty. One character could be given the meaning: "The next character is a phrase-generating character." The next character would then generate one from a list of 128 (or 256) phrases.

It is probably on commercial systems or systems with programs for carrying out highly specialized functions that one is most likely to find a need for canned responses. However, even on general-purpose systems where the operator writes programs at the keyboard, or where the terminal functions like an elaborate desk calculator, a set of fixed responses makes the device easier to use. JOSS, the RAND Corporation time-sharing system that became famous as one of the

How to Measure a Terminal's IQ

earliest conversational computing facilities, had 40 canned responses that made it considerably easier to use.

However, whereas all of JOSS's canned responses could be stored in about 2000 characters, we are going to need a large storage if we are to store the verbiage from many specialized programs. The question arises: Would it be economically viable to have a disk file or other large storage at the terminal control unit? In some cases, it would. For most such purposes the disk could be considerably smaller and slower than the disks and drums used for data files. For such purposes, it might be worth making a very inexpensive disk unit, with a slow seek mechanism and disks the size of a 45-rpm phonograph record with no high-precision packing.

IMAGE STORAGE AT THE TERMINAL

In some terminals, entire images are stored and will be produced on the receipt of appropriate commands from the computer. A variety of different media may be used for image storage. An entire man-machine dialogue may be constructed with fixed-frame responses. In some cases, the availability of photographs or diagrams from an image-display system entirely changes the realm of what is possible in terminal dialogue.

GRAPHIC IMAGE FORMATION

Many graphic applications in which the computer generates line diagrams are in use. In order to transmit the diagram, it is broken into separate lines that are transmitted in a coded form. Scanning the entire diagram like a television picture is scanned would need far too much communication bandwidth. The assembly of the picture from the coded information can be done by the display control unit. This unit will normally be a computer that the operator uses.

The question then arises: How much processing will be done in the terminal computer, and how much in the remote larger machine? As in nongraphic applications, the control computer may be used for library storage and large files of data, as well as for processing that is too lengthy for the terminal computer. Because of the fast response time needed with graphics and the complex nature of some responses, the graphics terminals will almost always use a local computer—it may be an inexpensive mini-computer that relies on a remote central machine for some of the operations.

INPUT CHECKING OPERATIONS

When the terminal has its own mini-computer, there are many operations it may carry out before data is sent onward to the central machine. One is the syntax or consistency checking of input data. The terminal computer may collect the data piece-by-piece from the operator, check it, and send in one message to the central machine.

DIALOGUE

It is very important to make terminal "conversations" as easy to use and unconfusing as possible, but this intention, can result in a large amount of verbiage being transmitted. With multidrop lines, the result can be a substantial increase in the network cost.

One solution is to make the terminal control computer take over at least part of the dialogue. For example, in an airline-reservation system, the dialogue in which details of the passenger and his booking are recorded could be carried out by a peripheral computer. The peripheral computer would check the data for completeness, date-time continuity, and journey continuity and then transmit it to the central computer for filing. Earlier, when the availability of seats is being checked, the dialogue must be with the central computer because it keeps the necessary data.

LANGUAGE PREPROCESSING

Where a terminal operator is programming a distant time-shared computer, some preprocessing of the language he uses may be done in the terminal computer. This step is worthwhile in remote job-entry systems in which the program will not be compiled and executed immediately but will wait in a conventional job stream. The preprocessing will detect some of the programmer's errors in real time and thus improve his chances of obtaining the results he needs the first time.

As with all these "intelligent terminal" schemes, the question must be asked: Where is it more economical to do this function, at the terminal location or at the central computer?

COMPUTING OPERATIONS NOT NEEDING FILES

It is possible that a common use for data transmission in the future will be to provide a backup for "the mini-computer in the office." The mini-computer may have no files. It may have a small disk. It may be able to read and write tape cartridges stored on its user's shelves.

If its user performs routine calculations, it may not need any external assistance. Often, however, it will need to call on a remote program library and obtain information from distant data banks. It may need the assistance of a remote computer to compile large programs or to execute programs that need a large quantity of storage or high computing power.

How to Measure a Terminal's IQ

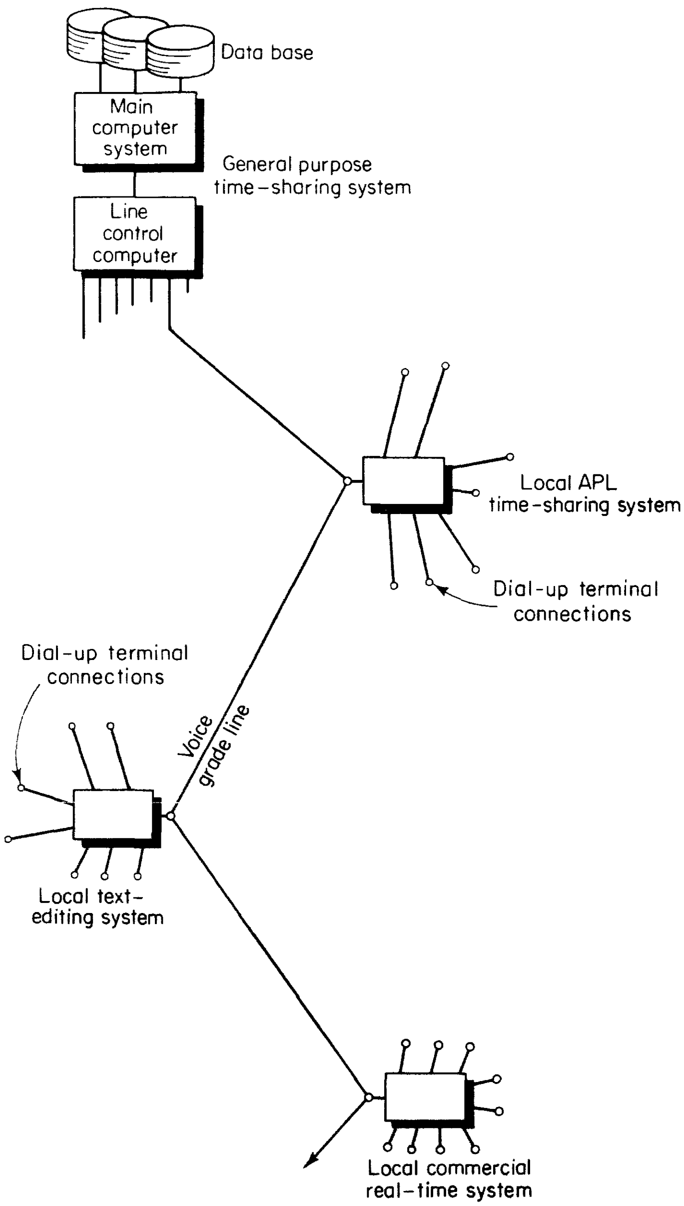


Figure 3. Small, local, special-purpose on-line systems linked to a large central system with wider capabilities and a large data base. The nodes of the network have become self-contained systems. If the path to the main computer is used only a small part of the day, the peripheral computers may dial up the main center rather than use a leased line

In some cases, the peripheral computers will be general-purpose machines that need to be programmed by their operators. In other cases, they will be machines for carrying on a highly specialized function for operators without programming skills. Secretaries may perhaps use such machines for text editing and document preparation, the storage of documents perhaps being on a central computer. Some of the countrywide banking systems in England use terminal computers that enable the bank clerks to carry out local accounting operations and that communicate with a central computer that has all the customer files.

TERMINAL COMPUTERS WITH FILES

In some cases, a small amount of file storage may be used on the terminal computers, so that records not needed centrally can be stored.

SPECIALIZED SUBSIDIARY PROCESSORS

In the rapidly evolving technology of time-sharing systems, some schools of thought favor general-purpose systems, as well as highly specialized systems and software. Some of the smaller and highly specialized systems have proven very efficient and have given an impressively fast response time. It seems probable that we will see more of them and that their cost per user will remain competitive with the larger, more general systems.

A corporation may have a number of such computers in its various locations, each capable of meeting a local need. Small, specialized systems are relatively easy to implement. These systems may be designed so as to pass on transactions that they cannot handle to a larger data processing system. Typical of such systems are those providing on-line computation with languages, such as APL and BASIC, those providing on-line computation with languages, such as APL and BASIC, those providing secretaries with text-editing and letter-typing facilities, and small commercial enquiry and data-entry systems. Figure 3 shows a possible arrangement with a line from the main data processing center wandering around several small real-time systems that are self-contained for their own functions. General-purpose terminals are used. The computer at the center polls the peripheral computers periodically to see whether they have any work for the main computer.

The network, in fact, may not have a distinguishable main computer. An organization may operate several computer centers primarily dedicated to different types of work but capable of interchanging some jobs or transactions. The centers may have different files of information available to them and different programs. Transactions are routed to the appropriate location in this complex.

In this report we have talked about "intelligence" in the terminal, or terminal control unit. This is not the only place where intelligence can reside. It may be built into nodes of the communication network, sometimes referred to as concentrators. Line-control computers, remote both from the central computer and from the terminal, may also carry out some of the functions we have discussed.□

Providing User-Oriented System Interfaces Planning

Problem:

What type of terminal provides the best interface between a computer user and a computer system? The answer depends upon who the user is and how that user wants to make use of the system. There are a multitude of answers ranging from direct input/output via CRT display to interposing human "information specialist" between the computer system and the user.

This report is a practical, down to earth discussion of man-machine dialog with emphasis on serving the user. The reader may be surprised at the value placed by the author upon human intervention in the retrieval of MIS information, and the dangers in some cases of direct user-machine dialog. If you ever thought that there would be an "ultimate terminal", read this report first.

Solution:

Terminal-based systems are often fully automated in that in normal operation there is no intervention by the staff associated with the computer. Many systems analysts and potential users of systems regard this as the proper state of affairs—and often it is. It would be wrong, however, to assume that all systems should be devoid of human intervention. On some, the best approach is a continuing combination of computer and human processing. Men and computers have entirely different talents. They should be linked together into the most effective combination for providing a service or doing a job.

The human expertise that forms part of an operational man-machine system will sometimes be gathered together in one place. There are numerous examples of such rooms on different systems and they have widely different purposes. They vary greatly in size and complexity, ranging from two men on teletype machines, to the Project Apollo control room at NASA's Manned Space Center. They are given various names such as "Management Control Room," "On-Line Service Center," and "War Room." They vary widely in function but in all cases have a group of men assisting the real-time operation of the system. Such rooms are likely to become an increasingly important part of systems in which men communicate directly with computers. In this report, we will examine examples of their use.

The file owners, who are responsible for the integrity, the accuracy, or the security of different files, may be in an information control room. They have a detailed knowledge of the data base and its control procedures. When errors are found in its data, they are responsible for having them corrected. This would usually not be a full time job, although on some existing systems it is.

Based upon *Design of Man-Computer Dialogues* by James Martin, Chpts. 25 & 26, pp. 425-445. © 1973 by Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Providing User-Oriented System Interfaces Planning

HUMAN PROCESSING

On some systems that are used for making operational decisions, not all of the decisions should be left entirely to the computer. There are some decisions for which it has insufficient information or for which it cannot be programmed adequately. Indeed, the key to effective operation is the degree to which man and machines can cooperate.

A good example of a system making operational decisions is an airline reservation system. On some airlines the decision of whether or not to give a potential passenger a seat on a given flight is made entirely by the computer. The airline agent at her terminal receives a response instructing her whether or not to sell that seat. This sounds like a simple enough decision: is there a seat or is there not? There are, however, complications. First, many seats are canceled. On some airlines, the average seat is sold between two and three times to different customers, because the cancellation rate is so high. Many of these cancellations are made the day before take off. Since the plane should take off as nearly full as possible and most cancellations occur late in the booking period, it is advantageous to overbook early in that period. If the number of bookings is never allowed to be greater than the number of seats on the plane, the late cancellations will result in some planes taking off with empty seats when they could have been full. The danger with overbooking, however, is that there is a possibility that the plane will remain overbooked. Then infuriated passengers will have to be switched to other planes or, worse, told that they cannot fly that day if the schedule has no other flights.

If overbooking is used (as it is), should it be left to the computer, or would an experienced flight controller make better judgments on the number of overbooked passengers that can be accepted on different flights at different times? If the latter is the case, then a three-way man-machine dialogue is needed in which the computer occasionally asks for assistance in dealing with a transaction from a booking agent's terminal. In practice, for most flights on most airlines, the best solution appears to be for the computer to overbook, with an overbooking limit that diminishes as the take off time approaches. The reservation is not confirmed while it is for an overbooked seat.

A more difficult complication attends multileg flights. Suppose that a flight travels around the world and stops at ten cities on the way: for example, London, Rome, Tel Aviv, New Delhi, and so on. Suppose now, that many passengers want seats for one particular flight from Rome to Tel Aviv. If they are accepted early in the booking period, the number of passengers that can be accepted for higher revenue journeys, such as London to Australia or New York to Tel Aviv, is lowered. The sales manager at Rome has been given the job of maximizing his seat sales, but if he sells too many seats to Tel Aviv, he will lower the overall revenue from the world-wide flight. The computer system must therefore make a probability calculation to decide whether selling the shorter distance flights is the best strategy or not. The calculation, however, turns out to be exceedingly complex. There are many possible combinations of cities on a flight and a high diversity of traffic patterns are involved.

On one airline an elaborate piece of operations research resulted in a program for deciding what booking limits should be set on multileg flights. The results of the program were then compared with the results achieved in practice by the human flight controllers still employed. To the secret delight of many people not in the operations research team, the flight controllers did better. The reason was that the traffic patterns were affected by highly varying factors. It takes many years of human experience (some would say "feel" or even "intuition") to spot the causes of traffic buildups and to predict when they will occur. Here was a case where the human decision-maker proved superior to the computer at certain points in what appears to be a simple operation—booking an airline seat.

The solution here was to retain the experienced flight controllers in the system and allow them to judge what booking levels should be set as the flight fills up. There are many other such situations in industry and elsewhere.

Human assistance can be built into the system in one of two ways. In the first way, any transaction that falls within a range deemed to need attention is intercepted and referred to a specialist at a terminal. The specialist often uses the same real-time system in making his decision about the transaction. In the second way, the specialist inspects the situation at intervals and sets figures in the system that enable it to take automatic action on the transactions. In the case of multileg flights, he periodically examines the traffic buildup on each flight and sets appropriate booking limits. He does not have to deal with each transaction individually and thus quickens the response time at the agent's terminals.

The specialists in an informaiton control room may have their own sources of data for decision making. These may be maintained in the same system. Each of the staff members in the room in Situation 1 has two terminals. One is a low-speed printer that gives unsolicited messages from the system, and the other is an inquiry terminal from which files can be inspected and changes made. With this combination, an urgent

Providing User-Oriented System Interfaces Planning

Situation Number 1—Unidentified control room

Site Description—Medium size crowded office

Equipment in use—Telephones, typewriter terminals, RO printers.

Operator activity—Operators are speaking on the telephones and simultaneously typing on the typewriter terminals. Unsolicited messages about errors or exception conditions can be simultaneously printed by the RO printers. Other personnel are standing, conversing while examining a large looseleaf notebook lying on a table.

Environmental description—Busy control center populated by shirt-sleeved men.

Comment—The typewriter terminals and RO printers are IBM Selectric-based, however it is not obvious that these are connected to a computer. They could be free standing typewriters.

Situation Number 2—New York City Police Vehicle Dispatching Center

Site description—Large crowded office.

Equipment in use—Telephones, time stamps, one or two CRT terminals.

Operator activity—Telephone (radio) conversations with police vehicle crews and other officers in the room, manually writing information.

Environmental description—Extremely busy. Many simultaneous conversations between operators on the phones and with other people in the room.

Comment—There is no evidence from the photo that any of these officers are carrying on a dialogue with a computer as stated in the text, nor are they able to observe from a CRT display, the status of any vehicle.

message from the system does not interrupt the specialist in the middle of a terminal dialogue.

THE OPERATIONS ROOM

It is often desirable to have one room from which a set of operations is directed or operational decisions made. Management, wanting to know what is happening or to change it, contacts this operations room or situation room. The information control room was designed primarily to be the location where errors are dealt with, but it may be much more than that. The radio and telephone links from the control room to the shop floor expediters are precisely what is needed, in addition to the data collection system, to provide one group of staff with an overall knowledge of the situation on the shop floor. This group, in addition to scheduling what takes place on the shop floor, forms a segment of a management information system for providing management with whatever shop information they request.

Situation 2 describes the operations room from which all police vehicles are dispatched in New York City. The dispatchers talk by radio to the crews of the vehicles and maintain details of the positions and assignments of the vehicles by means of the visual display terminals. The same terminals inform the dispatchers of all of the current emergencies and situations that need servicing. Details of major and special emergencies are broadcast by loudspeaker to the dispatchers. Police supervisors can monitor particular situations by walking around the room. The computer and dispatching crew thus work together to monitor and control the police activities in the city. In a similar way, the functioning of a factory, hospital and its ambulances, airport, or other organization can be monitored and controlled by a combined man-machine operation.

The operations room may have all manner of maps and wall charts like an RAF "Ops Room" in a World War II movie. Specially constructed displays with lights, counters and other devices operated by the computer can be used. There is much scope for development in computer-generated displays. On the other hand, computer printouts hung on the wall, or available for reference, serve very useful functions. Gantt charts and other changeable charts built with plastic or magnets can be updated periodically by the operations room staff.

When an operations room is first set up there is much experimenting with displays in an attempt to provide the most useful information. Initially, therefore, the displays should not be of a permanent nature. The evolution of requirements in the operations room will lead to an evolution of the computer system to provide it. However, between the computer system and the wall displays, there is often a human process of refining, clarifying and colorfully presenting the information.

Situation 3 describes a control room in which the information is displayed by means of conventional instrumentation. In Situation 4 there is no conventional instrumentation. Almost all of the information used by the staff participating in the control of a space mission is obtained by dialogue with the computer system. This form of information access is

Providing User-Oriented System Interfaces Planning

Situation Number 3—Nuclear Power Station Control Room

Site description—Large, clean, sparsley populated operations control room.

Equipment in use—Several semi-circular consoles containing meters, switches and multiple telephones. Around the walls are hundreds of meters, and chart recorders, plus clocks showing different times.

Operator activity—One operator at each console observing the instrumentation. Another walking about observing instrumentation and making notes on a clipboard.

Environmental description—Well controlled nerve center for a large and technically complex operation.

Comment—This is an example of man controlling an operation by reacting to conventional instrumentation rather than by interacting with a computer.

far more flexible and can provide a much more powerful mechanism for control.

A MANAGEMENT INFORMATION CENTER

Sometimes the purpose of an information room is not to direct a set of operations but rather to provide management with knowledge about what is going on. Management's information needs are very diverse, and the information room will often act as a clearing house. Specialists in the room will know how to obtain information which managers need, and the managers will telephone this room rather than attempt to use terminals themselves.

In some cases when a manager has a terminal in his office, it is used largely for delivering information, not for seeking it. The manager telephones the information room, and the specialist there accesses the relevent computer. He switches the required output to the terminal in the manager's office, where it is displayed or printed. Here again is a three-way conversation: manager to information room, information specialist to computer systems, computer to terminal in manager's office. Three-way conversations will be increasingly used to deal with situations in which the user cannot himself communicate directly with the computer systems for one reason or another.

The data requested by top management relates to all parts of the organization, and so telecommunication facilities link the information center to far-flung locations. These are used for telephone, facsimile, and data transmission. They may be public dial lines or the organization's internal telephone network. If a line to any particular location is used sufficiently frequently, say, more than two hours per day, it should be a private leased line. The break-even cost between dial and private lines depends upon the tariffs. Sometimes the information sources are service bureaus or time-sharing systems, and sometimes they are the organization's own computers.

Information requests are often made for data not stored in computers. The information center may act as a clearing house for these also by maintaining files of documents, letters, magazines, manuals, microfiche, and microfilm. Some of the information requests may require the use of a library. A video link from the information center to a company library may be used so that the services of the librarian can be enlisted when needed.

AN INPUT INFORMATION CENTER

Just as persons wishing to obtain information from a computer can telephone the staff in an information center, so persons wishing to enter information may use a similar process. Management might telephone and say, "make sure such-and-such an item is completed by next Monday." The staff enters this order into the data for computing work schedules.

In some cases a high proportion of the input data is obtained from persons telephoning an information

Situation Number 4—NASA Mission Control Center, Manned Space Headquarters, Houston

Site description—Large, heavily populated room containing several rows of CRT consoles all facing a large computer-driven mercator projected map of the earth which covers an entire wall.

Operator activity—All operators observe the wall map which shows the radar tracking stations around the globe, the orbital position on a real-time basis of the spacecraft under control, and the paths of just-completed orbits. Individual operators most of whom are wearing telephone headsets are monitoring via their CRT consoles specific phenomena about the spacecraft such as outside and inside temperature, cabin pressure, fuel supply, various back up systems, etc. in addition to the mission control directors' maintaining voice contact with the astronauts.

Environmental description—Extremely complex, well coordinated activity with numerous back-up safeguards.

Comment—This is probably the most easily recognizable and successful examples of complex man-computer dialogue on earth today. However it is unique in that the numerous control provisions which the center provides could never be economically justified for a commercial enterprise.

center. This is so, for example, with airline reservations. Situation 5 portrays a center at BEA (British European Airways), to which persons telephone to book or cancel seats. The terminal operator talks to the passenger, gives him the facts he wants about flight schedules and seat availability, and then enters details of his booking into the computer. In the future, many types of firms will handle requests from customers in this manner.

Situation Number 5—British Airways Reservation Office, London

Site Description—Enormous room with row after row of desks containing terminals and operators.

Equipment in use—CRT terminals, telephone headsets.

Operator activity—Operator agents speak on the telephone with customers wanting to make reservations, modifications, inquiries or confirmation of flights. The agent carries on a dialogue with a large computer via the CRT terminals. The caller may, or may not know that the agent is interacting with a computer.

Environmental description—Extremely busy, but well controlled service operation.

Comment—This activity operates 24 hours per day, but with reduced staff after normal business hours.

OPERATOR ASSISTANCE CENTER

With the more elaborate terminal applications, users sometimes require help. This is especially so when they are "casual" users, who only occasionally make use of the terminal. It is also true when there is a wide diversity of applications that can be tackled from a terminal.

Help can come from the system itself, which might be switched into an instructional mode. Alternatively, the bewildered terminal user can telephone an assistance center where an operator can monitor what he does at his terminal and instruct him in the correct procedure.

The value of such human assistance should not be underrated. It is only too easy for the inexperienced terminal user to become confused at the terminal and feel too foolish to ask for help from his colleagues. Anonymous assistance and monitoring from the assistance center is greatly appreciated and may avoid his rejecting the system—as has often happened.

PUBLIC USES OF COMPUTERS

The building of human assistance into certain computer systems is likely to become a vitally important aspect of their design. It is not only in industry where such assistance will be needed. In the society we are building, people in all walks of life will need to obtain information from computer terminals. In many cases, they will need human help in so doing.

Since the earliest days of the computer, there has been a trend from military usage of information systems to industrial usage. SAGE used real-time processing, visual display screens, and light pens; ten years later these features were common in industrial systems. Now the technology of military command-and-control systems can be found in industry. Data transmission networks are spanning commercial organizations. Analogues in industrial and civil situations are being found for the simulations used in "war games."

There may be a second stage to this trend. The schemes that were first used in the military at enormous cost reach industry when the cost falls substantially, and then in some cases reach the general public when the cost further drops. The data banks containing information of value to industry today will hold information sought by the public tomorrow. The public is beginning to use computers in helping to find a house, book theater tickets, provide stock information, give advice on optimizing tax returns, and so on. The terminals that spread from the military to industry will eventually spread from industry to the home. Just as the industrial manager needs help in using his terminal, so, to a greater extent, will the public. The lessons now being learned about the man-machine interface will be essential to home use of terminals. They are the key to their acceptance.

The information room in industry needs human skills as well as machines. Similarly in systems serving the public it would be a mistake to attempt to automate everything. The public will need not only "directory assistance" but many forms of more elaborate assistance in using their terminals. As in industry, a human operator will often generate an appropriate display and switch it on to the screen in the inquirer.

In the early days of the telephone, the public received much more help from the staff at the telephone exchange than they do today. They could ask their "friendly operator" for the time, the weather forecast, how to get a nurse, or what film was at the local moving-picture house. Now they usually get a recorded voice with a tone of monotonous indifference more likely to incense than soothe. Some countries still have friendly human help available by telephone. France, whose telephone system is notoriously underdeveloped, has a service called "Q.E.D." that enables the public to ask questions about almost anything.

Providing User-Oriented System Interfaces Planning

When the general public first uses computer terminals, it will be desirable to bring back the "friendly human operator" where possible. As the average person uses the terminal for shopping, attempting to book a journey or a vacation, balancing his bank or credit account, searching for literature on a particular subject, performing calculations, or carrying out any number of other functions, he will occasionally need assistance. The terminals should have a "HELP" button, and the computer rooms that provide such services will need to have a staff for this. Information rooms with skilled personnel like those for management assistance will be needed in many other areas of terminal usage. Not least they will be needed for helping obtain the switched connections to the requisite computers. Let us end this report then, with a plea for the skilled use of people in computer systems. The shift in employment from the manufacturing industries to the service industries will continue, and one segment that will need people will be that of the rapidly growing information services.

Total automation is not an end in itself.

Today many firms are trying to build some form of "management information system." What is understood by this term differs greatly from one organization to another. This is not surprising because of the nature of "managers" and their needs differ greatly. Some firms give the initials "MIS" to a simple inquiry system. Others define it in such all-embracing terms that it can never be more than an unattainable dream. Some have purged the term from their vocabulary in a change of policial fashion. Management information systems, real or hypothetical, have one property in common: when a manager asks for certain types of information he must receive an answer quickly.

To answer a manager's questions, terminals are used but not necessarily by the manager himself. The system designer must decide whether or not the manager should have his own terminal, and if he has his own terminal, should he work it himself or should he have an assistant to work it? Sometimes the terminals are employed only by management staff, and sometimes management staff have assistants to work the terminals. The debate as to which managers should use terminals will continue for many years hence. Some managers who once refused to have terminals in their office now have one, and others who had a terminal installed subsequently rejected it. A major factor in these decisions (apart from the cost of the terminals) is whether the necessary man-machine dialogue can be carried on easily by the manager in question. Some levels of management, usually lower levels, can be trained to use a fairly complicated dialogue. However, the majority of managers today and often their staff also can only be expected to interact with a terminal if it involves a very simple dialogue. This may change with new generations of management, but for the present it must be a major factor in the system design. If managers, especially older ones, are asked to do something too complicated with a terminal, they will become frustrated, annoyed, critical of the system, and will probably not use the terminal—certainly not to its full extent. Many schemes for giving terminals to management or their staff have foundered on this rock.

The managements in different organizations differ very widely in their style of operating. In many organizations it would be unthinkable for high management to use computer terminals directly, and the discussion in this chapter applies entirely to their staff rather than to the managers themselves. The reason for this distinction has nothing to do with the man-machine interface but with the customs and methods within the organization.

In practice, the situation is further complicated in many firms that make advanced use of computers by the fact that there is more than one source of management information. The different sources have evolved separately, on separate computers, and they often use quite different dialogue structures. Sometimes the different systems are dialable from the same terminal, sometimes not. The informed manager or staff assistant often needs access to information from separate and incompatible systems.

THREE OPTIONS

The designer of management's information sources has three options open to him. First, he can provide management with terminals and design a dialogue structure that is suitably simple.

The set of user characteristics that apply to most of this category of user are:

- high intelligence
- requires a high information bandwidth in the dialogue
- too busy for a training course
- will not remember mnemonics
- highly impatient
- nonrugged (will reject the dialogue if confused by it or put off by unintelligible error messages)
- requires worthwhile results

Providing User-Oriented System Interfaces Planning

In many cases, a major obstacle to managers and their staff using their own terminals is the diversity of information sources they should have access to, and the fact that these sources differ widely in their dialogue procedures. Some authorities hope that greater standardization will be achieved, but it seems more likely that the proliferation will continue in the years immediately ahead as the sources of terminal information available to management become more numerous. This factor is further compounded by the increasingly complex nature of the data banks in use. Managers usually do not know the content of the data banks, nor how to search them.

DATA SECRETARIES

The second of the three options is to provide management or their staff with specialized assistants. Firms can establish corps of trained personnel whom we will refer to as data secretaries. Their function is to inform and assist management in the use of the computers. Some managers can have their own personal data secretary; others can summon one from a pool when needed. The data secretaries may show the managers or their staff how to carry out simple functions at their terminals and perform the more complex ones for them. They should be trained to understand that managers differ widely in their capability or willingness to interact directly with the computer system. They explain to managers the operations possible with the system and encourage and guide them in its use. The data secretaries assist in the filing and retrieval of information just as a clerical secretary does. They obtain answers to those questions that can be answered with the computers, and they attempt to inform the management of the types of questions that should be asked.

If corporate models are available with graphics output, the data secretaries may arrange for a manager to use the light pen, and ensure that he is on a right track.

A communication gap often develops between management and the data-processing team because of their entirely different ways of thinking. Many managers are naturally somewhat frightened of the computers because they do not understand the technology. The data secretaries should attempt to bridge this gap. They require a high level of diplomacy. They need detailed training in how to use the systems, not how to program or design the systems. They must be very sensitive to management's needs and should probably be part of the management's staff, not the data processing staff. This pliable human link can greatly enhance the use of computers.

Most systems currently providing information to management are fairly simple. The questions that can be asked of them are fairly simple—mostly direct enquiries of records in the files, and questions involving searches through the files. Where this is the case, the job of the data secretary is not complex. A nongraduate would generally suffice. However, in the decade ahead information systems are going to become increasingly more powerful and sophisticated. The questions will require much more elaborate tools, such as simulation, linear programming, mathematical models of corporations, and other operations research techniques. Management will ask questions about optimization and "what if . . ." questions: "What effect will it have on production if I accept this order?" or "What would be the effect on the cash flow if this project slipped by one year?" Assisting management to interact with its computer system will become an increasingly more professional job. Perhaps a new profession, or new level of professional staff, will arise, trained to understand the use of computers and their complex programs, and to assist management to make the best use of them.

MANAGEMENT INFORMATION CENTER

The third way to solve the manager-machine interface problem uses an information room. Here it is designed to provide a source of expertise that will enable managers to obtain the information they want from their computers.

From their offices, managers may contact the information room by telephone, by closed-circuit television, by the use of terminals, or in some cases by Picturephone. The staff in the information room listen to their requests for information and try to provide the answers as quickly as possible. They convey the answers verbally, by messenger, by closed-circuit television, or by switching to a terminal in the manager's office where the answers are printed or displayed.

In some organizations, a hierarchy of information rooms will probably develop—some simple and serving limited needs; others highly complex. A factory may have one information room relating to the flow of work through the shop floor. A regional sales headquarters may have one relating to its customer and order situations. Subsidiary companies will probably have their own, separate from their parent company. A head office location may have a group of skilled operations research staff capable of dealing with more complex types of questions.

The manager of the future may have a "hot line" to his local information room. He picks up a red telephone on his desk and is immediately switched to personnel there who know where to find the answers to various types of questions. If possible, they will answer his question verbally or will display relevant information on the screen in his office or prepare a printout for him. Sometimes they will route the query to another information room, that is more specialized or perhaps in a distant part of the organization.

Board rooms and conference rooms will probably become equipped with screens and perhaps printing terminals. As the discussions proceed, information can be obtained quickly from the information room staff, who will flash appropriate displays on the board room screens.

LEVELS OF MANAGEMENT

It is important to differentiate between levels of management, because the way they are likely to use systems is very different. Systems in which managers have used terminals have tended to serve the needs of lower, rather than higher, management. Operational management has specific, well-defined requirements. Programs can be written and files set up to meet these requirements. The requirements of higher level management are less easy to anticipate, much more varying in nature, and less easy to fulfill. A shop foreman can fairly easily be given a terminal that provides details about the work schedule in his shop. The sales manager can be given one that provides details about products, orders, and customers. It is not difficult to supply a real-estate broker with a teleprinter and microfilm viewer, which together permit him to search a wide area to satisfy his customer needs. However, one has to look far and wide to find strategic, rather than operational, management making valuable use of terminals in their offices.

Apart from the reason that their needs are more complex and less structured, higher levels of management generally have little inclination to learn how to use a terminal. Therefore, the best way to organize communication between higher management and the computer usually is to employ a specialized third party in some way or other.

A SYSTEM FOR TOP MANAGEMENT

IBM corporate headquarters in Armonk has a successfully operating information system for top management—or rather we should say strategic management, because as well as being a tool of top-line executives it is greatly used by staff executives concerned with corporate strategy but who are far from the "top." The information requirements of its users are highly diverse as one would expect for strategic planning, and they are often highly complex. The system is not concerned with day-by-day operations. There are other quite separate systems to meet these needs in the areas of sales, manufacturing, engineering, product development, and research.

The realization came early in the development that managers at this level are generally not going to work their own terminals. They have never learned to type and regard that as a secretary's job. They are very short of time, and much time is needed to learn the terminal techniques. Furthermore, men of high prestige do not like a machine spitting back error messages at them. This would be more than unwelcome if other executives, especially from another corporation, were standing over their shoulder at the terminal.

There are exceptions, of course. One vice-president had a terminal at which he wrote elaborate programs in an advanced language (APL). The personnel department at Armonk makes elaborate use of their own terminals for obtaining personnel statistics. One president of a different corporation has an entire IBM System 3 computer in his office. However, the corporate information system cannot be designed for the exceptions. And even the vice-president who programmed in APL would not know how to access most of the information the system can provide.

The solution used, and this may be the right solution for other corporations—was to set up an information center from which management's requests for information could be quickly answered. The information center at Armonk is described in Situation 6. In it is a staff of information specialists who have the skills needed for locating, retrieving, and, if necessary, processing the information required.

Situation Number 6—IBM's Corporate Information Center, Armonk

Site description—Office size or layout not evident.

Equipment in use—Closed circuit TV camera and console monitor, graphics display terminal with light pen, large 13 x 15 matrix routing switch, microfiche viewer, 16 mm sound motion picture projector, 35 mm slide projector, library of looseleaf manuals, telephones.

Operator activity—Operators are information specialists who accept requests over the telephone for information from executives and respond by presenting it to the requestor over closed circuit television using the most appropriate medium available.

Environmental Description—Friendly, efficient, service activity.

Comment—It is interesting to note that in the executive offices of *the* largest computer manufacturer, who obviously *could* implement any type of man-machine dialogue imaginable, both computerized and non-computerized storage and information display as well as a human interface are employed.

These men are professional and have taken some time to learn to do their job well. They have detailed knowledge of the available data banks and associated programs. They use terminals remotely connected to a variety of different computer systems throughout the corporation and are familiar with the pertinent computer techniques.

When managers have questions they phone the information center and talk to their "friendly information specialist." He gives whatever help he can, advising management of the types of questions that can be answered. Often he is able to obtain the answer to a question very quickly. Sometimes he needs to initiate a search through computer files, run statistical programs such as linear regression analyses, send cables to other locations, research through documents, or even write a program.

When the information specialist has obtained an answer, he can give it to the requester in one of several ways. First, he can telephone it. Second, he can send a messenger to the requester's office with a printout. Third, some managers have screens in their offices, and by means of a video console the specialist can switch an image of the information onto the manager's screen. A fourth way would be to switch the data to a printing terminal in the manager's office.

The information center does not have a computer in it. It has only terminals through which information sources can be contacted and calculations performed. Security is of great importance in accessing data banks containing confidential or personal information. Each terminal user has a security code and can only obtain information that he has been authorized to have from the data banks.

The information specialist forms a human interface between requests for information and sources of information. As information systems and information processing becomes more complexed the need for this human interface will grow.

One lesson learned from the difficulties that have been encountered in setting up management information systems is that total automation does not work. Intelligent and specialized human beings are an essential part of the operational system. This comment applies to other types of systems also. In corporations the style of management differs to such an extent that what works well in one corporation does not succeed in another. It is very difficult for the systems analyst to change management style; rather he must adapt himself to it.

We thus return to our earlier theme that the best approach to some systems combines computer and human processing. The widely different talents of men and computers should be linked together in whatever is the most effective manner.

COMMUNICATION BETWEEN DIFFERENT MANAGERS

In some cases, a major advantage to management employing computer terminals has been the greater degree of precision in the communication between different managers. The first application of data processing has revealed the lack of precision or control in many commercial operations. For example, application of computers to stock control revealed how badly the stock was kept before. Attempts to model decision-making processes have revealed how poorly the important parameters were understood. Where managers of different functions are brought together to use a common computer dialogue, there is less change of imprecision.

In a system at Westinghouse[1], a graphics terminal is used for production scheduling based on sales forecasts of washing machines. This is a complex operation because Westinghouse makes over one hundred models, all available in several colors. Once a month, the production and marketing managers travel to Pittsburgh to work together on the display console. The marketing managers evaluate market forecasts and assist the production managers in working out the production schedule. The use of the terminal permits more options to be explored than were possible before. Before, according to Reference 1, a "seat of the pants" approach was necessary. Now the two groups of managers can communicate with precision. The managers involved, once experienced with the technique "wouldn't want to do their scheduling any other way"[1].

REFERENCE

[1]William E. Workman: *Which Color Washer Will They Choose?* Computer Decision, Dec. 1969. □

Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path

Problem:

Data produced by a terminal or a computer is generally connected directly into a modulator/demodulator (modem). Once the data passes through the modem, it is no longer linked to a discrete set of data channel problems, but it enters a totally new environment of network problems. Before we leave this narrow universe of the terminal/computer-modem data path, we offer this report to summarize a variety of "black boxes" that supply specific solutions to specific problems such as asynchronous-to-synchronous data conversion, error control, and encrytion.

Solution:

In this report, several gadgets are described that are rather ingenious solutions to some common data communications system problems. All the devices are connected between a terminal or a computer and its modem and operate on the digital data stream to provide compatibility, to improve throughput, or to enhance security.

ASYNCHRONOUS,—TO,—SYNCHRONOUS DATA CONVERTERS

One of the basic tenets of data communications is that long-distance transmission of asynchronous data over voice-grade lines is limited to data rates of 1800 baud or less. However, there are many asynchronous terminals that operate at speeds of 2400, 4800, and 9600 baud. Most such terminals are physically close to their computers, but quite often a need arises to use them remotely. A Hobson's choice then must be made whether to cut speed to 1800 or 1200 baud or to replace the terminals with more expensive synchronous terminals that can run faster but that require new front-end hardware and software support.

"That Old Black Box Magic" from *Basic Techniques in Data Communications* by Ralph Glasgal. © 1977 Artech House. Reprinted by permission.

An inexpensive solution to this problem is the asynchronous-to-synchronous converter. Such a device receives the characters asynchronously by looking for the start bit and stores the characters in a buffer so that the modem can clock them out, usually including the start and stop bits, and the clock. The same device can be used to couple asynchronous terminals to the DDS network.

The data as finally transmitted by the modem or DSU is isochronous, synchronous, and asynchronous all at the same time. Thus, the synchronous data received by the synchronous modem or DSU can go directly into the asynchronous terminal without conversion. There are, though, some precautions that must be observed when using such converters. If the asynchronous terminal is slower than the synchronous modem (as when an 1800 baud terminal is put on a 2400 bps DDS line), the converter buffer empties faster than it can be filled and a buffer-full condition never occurs. If, however, a 2400-baud terminal is put on a 2400 bps line, data may accumulate in the buffer if the asynchronous clock is slightly faster than the synchronous clock. Eventually the buffer will overflow, and data will be lost. To prevent this occurrence, the asynchronous data should be generated with periodic pauses to allow the buffer to empty. For example, if the clock speed difference is

Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path

.05%, one character will be accumulated for every 2000 transmitted. If a four-character buffer is provided, 8000 characters can be transmitted head to tail without pause before an error will occur. In normal systems, buffer overflow is seldom a problem. A Request to Send stretching circuit prevents the lowering of Request to Send to the modem before the buffer is emptied.

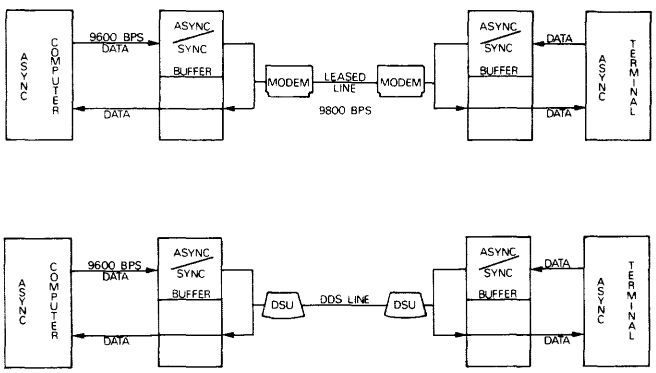


Figure 1. Asynchronous data transmission at rates up to 9600 bps are possible by using a simple conversion unit on synchronous lines

ERROR CONTROL

One disadvantage of using high-speed asynchronous terminals at remote locations (such as in Figure 1) is that there is usually no error control. Since the system probably was designed with only local terminals in mind, transmission error control was not needed, and provisions were not made for it. The most common way of getting around this problem (getting a bisync or SDLC terminal, new front end, and software support) is expensive and time consuming. The alternative is quite simple in concept. Use an error control black box between the terminal and modem, as shown in Figure 2. In this manner, most of the advantages of the synchronous block transmission formats can be obtained at a fraction of the cost and without modifications to terminal or software.

There are basically two types of error control—retransmission of data received in error (ARQ) and correction of errors by the receiver (Forward Error Correction, or FEC). The ARQ type of error control unit takes data from the terminal, formats it into blocks, and adds a header and check bits at the beginning and end of each block, respectively. A relatively small number of check bits allows detection of virtually all types of errors in a block. When a block is found to be in error, a request is made for retransmission of the block. Data throughput is determined primarily by the number of blocks that must be retransmitted. Additional time may be lost in formats such as bisync that require that each block be acknowledged as correct before the next block is sent. The new Synchronous Data Link Control (SDLC) eliminates the time lost in waiting for the acknowledgments by sending acknowledgments while subsequent blocks are sent; however, SDLC cannot make up for the loss of throughput due to noisy lines, nor can any other ARQ system. Thus, ARQ, if applied to a 9600 bps modem link operating with a burst error rate of one burst of errors in every 10,000 bits and having blocks of 3000 bits, would result in a throughput of only 5,568 bps, because some 42% of the blocks would need to be retransmitted. At this rate, a 4800 bps modem might as well be used to save some money and eliminate many of the errors.

Thus ARQ or bisync is not always the best answer to error control for modems, multiplexers, or terminals that do not already include error control; forward error control has some real advantages in such situations. By adding some 140 check bits to a block of 4000 bits, it is possible not only to detect but also to correct for burst errors up to 50 bits wide. On real lines, well over 90% of the errors that occur can be corrected.

In a real system, though, 100% correction is desired. The answer is to request the retransmission of the few uncorrected blocks and to combine FEC and ARQ techniques so as to optimize the throughput. In the combined ARQ-FEC units available today, it is possible to improve the error rate by a factor of 10^6 at an overhead cost of only 15%. Thus, a 9600 bps modem on a 10^{-4} line could deliver 8160 bps with an error rate of 10^{-10}. The 1,440 overhead bits actually need not be lost if it is considered that, with the exception of a few sync characters, the terminal in theory need not waste any bits on protocol or check characters of its own. The throughput can be enhanced further if the error control units use a full-duplex method similar to SDLC to acknowledge correct blocks or to request retransmission of wrong

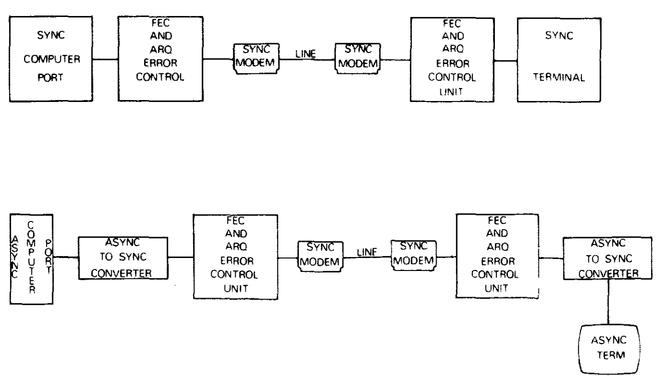


Figure 2. Error control units improve throughput and may eliminate the basic need for elaborate error control protocols such as sdlc or bisync

Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path

blocks. Note that on occasion, as in any ARQ system, if too many retransmissions occur in a given period of time, the buffers of the transmitting unit may fill up. Therefore, there must be a mechanism to halt temporarily the flow of data from the terminal or computer. For asynchronous terminals, a lowering of Clear to Send may be adequate; for synchronous terminals, an interruption of the send clock may be permissible.

FEC can be used to advantage with TDMs. By using smaller block sizes, delay in the error control system can be held to less than 100 ms. The error control unit must clock the multiplexer, however, and therefore the multiplexer must be of a type not sensitive to the fact that the clock supplied is nominally 15% less than each of the standard modem clock rates of 2400, 4800, and 9600 bps.

INCREASING THROUGHPUT OF BATCH TERMINALS ON SATELLITE LINES

New satellite leased lines are attractive for data communications because of their lower cost, higher reliability, and sometimes wider bandwidths than typical domestic or international terrestrial circuits. However, the propagation delay on such a circuit is typically a third of a second. Thus a terminal using a half-duplex block transmission protocol, such as bisync, would have to wait two-thirds of a second to find out if the last block it sent was received properly. If 5,000-bit blocks were being sent at 9600 bps, more than half the time would be spent in waiting; the throughput might be as low as 2400 bps if retransmissions due to errors were taken into account.

The system could be converted to a protocol such as SDLC that was designed to avoid this problem, but a much more convenient solution is to use the ARQ/FEC box described above in combination with a bisync protocol simulator. The final satellite error control unit works as follows. After a terminal has been polled by the computer, it outputs a block of data to the satellite error control unit, which gives

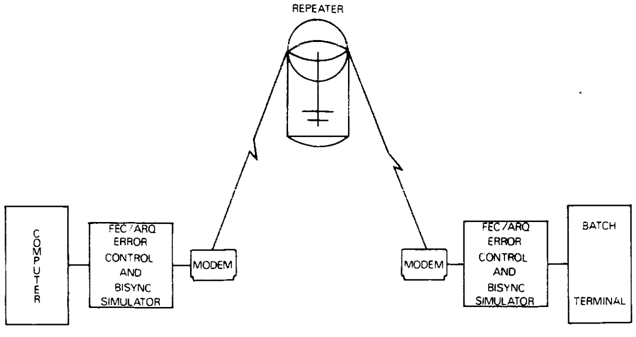


Figure 3. Elimination of satellite delay effect on throughput by error control and protocol simulation

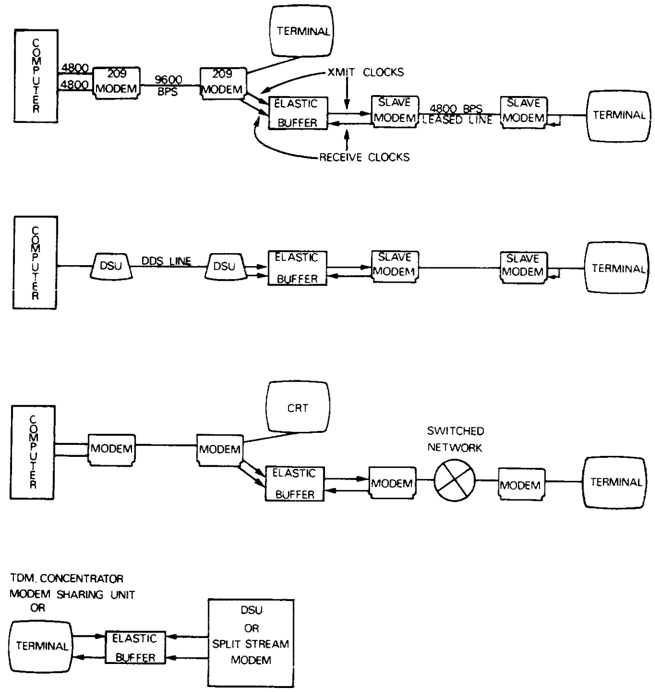


Figure 4. Applications of the elastic interface unit

back an immediate acknowledgment. The terminal then can begin immediately to release the next block of data. The first block meanwhile is formatted, transmitted, forward error corrected (or ARQed, if necessary), and delivered to the remote computer without error. The computer's acknowledgment is then ignored. On those rare occasions when too many blocks need retransmission, the terminal can be given a negative acknowledgment to keep it from outputing data until the backlog is cleared.

Any batch transmission protocol can be fooled in this manner, and improvements in throughput of 300% have been observed. The shorter the batch block and the faster the data rate, the greater the improvement. This technique is not as suitable, though, for multidrop systems using polling protocols. In such systems, one still must wait for the response to a poll from each controller or terminal; therefore the improvement in throughput is smaller, depending on the ratio of polls to data blocks.

ELASTIC INTERFACE UNIT

There are many times when it is necessary to feed data between two synchronous devices, one of which cannot be clocked externally. Some common examples are shown in Figure 4; they include the analog extension of a digital network such as DDS, the extension of unbuffered split-stream modems or synchronous time-division multiplexer channels, the connection of internally clocked terminals to the DDS network or to the second channel of a split-

Hardware Solutions to Problems in the Terminal/Computer-to-Modem Data Path

stream modem, and any half-duplex extension of DSUs, modems, or TDMs (such as a dial network link).

The basic problem in all these applications of the elastic interface unit is that the data crossing the interface is not in step with the clock that is asking for it. In the case of a dial network extension, even the data rates may be different. In particular, a DSU asks for data to transmit on pin 2 by outputting a transmit clock on pin 15. If this data is receive data from a modem, it emerges from pin 3 of the modem at a rate and phase determined by the receive data clock provided by the modem on pin 17. An elastic interface buffer for this application must accept data into a register using the modem clock on pin 17, then allow the DSU clock on pin 15 to remove it from the register. The buffer also twists the data leads so as to match data inputs and outputs.

Since the elastic interface buffer is a relatively inexpensive device, it usually only buffers eight bits or less. Its buffer is set to the midpoint when power is turned on; in the better units, the buffer is also reset to its midpoint when modem carrier comes on. This option is only significant in the dial network extension case in which the modem receive clock at the remote end of the line cannot be used as an external send clock. Thus, it would be possible for the buffer to underflow or overflow if it were not reset, preferably at the start of each new block of data.

Usually one channel of a split-stream modem can accept an external clock. Therefore the extension circuit modem or DSU can be connected to this port directly if a properly twisted cable adapter is used. However, if two or more tail circuits have to be driven from the same split-stream modem, the buffer units must be used on the second or subsequent lines if the modem port is unbuffered or cannot accept an external clock.

DATA ENCRYPTION

There are many data communications systems these days in which theft by someone tapping a phone line is a very real possibility. Such sensitive applications include credit card transaction verification, electronic fund transfer, cash dispensing machines, and confidential financial data processing. Fortunately, putting data scramblers as close to the outputs of terminals and computers as possible (see Figure 5) is a relatively inexpensive way to prevent theft or spoofing by someone tapping phone lines or EIA cables.

A data scrambler consists of a random bit generator, which, when digitally combined with data, produces an encrypted data stream (cipher) that appears to be

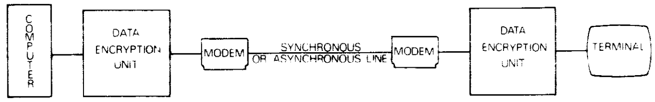


Figure 5. Data scramblers prevent theft or spoofing of data

completely random. The decryption unit generates an identical random bit pattern, which is subtracted from the encrypted stream to recover the original data.

The trick in data encryption is to make such a pseudo-random pattern, or key as it is called, so long and complicated that a thief—even with the aid of the fastest computer in existence—could not unscramble the message in a period of time short enough for the information recovered still to be of value. Many of the techniques used to generate keys are confidential, but a key usually is formed in a long shift register. The bits entering or passing through the shift register are modified by feedback from the end or several intermediate stages. By providing switches to change the feedback points, the key can be varied—which should be done at regular intervals to provide additional security. Sophisticated encryption units put an additional level of logic in the feedback paths to make the feedback nonlinear; that is, the state of one feedback signal influences the feedback to another point in the register, either at that instant or several clock cycles later. A modern key generator typically can develop 4,000 trillion different pseudo-random key patterns with key lengths of up to 10^{52} bits. Another feature of data scramblers that enhances security is a random start capability, so that the key does not always begin at the same point each time power is applied or transmission initiated.

From the communications standpoint, there are several important factors to consider. First, data scramblers must be synchronized to each other because the descrambler must know exactly what portion of the key to use to decrypt the message. There are two methods of synchronizing key generators. In the first, a synchronization pattern is sent to initialize the descrambler. Thus, every time transmission is begun, a synchronization preamble must be sent ahead of the data. The disadvantage of this type of synchronization is that, if synchronization is lost due to noise or clock drift, recovery is not automatic, and a considerable amount of data may be lost before resynchronization occurs. In polling systems, the resynchronization delay can increase the polling time substantially. However, an advantage of this technique is that simple transmission errors after synchronization are not multiplied in the decryption process.

The second synchronization method is an inherent part of the encryption scheme. If the data itself is passed through the shift register that generates the key, the data history is an element in the formulation of the key. The decryption device then becomes self-synchronizing—if enciphered data passes through the register in the decrypter, it eventually contains the exact pattern that the encrypting register did, and the registers therefore must come into lockstep. This technique eliminates the need for transmitting the key starting point; therefore, it yields less information to the thief.

The key pattern, being dependent on data, seldom repeats and is not determined solely by the feedback switch settings, which could be compromised. Attempts to modify encrypted messages en route (e.g., to change dollar amounts or account numbers) result in improper decryption because the data history is not correct. Any synchronization error is corrected automatically as more enciphered data is sent. However, transmission errors are multiplied—single bit errors usually garble the next four or five characters. The data encrypter automatically delays Clear to Send to the terminal until after synchronization is assured.

There are both synchronous and asynchronous data scramblers. In the synchronous case, all the data bits usually are enciphered so that the data scrambler need not be protocol sensitive. Where packet switching or satellite turn-around-delay elimination techniques are used, it may be necessary to leave some header characters unencrypted. For asynchronous data, it is usually the practice to leave unencrypted all control characters, such as carriage return, line feed, shift, and break signal. The encryption unit also monitors its own cipher to ensure that the data characters it encrypts do not look like such control characters. If transmission is asynchronous, the start-stop bits also must be left intact. It is possible, though, to encrypt start and stop bits if the stop bits are made integral so that clock can be recovered from the data stream to clock the decoder.

Another useful feature of asynchronous systems is a remote switching feature to shift a receiver automatically from clear to the encrypted mode. This transition is done by sending an impossible data character combination, such as QQ or KK, which causes the remote end to shift its mode. It is also possible in some encryption equipment to change the code settings electronically rather than manually. In such systems it is possible for a computer to determine—randomly and without operator knowledge—the particular key to be used to transmit data. The commands to change the key at the remote end are encrypted using the previous key setting, thereby insuring maximum protection.

CONCLUSION

There are, of course, many other black boxes that from time to time are quite useful; among these are the interface and code converters. Boxes are available to convert from EIA RS-232 to current loop, military standard, or any of various European CCITT or wideband interfaces. Also available are units that convert Baudot-coded data to ASCII or IBM code; most such devices also convert speed as well. For instance, they accept Baudot-coded data at 75 baud and output ASCII data at 1200 bps. These units usually have considerable storage capabilities so that they can accept high-speed data for output at a lower speed. Such units with large solid-state memories often are used in place of paper tape devices. In the future, interface protocol converters are likely as well. Protocol converters, such as asynchronous to bisync, or bisync to SDLC, are now practical using microprocessors.

Other devices, such as multiplexers and concentrators are dealt with later in this section.□

Planning Considerations for Dial-Up Versus Leased-Line Facilities

Problem:

There are only two parts to any communications system, no matter how complex—the termini and the lines that link the termini together. In spite of the fact that a terminus may actually consist of millions of dollars worth of computer and computer-related equipment and the lines may be just a network of plain old copper wires, the problems of planning for the right kinds of lines can be horrendously difficult because of the many options available to the users, and the wrong choice can be potentially very costly because transmission costs are a running, never ending expense that can become a terrific financial drain if the wrong service is selected.

This report explains some of the basic options available to you through the very convenient but very confusing telephone lines supplied by AT&T. Unfortunately, some of the more complex evaluations lean heavily on advanced mathematical techniques, which are touched on briefly in this report, but the techniques are necessary design tools that must be used eventually to reach an optimum cost/service configuration.

Solution:

A basic decision that must be made early in the design process is whether the terminals should be on leased or dial-up lines. If line usage is high—more than 3 hours per day, for example—then the leased-line tariff will normally prove the less expensive. If it is used infrequently, it will be cheaper to dial whenever the connection is needed.

On some systems, dialing would be quite unsatisfactory because of fast response-time requirements. Sales agents using a terminal to answer customers' telephone requests may not have the time to dial the computer; they want access to it almost immediately. On airline reservation systems, for example, the delay associated with dialing has generally been considered unacceptable. Dialing may add 20 seconds or so onto what would otherwise be a response time of 2 seconds. On the other hand, many systems do not need this frantic speed. On some systems, a lengthy man-machine conversation is going to ensue once the connection is made, and the addition of 20 seconds dialing time to a 20-minute conversation is negligible. Where batches of work are to be sent, a dialed connection is often used. However, the transmission speed on a public line is generally less than on a leased line, and so here, too, the overall costs will favor the leased line if more than a certain quantity of data is to be sent.

In this report we will discuss systems where a dial-up connection would be used if it were cheaper; there is no objection to it in terms of response time or the inconvenience of dialing.

From *Systems Analysis for Data Transmission* by James Martin. Chpt. 41, pp 711-737, p 257 and pp 700-708. © 1978 by Prentice-Hall Inc. Published by Prentice-Hall Inc., Englewood Cliffs, New Jersey. Reprinted by permission.

Planning Considerations for Dial-Up Versus Leased-Line Facilities

There are two types of systems in which public dial-up lines are used. The most common type is where the operator originates the calls and dials the computer (sometimes dials another terminal). The second is where the computer originates the call and an automatic dialing unit establishes the connection. In the first case, if the computer has no out-dialing facility, it cannot send unsolicited messages to a terminal until the operator of that terminal dials in. In the second case, the system may be designed so that the operator can dial in addition to the computer dialing out. In some systems, however, this situation does not exist, and the computer originates all calls. The operator may load cards or tapes at the terminal, and the terminal awaits a dial-up signal from the computer. This practice occurs, for example, in some message-switching systems. The computer is programmed to dial the terminals one by one to see if they have anything to send—a scanning operation that is slow because of the time taken to dial.

When the computer originates the calls, WATS lines are sometimes used, for they allow an unlimited number of calls to a given area. When the operator originates the calls, the network may be designed with a variety of different line types—Telex, TWX, INWATS, and private branch exchanges with leased lines. Often telephone lines are used because of the ubiquitous convenience of the public network.

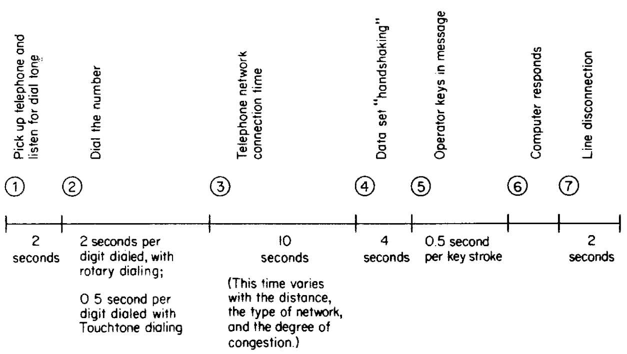


Figure 1. Typical times involved in an operator dialing a computer and obtaining a response. The connection and data set hand-shaking times should be assessed specifically for the data set and connections in question

TIMING CALCULATIONS

Figure 1 shows the elements of time that must be added when an operator dials a computer on the public network.

1. The operator picks up the telephone and receives a dial tone. On some data sets, she will press the TALK button at this time.

2. She dials the number of the computer. The time taken to do so varies with the number of digits dialed and with operator dexterity. A conservative time to use in the calculations might be 2 seconds per digit for a device with a rotary dial. Timing myself on my office telephone, I find that I take 1.5 seconds per digit dialing random numbers, 2.1 seconds per digit dialing all 9s, and 0.75 second per digit dialing all 1s.

With a Touch-Tone telephone, a reasonable time to allow is 0.5 second per digit.

3. After the dialing is completed, the equipment in the switching centers takes an additional period of time to complete the connection. This period can vary from about 1 second up to 20 seconds or more. From my telephone in Manhattan, I find that the interval between the end of dialing and the first ring of the dialed party is usually between 1 and 5 seconds. Occasionally it is longer. When calling numbers more than 500 miles away, the delay is typically about 10 seconds, but again it will occasionally be longer, sometimes 30 seconds or more. An appropriate figure can be established for the dial-up links in question.

4. On many data sets the operator hears a "data tone" when the connection is completed; she presses the DATA key and hangs up. A time of 4 seconds might be used in the calculations for this action.

5. and 6. In Figure 1, (5) and (6) are the normal times for message transmission and computer response.

7. The time for disconnecting the line. Until this time element is complete, the computer cannot accept another call on the same line.

The response time may now be defined to include time elements 1 through 6 above. The total time the line into the computer is occupied includes elements 4 through 7. During this time, the line appears busy to another incoming call. This is the time period that would be used in calculations of the number of lines needed. Knowing this time period, the systems analyst can calculate the probability of succeeding in obtaining service from the computer—the "grade of service."

A user might specify that the response time, including dialing, should not exceed 30 seconds, for example, and that the probability of obtaining a busy signal from the computer should not exceed 0.01. The systems analyst would use the foregoing timing figures to design a system that meets these requirements.

COMPUTER DIALING OUT

The computer may be equipped with an automatic dialing unit that will enable it to send unsolicited messages to terminals over the public network and to scan terminals on WATS or other suitably tariffed lines to see whether they have data ready to send.

Planning Considerations for Dial-Up Versus Leased-Line Facilities

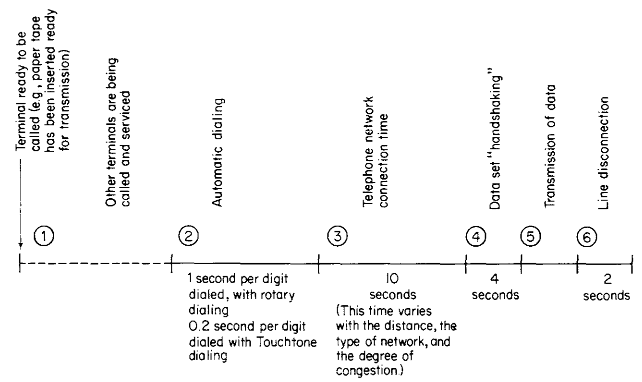


Figure 2. Typical times involved when the computer automatically dials terminals. The connection and data set handshaking times should be assessed specifically for the data set and connection in question

Figure 2 shows typical times involved when the computer dials out.

1. At the beginning of time element 1, the terminal is ready to transmit. The computer, or its transmission control unit, however, is busy dialing other terminals. Eventually it dials the terminal in question.

2. The dialing time, as in the previous case, varies with the number and value of digits dialed. It is at least twice as fast as manual dialing. If the systems analyst cannot obtain exact figures for the equipment in question, he might assume 1 second per digit for rotary dialing and 0.2 second per digit for Touchstone.

3. The same comments as before apply to the telephone network connection time.

4. When the connection is established, the terminal must respond to the "ringing," and a data set handshaking operation takes place. The time for thus establishing the data path will vary from one type of equipment to another. A typical figure of 4 seconds is given in Figure 2.

5. The time for the transmission of data must include the necessary control characters and timeout intervals, if any.

6. Finally, there is the disconnection time, from the end of transmission of data to the time the next call can be initiated—typically about 2 seconds.

The line-control equipment at the computer is tied up for time elements 2 through 6 above. The transaction response time for incoming transactions will include items 1 through 7.

The mean response time can be found in the following way.

Let us consider a system in which the computer dials M terminals. It dials them in sequence to see whether they have data to send.

Let t_M equal the time for dialing a terminal and reading its data (in seconds), t_0 equal the time for dialing and disconnecting a terminal that has nothing to send (seconds), and n equal the number of transactions per hour handled by the system.

The proportion of the time spent servicing terminals that *have* data to send is

$$\frac{nt_M}{3600}$$

The proportion of the time spent scanning terminals that do not have data to send is

$$1 - \frac{nt_M}{3600}$$

The mean time to service one terminal is therefore

$$\frac{nt_M}{3600} t_M + \left(1 - \frac{nt_M}{3600}\right) t_0 = \frac{nt_M(t_M - t_0)}{3600} + t_0 \quad \text{Eq. 1}$$

A terminal having data to send may be lucky and not have to wait before it is serviced. On the other hand, it may be unlucky and have to wait while $M - 1$ terminals are serviced. The probability of it having to wait while J terminals are serviced (where $0 \leq J \leq M - 1$) is $1/M$.

The mean waiting time is then

$$\sum_{J=0}^{M-1} \frac{1}{M} J \left[\frac{nt_M(t_M - t_0)}{3600} + t_0\right] = \frac{M-1}{2} \left[\frac{nt_M(t_M - t_0)}{3600} + t_0\right]$$

The mean response time is therefore

$$E(t_R) = t_M + \frac{M-1}{2}\left[\frac{nt_M(t_M - t_0)}{3600} + t_0\right] \quad \text{Eq. 2}$$

Example 1. A System with Computer-Originated Dialing. *A sales-order entry system has 200 terminals in different sales offices. Data and queries about orders are entered into these terminals. The computer will dial the terminals, read what data has been entered, and send back a response. When a terminal with no data is dialed, the line into the transmission control unit will be tied up for 17 seconds. When a terminal with data is dialed, the operation, including the response to the terminal, takes 35 seconds on the average. The peak volume expected is 400 transactions per hour. How many lines into the transmission*

Planning Considerations for Dial-Up Versus Leased-Line Facilities

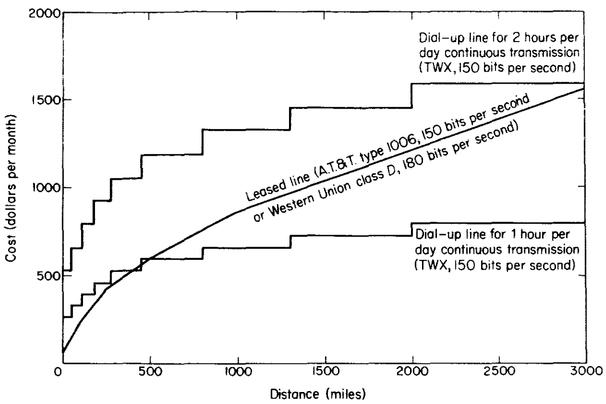


Figure 3. A comparison of leased and dial subvoice-grade lines. Note: The costs are likely to change

control unit are needed in order to achieve a mean response time no greater than 5 minutes?

We have

$$t_0 = 17 \quad \text{and} \quad t_M = 35$$

Let L be the number of lines that are used. The maximum number of terminals per line is

$$M = \left[\frac{200}{L}\right]$$

The maximum number of transactions per hour per line is

$$n = \left[\frac{400}{L}\right]$$

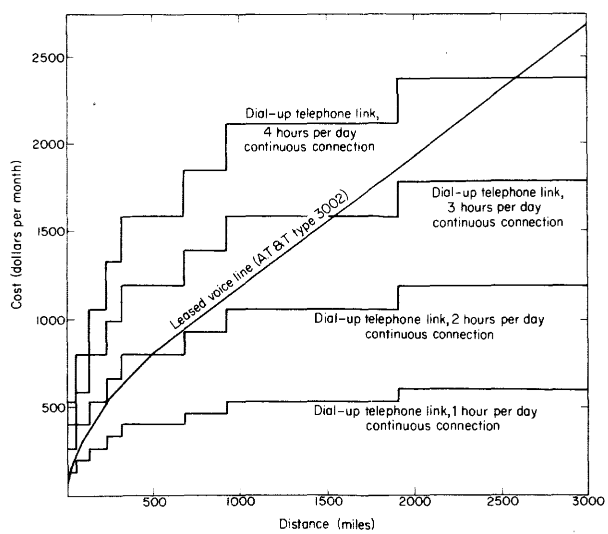


Figure 4. A comparison of leased and dial voice-grade lines. Note: The costs are likely to change. Dial-up costs are lower at off-peak hours

[1]Note that all charges quoted are subject to change.

(assuming that the traffic volume can be distributed evenly between the lines. A detailed knowledge of the traffic breakdown would give an exact figure for the highest value of n.)

From Eq. 2 we have the mean response time

$$E(t_R) = 35 + \frac{[200/L] - 1}{2}\left\{\frac{[400/L] \times 35 \times (35 - 17)}{3600} + 17\right\}$$

With $L = 9$, $E(t_R) = 308.6$ seconds—slightly too high. With $L = 10$. $E(t_R) = 263.0$ seconds.

Therefore 10 lines into the transmission control unit are needed.

RELATIVE COSTS

Assuming that there is no systems objection to dial-up lines, the systems analyst must calculate whether they are cheaper than leased lines.

Figures 3 and 4 indicate the variation in cost with distance of voice-grade and subvoice-grade lines. The tariffs are likely to change, and these diagrams should only be taken as typical of relative costs. The dial-up lines in Figure 3 are TWX lines used for continuous transmission (i.e., not for short, separate transmissions). On this basis, it will be seen that the leased line is always cheaper than the dial-up connection for 2 hours worth of transmission per day. For one hour of transmission per day, it is still cheaper below distances of about 400 miles. Figure 4 is a similar plot for voice lines. It will be seen that the breakeven point varies considerably with distance.

Surprisingly, perhaps, the two figures show that the dial-up voice-grade line is cheaper than the subvoice line, although it has a much higher capacity. The TWX line should therefore not be used for continuous transmission.

Telex rates are lower than toll telephone rates for most locations in the United States, but are sufficiently close for it to be generally uneconomical to use Telex for continuous transmission. Normally Telex is used for sending relatively short transmissions of not more than a few hundred characters. The minimum charge[1] for such a connection is 20 to 60 cents for one minute of transmission, depending on distance (at 15 characters per second, this is 900 characters). We should therefore evaluate the cost of the subvoice-grade lines on the basis of the number of short calls. Figure 5 compares dial-up cost with leased line cost on this basis, and Figure 6 does the same for a voice line. The minimum charge on a voice line is 30 cents to $1.70 for 3 minutes. On this basis, the TWX line is substantially cheaper than toll telephone, especially for long

Planning Considerations for Dial-Up Versus Leased-Line Facilities

distances. Once again the reader should note that the tariffs are likely to change.

Figures 7 and 8 show similar costs for lines in Great Britain. Here the Telex line is more expensive than the voice line for continuous transmission. The minimum charge for a short connection is 1p for either Telex or STD telephone lines. The time that this 1p pays for is shown in tabular form:

Distance (miles)	Telex (seconds)	STD Telephone (seconds)
Up to 35	60	30 (6 minutes for local calls)
35 to 50	30	15
50 to 75	30	10
Over 75	15	10

A "short" connection over 75 miles on Telex is therefore 15 seconds or less, in which 100 characters or less can be transmitted.

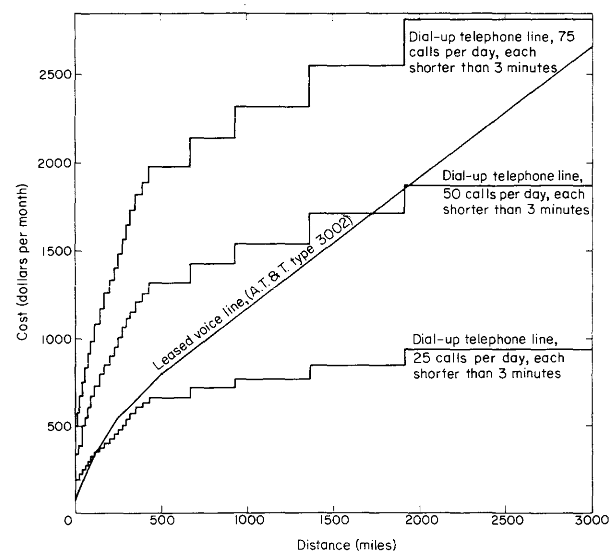


Figure 6. A comparison of leased and dial telephone charges for short calls. Note: The charges are likely to change. Dial-up charges are lower for off-peak hours

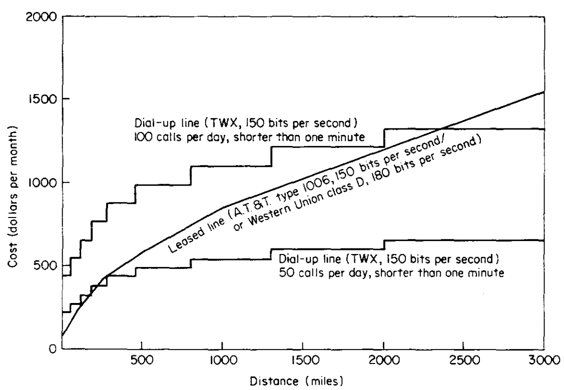


Figure 5. A comparison of leased and dial subvoice-grade lines for short message transmission. Note: The costs are likely to change

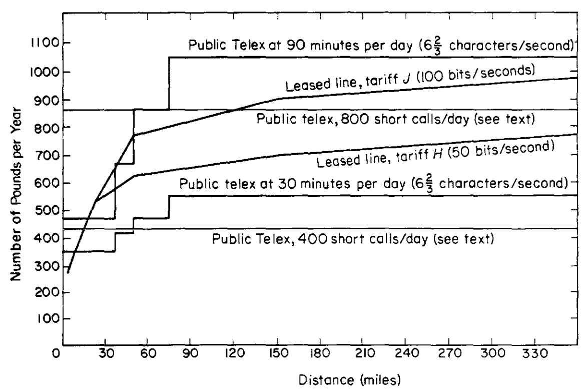


Figure 7. Cost comparison of dial-up Telex and leased telegraph lines in Britain

The breakeven point in cost between private and public lines varies with the distance of the link as well as with the amount of time the line is used. Figures 9 and 10 illustrate this. Figure 10 charts the breakeven point between leased and dial voice lines. Figure 9 shows the breakeven point between leased subvoice-grade lines (Type 1006) and dial voice lines. Voice lines were used in this plot rather than TWX lines because they are cheaper for continuous transmission. In both cases the dial-up tariff looks more favorable at a long distance.

Figure 11 compares the WATS tariff cost with leased voice line costs. Except at long distances, the monthly flat rate for the WATS line is higher than that for the leased voice line.

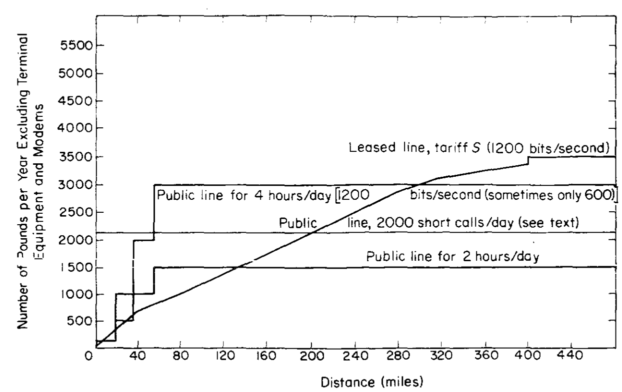


Figure 8. Cost comparison of dial-up and leased telephone lines in Britain

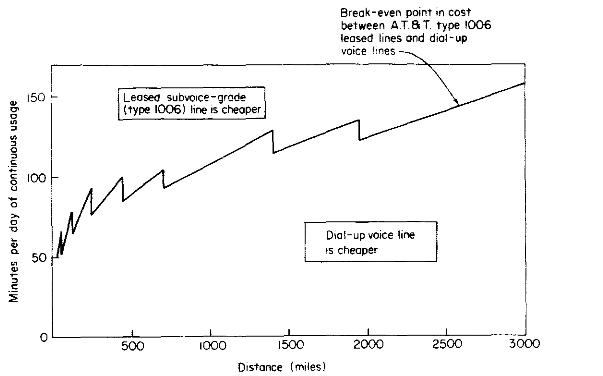


Figure 9. A curve showing the break-even point in cost between leased subvoice-grade (Type 1006, 150-bit-per-second) lines and dial-up voice lines. Note: The calculations assumed 22 working days in the month. Dial-up voice lines were plotted rather than TWX because they are cheaper for continuous transmission. The price figures used are subject to change

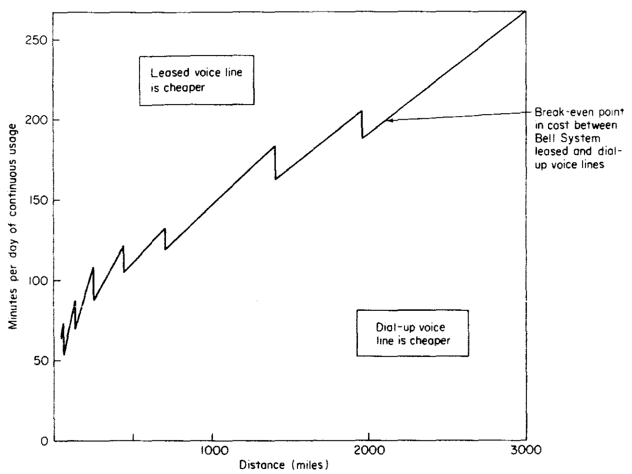


Figure 10. A curve showing the break-even point in cost between leased and dial-up line costs on Bell System interstate voice-grade lines. Note: The calculations assumed 22 working days in the month. The price figures used in plotting the curve are subject to change

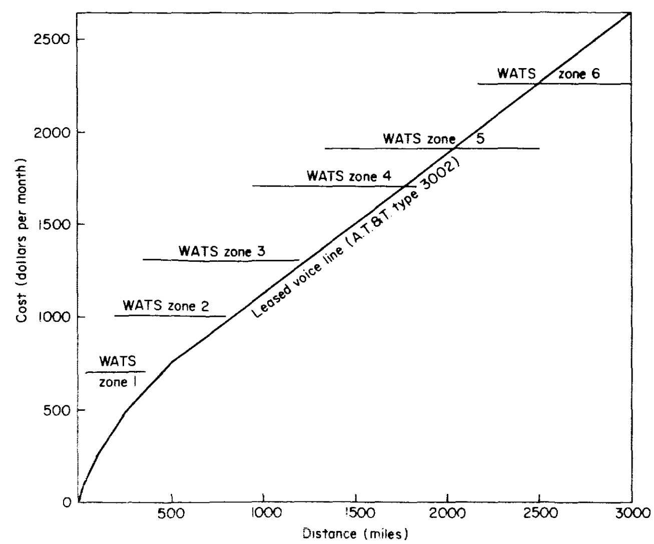


Figure 11. A comparison of WATS line costs with those of a leased voice line. WATS lines to Southeast New York were used

COMPLEX NETWORKS

For a point-to-point link, there is no difficulty in working out relative costs if the volume of data to be transmitted is known. Where many terminals are involved, however, the situation is more complex.

Sometimes the tariffs and networks indicate clearly that all the links should be dial-up. Sometimes it is clear that all of them should be leased lines. Often, however, the lowest-cost solution is to have some dial-up connections and some leased. This solution is sometimes not considered by the network designers. There is a tendency to make the network entirely one or the other.

In this next example we will demonstrate a unified leased-line design approach to a complex sample problem, and we will then re-examine the same problem from a mixed-design point of view.

Example 2. A Network Too Complex to Optimize by Hand. *Again a network is to be designed to handle data from terminals connected to a data processing center in Chicago. Now, however, there are 187 terminal locations, as shown in Table 1.*

City	Coordinates		Traffic Volume (messages/hour)	
	V	*H*	Input	Output
MONTGOMERY ALA	7692	2247	2.09	3.10
MOBILE ALA	8167	2367	1.92	2.84
BIRMINGHAM ALA	7518	2446	2.17	3.20
MERIDEN CONN	4740	1358	2.20	3.26
WATERBURY CONN	4761	1391	2.67	3.95
HARTFORD CONN	4687	1373	3.72	5.51
NORWICH CONN	4668	1263	5.30	7.84
NEW BRITAIN CONN	4715	1373	3.05	4.51
NEW HAVEN CONN	4792	1342	0.43	0.64
DANBURY CONN	4829	1423	3.51	5.20
BRIDGEPORT CONN	4841	1360	5.85	8.65
CANAAN CONN	4703	1485	2.07	3.06
WILMINGTON DEL	5326	1485	2.81	4.16
WASHINGTON DC	5622	1583	7.21	10.68
ORLANDO FLA	7954	1031	1.44	2.13
WEST PALM BEACH FLA	8166	0607	2.90	4.29
TAMPA FLA	8173	1147	1.60	2.37
TALLAHASSE FLA	7877	1716	1.43	2.12
ST. PETERSBURG FLA	8224	1159	0.57	0.85
MIAMI FLA	8351	0527	7.16	10.60
JACKSONVILLE FLA	7649	1276	0.57	0.85
FORT LAUDERDALE FLA	8282	0557	5.40	8.00
MACON GA	7364	1865	0.43	0.64
SAVANNAH GA	7266	1379	2.13	3.15
AUGUSTA GA	7089	1674	0.63	0.94
ATLANTA GA	7260	2083	0.89	1.32
CENTRALIA ILL	6744	3311	1.99	2.94
JOLIET ILL	6088	3454	2.75	4.07
HIGHLAND PARK ILL	5940	3480	3.01	4.45
SPRINGFIELD ILL	6539	3513	3.54	5.23
ROCKFORD ILL	6022	3675	3.44	5.09
QUINCY ILL	6642	3790	0.98	1.44
PEORIA ILL	6362	3592	3.50	5.18
GALESBURG ILL	6369	3732	2.16	3.20
DECATUR ILL	6478	3413	2.22	3.28
DANVILLE ILL	6322	3245	2.97	4.39
CHICAGO B ZONE 1	5986	3426	5.25	7.77
CHICAGO D ZONE 2	5971	3443	4.52	6.69
CHICAGO F ZONE 2	5971	3443	2.70	3.99
CHICAGO H ZONE 3	5979	3455	3.02	4.47
CHICAGO J ZONE 4	5981	3437	8.10	11.99
CHICAGO L ZONE 5	5991	3449	3.65	5.41
CHICAGO N ZONE 6	5998	3431	4.22	6.24
CHICAGO P ZONE 7	6007	3412	3.45	5.11
CHICAGO R ZONE 8	6011	3424	4.14	6.13
CHICAGO S ZONE 9	6014	3397	7.46	11.04

Table 1. Terminal locations and volumes for design example

Planning Considerations for Dial-Up Versus Leased-Line Facilities

City	Coordinates V	Coordinates H	Traffic Volume (messages/hour) Input	Traffic Volume (messages/hour) Output
CHICAGO U ZONE 10	6022	3407	3.16	4.67
CHICAGO X ZONE 11	5988	3474	0.69	1.02
CHAMPAIGN ILL	6371	3336	3.37	4.99
BLOOMINGTON ILL	6358	3483	1.69	2.49
MICHIGAN CITY IND	5962	3301	4.01	5.94
KOKOMO IND	6135	3063	1.10	1.63
FORT WAYNE IND	5942	2982	5.65	8.36
ELKHART IND	5895	3168	2.29	3.39
BLOOMINGTON IND	6417	2984	1.90	2.81
MASON CITY IOWA	6136	4352	0.57	0.85
CLINTON IOWA	6180	3793	0.43	0.64
FORT DODGE IOWA	6328	4438	0.59	0.88
IOWA CITY IOWA	6313	3972	1.44	2.13
DES MOINES IOWA	6471	4275	3.38	5.01
DAVENPORT IOWA	6273	3817	2.08	3.08
CEDAR RAPIDS IOWA	6261	4021	0.98	1.44
WICHITA KAN	7489	4520	0.43	0.64
TOPEKA KAN	7110	4369	2.05	3.04
MADISONVILLE KY	6845	2942	0.59	0.88
LEXINGTON KY	6459	2562	2.87	4.25
WINCHESTER KY	6441	2509	2.20	3.26
ASHLAND KY	6220	2334	0.63	0.94
PADUCAH KY	6982	3088	0.63	0.94
LOUISVILLE KY	6529	2772	0.89	1.32
MONROE LA	8148	3218	1.44	2.13
SHREVEPORT LA	8272	3495	0.43	0.64
NEW ORLEANS LA	8483	2638	0.57	0.85
BATON ROUGE LA	8476	2874	1.56	2.31
HAGERSTOWN MD	5555	1772	3.39	5.01
CUMBERLAND MD	5650	1916	2.72	4.03
BALTIMORE MD	5510	1575	8.95	13.24
ANNAPOLIS MD	5555	1519	1.66	2.46
SPRINGFIELD MASS	4620	1408	4.45	6.59
BOSTON MASS	4422	1249	5.63	8.33
GREENFIELD MASS	4537	1475	2.04	3.03
WORCESTER MASS	4513	1330	3.47	5.13
SALEM MASS	4378	1251	2.75	4.07
JACKSON MICH	5663	3009	3.72	5.51
ANN ARBOR MICH	5602	2918	3.55	5.26
MUSKEGON MICH	5622	3370	1.93	2.86
LANSING MICH	5584	3081	2.52	3.74
KALAMAZOO MICH	5749	3177	2.27	3.35
GRAND RAPIDS MICH	5628	3261	2.40	3.55
FLINT MICH	5461	2993	3.58	5.31
SAGINAW MICH	5404	3074	0.57	0.85
DETROIT 1 MICH	5536	2828	9.43	13.95
DETROIT 2 MICH	5536	2828	4.93	7.30
ROCHESTER MINN	5916	4326	0.71	1.04
DULUTH MINN	5352	4530	0.63	0.94
JACKSON MISS	8035	2880	0.43	0.64
ST LOUIS MO	6807	3482	1.23	1.82
JEFFERSON CITY MO	6963	3782	2.87	4.25
JOPLIN MO	7421	4015	1.49	2.21
SPRINGFIELD MO	7310	3836	3.58	5.29
KANSAS CITY MO	7027	4203	7.19	10.65
OKLAHOMA CITY OKLA	7947	4373	1.76	2.61
TULSA OKLA	7707	4173	0.85	1.26
ATLANTIC CITY NJ	5284	1284	3.14	4.64
NEWARK NJ	5015	1430	6.10	9.04
PATERSON NJ	4984	1452	3.79	5.61
TRENTON NJ	5164	1440	2.86	4.23
AUBURN NY	4858	2030	1.23	1.82
PEEKSKILL NY	4894	1470	3.74	5.54
WHITE PLAINS NY	4921	1416	6.35	9.40
NIAGARA FALLS NY	5053	2377	3.26	4.82
NEW YORK CITY 1	4972	1408	6.09	9.01
NEW YORK CITY 2	5004	1392	3.08	4.55
NEW YORK CITY 3	4975	1387	4.75	7.02
NEW YORK CITY 4	4970	1379	7.43	11.00
NEW YORK CITY 5	5000	1358	6.29	9.30
NEW YORK CITY 6	5054	1407	5.17	7.66
UTICA NY	4701	1878	5.46	8.08
SYRACUSE NY	4798	1990	3.43	5.07
SCHENECTADY NY	4629	1675	4.67	6.91
ROCHESTER NY	4913	2195	8.95	13.25
POUGHKEEPSIE NY	4821	1526	4.62	6.84
ITHACA NY	4938	1958	2.25	3.34
BUFFALO NY	5075	2326	9.58	14.18
BINGHAMTON NY	4943	1837	4.99	7.38
ALBANY NY	4639	1629	0.63	0.94
GOLDSBORO NC	6352	1290	2.07	3.06
ASHEVILLE NC	6749	2001	2.61	3.86
WINSTON SALEM NC	6440	1710	2.08	3.07
RALEIGH NC	6344	1436	2.31	3.42
FAYETTEVILLE NC	6501	1385	1.62	2.39
CHARLOTTE NC	6657	1698	0.43	0.64
MANSFIELD OHIO	5783	2575	2.26	3.35
CLEVELAND OHIO	5574	2543	3.09	4.58
YOUNGSTOWN OHIO	5557	2353	3.66	5.42
STEUBENVILLE OHIO	5689	2262	2.35	3.48
SPRINGFIELD OHIO	6049	2666	3.13	4.64

Table 1. Terminal locations and volumes for design example (continued)

City	Coordinates V	Coordinates H	Traffic Volume (messages/hour) Input	Traffic Volume (messages/hour) Output
SANDUSKY OHIO	5670	2682	1.58	2.33
LIMA OHIO	5921	2799	1.88	2.78
DAYTON OHIO	6113	2705	2.72	4.03
COLUMBUS OHIO	5972	2555	4.26	6.31
CINCINNATI 2 OHIO	6263	2679	2.52	3.73
CINCINNATI 1 OHIO	6263	2679	4.56	6.75
CANTON OHIO	5676	2419	0.43	0.64
ASHTABULA OHIO	5429	2462	1.61	2.39
ALLIANCE OHIO	5629	2395	3.14	4.65
AKRON OHIO	5637	2472	0.89	1.32
WILLIAMSPORT PENN	5201	1876	0.43	0.64
SCRANTON PENN	5042	1715	1.72	2.54
HARRISBURG PENN	5363	1733	1.76	2.61
BRADFORD PENN	5221	2182	1.68	2.49
WILKES-BARRE PENN	5093	1723	2.14	3.17
PHILADELPHIA PENN	5251	1458	8.91	13.19
PITTSBURGH PENN	5621	2185	8.85	13.09
ROCHESTER PENN	5589	2251	0.85	1.26
NEWPORT RI	4596	1160	3.45	5.10
PROVIDENCE RI	4550	1219	4.13	6.11
FLORENCE SC	6744	1417	0.86	1.28
SPARTENBURG SC	6811	1833	0.63	0.94
CHARLESTON SC	7021	1281	2.45	3.62
GREENVILLE SC	6873	1894	1.99	2.95
COLUMBIA SC	6901	1589	2.17	3.21
NASHVILLE TENN	7010	2710	2.85	4.21
KNOXVILLE TENN	6801	2251	0.63	0.94
MEMPHIS TENN	7471	3125	1.23	1.82
CHATTANOOGA TENN	7098	2366	2.54	3.76
AUSTIN TEX	9005	3996	1.53	2.26
SAN ANTONIO TEX	9225	4062	2.89	4.28
HOUSTON TEX	8934	3536	2.19	3.25
FORT WORTH TEX	8479	4122	1.05	1.55
CORPUS CHRISTI TEX	9475	3739	1.97	2.92
DALLAS TEX	8436	4034	2.10	3.11
DANVILLE VA	6270	1640	0.63	0.94
STAUNTON VA	5953	1781	1.19	1.76
RICHMOND VA	5906	1472	3.81	5.64
PETERSBURG VA	5961	1429	2.06	3.05
NORFOLK VA	5918	1223	3.14	4.65
LYNCHBURG VA	6093	1703	1.62	2.40
WHEELING W VA	5755	2241	1.43	2.12
CHARLESTON W VA	6152	2174	1.85	2.74
EAU CLAIRE WIS	5698	4261	1.39	2.06
GREEN BAY WIS	5512	3747	1.49	2.21
APPLETONN WIS	5589	3776	1.44	2.13
MILWAUKEE WIS	5788	3589	9.52	14.08
MADISON WIS	5887	3796	0.63	0.94
RHINELANDER WIS	5394	4053	1.23	1.82
FORT SMITH ARK	7752	3855	1.23	1.82
LITTLE ROCK ARK	7721	3451	1.72	2.54
HOT SPRINGS ARK	7827	3554	1.19	1.76

Table 1. Terminal locations and volumes for design example (continued)

Table 1 gives the coordinates of the terminal locations, plus the traffic volumes that the network is to be designed for. In this case, there are more output than input messages. The reason is that some input messages cause a response to be sent to more than one terminal. Also the computer sends some unsolicited messages.

The terminals that are to be used operate over a half-duplex line at 14.8 characters per second. Not more than 26 terminals can be attached to one such line. Line loading and queuing calculations have resulted in the traffic rate table shown in Table 2.

A first step in laying out the network is to determine how many terminals will be in use simultaneously at each location. A study of the operator's work in this example led to the criterion that if there are less than 24 messages per hour, only one active terminal will be needed. Similarly, if there are between 24 and 48

Number of Terminals per Line	Maximum Traffic Rate (messages/hour)
1	350.300
2	343.300
3	336.200
4	329.400
5	322.800
6	318.600
7	310.400
8	304.500
9	298.600
10	293.000
11	287.600
12	282.300
13	277.200
14	272.200
15	267.400
16	262.700
17	258.100
18	253.500
19	249.000
20	244.800
21	240.700
22	236.800
23	232.800
24	228.800
25	225.100
26	221.500

Table 2. Traffic rates for design example

messages per hour, 2 terminals will be needed. If there are between 48 and 72 messages per hour, 3 will be needed and so on. A few locations will have an extra terminal for reliability, but this factor does not affect the traffic rate table, for only one will be connected to the line at one time. With the traffic volumes listed above, no location need have more than one terminal in use at one time. However, we will repeat this example with higher traffic volumes and then some locations will need several terminals.

The network was designed for Bell System type 1006 lines. The resulting configuration is shown in Figure 12. Eight lines are needed, some of which have the maximum of 26 locations connected to them. The cost of the network (at the time of writing), including the channel terminal charges, is $25,890 per month.

Increasing the Traffic Volume

Let us repeat our Example 2 with different traffic volumes. First, we will double the traffic volume. The resulting network is shown in Figure 13. Now 11 lines are needed, including one that does not leave Chicago. Although the traffic volume has doubled, the network cost has increased by only about 4.4 percent. The reason is that some of the lines with 26 drops were not fully loaded in Figure 12. Furthermore, the new interconnections shown in Figure 13 have increased the total mileage by only 10.5 percent.

Next the traffic volume of Figure 12 is multiplied by eight, which means that many of the locations must now have more than one terminal. The resulting network is shown in Figure 14. There are now 38 lines, including 3 from New York and 3 that do not leave Chicago. The total line cost is $39,000. Although the traffic volume has gone up by 800 percent, the total line mileage has increased by only slightly more than double. The number of line terminations, including those at Chicago (and hence the channel terminal charges), have increased by only 47 percent.

Figure 15 plots the variation in line cost for interconnecting these cities against the traffic volume. It will be seen that the cost rises only slowly at low-traffic volumes. It will then perhaps be worthwhile to design the network with some slack in it so that the network can handle abnormally high peak loads.

Example 3. A Network with Both Dial-Up and Leased Lines. *Consider Example 2 again, with the*

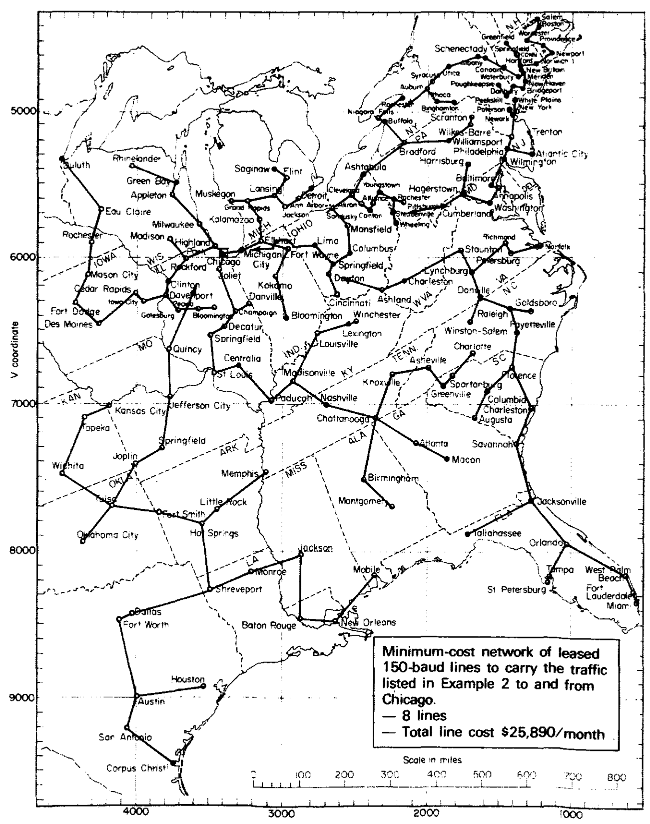


Figure 12. The minimum-cost network for Example 2

Planning Considerations for Dial-Up Versus Leased-Line Facilities

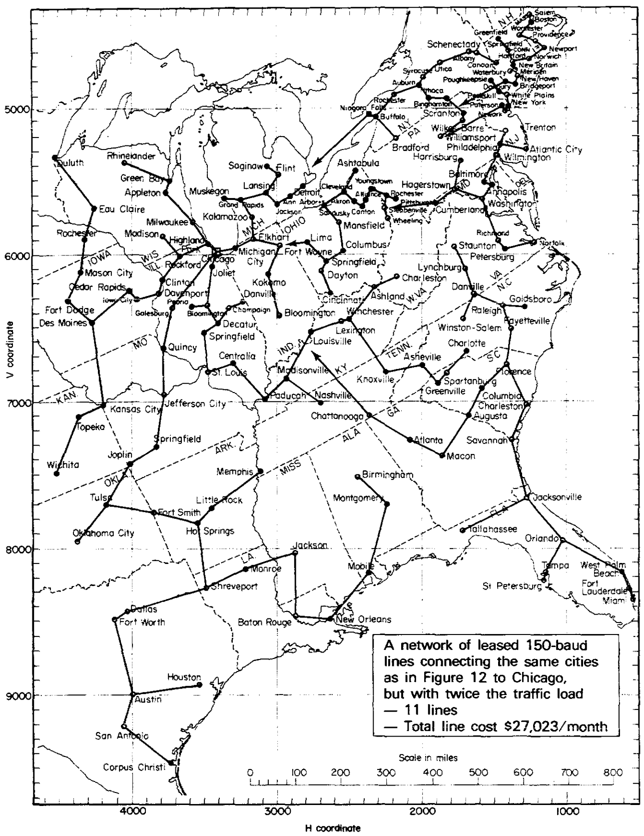


Figure 13. A network for the same cities as in Figure 12, but with twice the traffic load

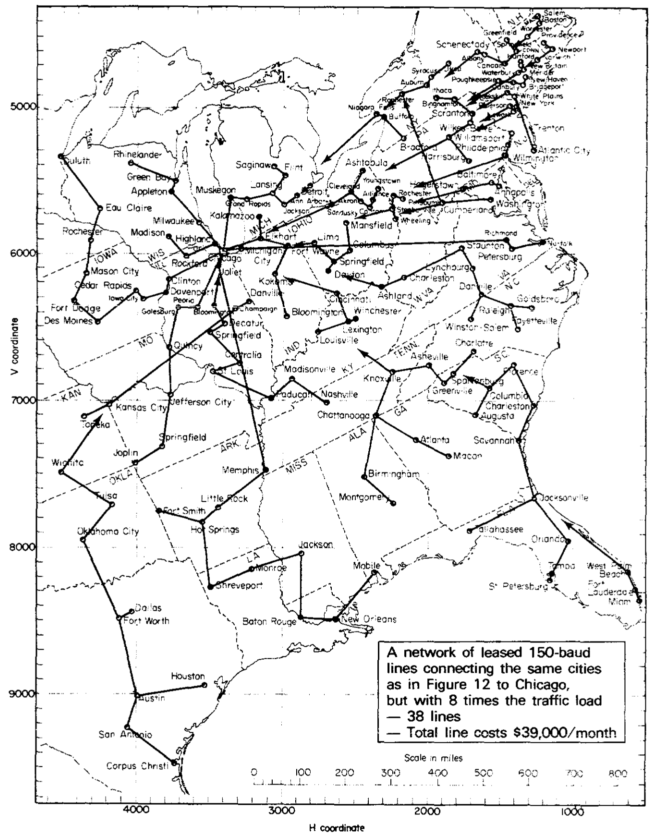


Figure 14. A network for the same cities as in Figure 12 but with eight times the traffic load

possibility of using dial-up TWX connections from some of the cities to the computer center at Chicago, thereby excluding them from the leased-line network.

All the transactions are sufficiently short so that a message and its response will not occupy a 150-baud line for more than one minute. If TWX lines were used, we would therefore pay for one minute of time for each message plus response. Some of the messages are unsolicited output from the computer, and these messages would also need a payment for one minute of line time. The volume figures quoted in Example 2 are for the peak hour. The peak-hour volume in this example is one-fourth of the daily volume.

In deciding whether a city should be on a dial-up line, we must compare the monthly cost of its dialing Chicago with the incremental cost of including it on the leased-line network. This incremental cost will depend on which nearby city it might be connected to on the leased-line network. Unfortunately, this factor cannot be assessed simply because the structure of the leased-line network is going to change severely.

It would be useful to have a program that tackles calculations of this type. A development of the CNDP algorithm* could do just that. At each step it would examine the trade-off between making a leased-line interconnection and removing the city in

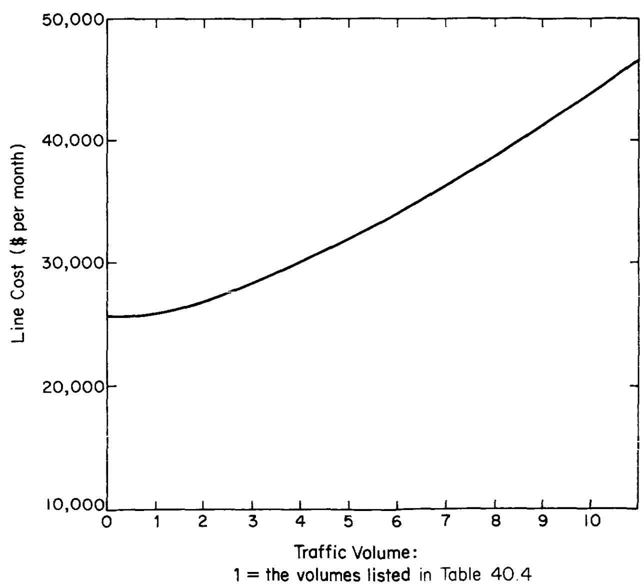


Figure 15. The variation of cost with traffic volume for a leased line network connecting the cities in Figure 12 to Chicago.

*The CNDP algorithm as described by L.R. Esau and K.C. Williams in "A Method for Approximating the Optimal Network," *IBM Systems Journal*, Vol. 5, No. 3, Armonk, 1966.

Planning Considerations for Dial-Up Versus Leased-Line Facilities

question from the leased-line network. No such program existed at the time this report was written. The systems designer, indeed, is often faced with situations in which the exact design tools he needs do not exist. He can either produce such tools himself, which in this case is a very lengthy process, or he can find a quicker but approximate expedient.

An approximate solution was produced in this case by H.M. Krouse at the IBM Systems Research Institute. He used an algorithm that assigns locations to dial-up lines in the following manner:

1. Starting with the city farthest from the computer center and not yet assigned to a dial-up line, find the closest neighbor that has also not yet been assigned to a dial-up line.

2. Calculate the cost of a leased line interconnecting these two cities. Only one channel-terminal charge should be added to this calculation.

3. Calculate the cost of using a dial-up line from the city in question to the computer, taking the daily volumes into consideration.

4. If the dial-up line is the less expensive of the foregoing alternatives, then assign the city to a dial-up line.

5. Repeat the preceding four steps for all the cities that will have a terminal.

6. It is possible that some cities not assigned to a dial-up line in the foregoing process have now become eligible for one. Their nearest city in step 1 may itself have now been assigned to dial up, and earlier steps must be done again, taking this factor into consideration. Therefore repeat steps 1 to 5.

7. Repeat the preceding steps until a cycle is reached in which no more cities become assigned to a dial-up line.

8. Make a list of the leased-line interconnections that were finally assumed in the foregoing calculation.

9. The minimum-cost leased-line configuration must now be found for the cities not assigned to dial-up lines. This step is done with a program using the CNDP algorithm.

The reason why this algorithm is not exact is that occasionally a leased-line segment does not connect a location to the nearest neighbor that is also on a leased line. A glance ahead at Figure 18 makes this point clear. Cumberland, Maryland, is not connected to Pittsburgh, Pa., for example. The city interconnection listing produced in step 9 is nevertheless close to the optimum network. As we shall see below, additional minor refinements can be made by hand.

In the program used for this algorithm, the costs are evaluated as follows (where D is the distance in miles and LCOST is the leased-line cost in dollars per month):

In PL/1:

```
IF D>1000 THEN DO;
  COST=862.0 + CEIL(D-1000)*.350;GO TO B3;END;
IF D>500  THEN DO;
  COST=597.0 + CEIL(D-500 )*.53 ;GO TO B3;END;
IF D>250  THEN DO;
  COST=422.0 + CEIL(D-250 )*.700;GO TO B3;END;
IF D>100  THEN DO;
  COST=237.5 + CEIL(D-100 )*1.23 ;GO TO B3;END;
  COST= 62.5 + CEIL(D)*1.75;
B3:
```

IN FORTRAN:

```
      DIMENSION DISTC(5),PRICEC(5),PRCNTC(5)
      DATA DISTC /1000.,500.0,250.,100.,  0.0/,
     1     PRICEC/862.0,597.0,422.0,237.5, 62.55/,
     2     PRCNTC/ .350, .530, .700, 1.230, 1.75/
C
      DO 140 I=1,5
      IF (D.LE.DISTC(I)) GO TO 140
      LCOST=PRICEC(I)+(D-DISTC(I))*PRCNTC(I)
      GO TO 150
  140 CONTINUE
  150 CONTINUE
```

For the dial-up, TWX lines, TIME is the time of connection and DCOST is the cost in dollars per month. In PL/I:

```
IF D>2000 THEN DO;DCOST=.60*CEIL(TIME);GO TO B4;END;
IF D>1300 THEN DO;DCOST=.55*CEIL(TIME);GO TO B4;END;
IF D>800  THEN DC;DCOST=.50*CEIL(TIME);GO TO B4;END;
IF D>450  THEN DO;DCOST=.45*CEIL(TIME);GO TO B4;END;
IF D>280  THEN DO;DCOST=.40*CEIL(TIME);GO TO B4;END;
IF D>185  THEN DO;DCOST=.35*CEIL(TIME);GO TO B4;END;
IF D>110  THEN DC;DCOST=.30*CEIL(TIME);GO TO B4;END;
IF D>50   THEN DO;DCOST=.25*CEIL(TIME);GO TO B4;END;
                  DCOST=.20*CEIL(TIME);
B4:
```

IN FORTRAN:

```
      DIMENSION DISTD(9),PRCNTD(9)
      DATA DISTD /2000.,1300.,800.,450.,280.,185.,110., 50., 0./,
     1     PRCNTD/ .60, .55,.50,.45,.40,.35,.30,.25,.20/
C
      DO 180 I=1,9
      IF (D.LE.DISTD(I)) GO TO 180
      DCOST=PRCNTD(I)*MINS
      GO TO 210
  180 CONTINUE
  210 CONTINUE
```

Similar segments of code could be used for other tariffs. Here, for example, are subroutines for voice-grade lines used in plotting Figure 10.

For toll telephone connections, where MINS is the number of minutes (of continuous transmission) per day: In PL/I:

```
TIME=MINS-3;
IF D>1910 THEN DO;DCOST=1.35+.45*CEIL(TIME);GO TO B2;END;
IF D>1360 THEN DO;DCOST=1.25+.40*CEIL(TIME);GO TO B2;END;
IF D> 925 THEN DO;DCOST=1.15+.35*CEIL(TIME);GO TO B2;END;
IF D> 675 THEN DO;DCOST=1.05+.35*CEIL(TIME);GO TO B2;END;
```

Planning Considerations for Dial-Up Versus Leased-Line Facilities

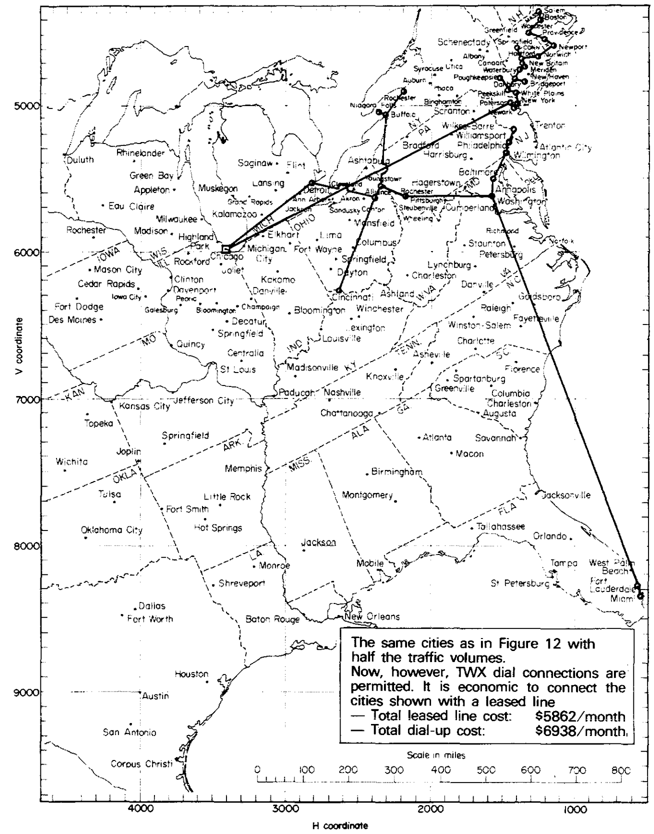


Figure 16. The lowest-cost network. The cities marked on the map with no line going to them use TWX dialing

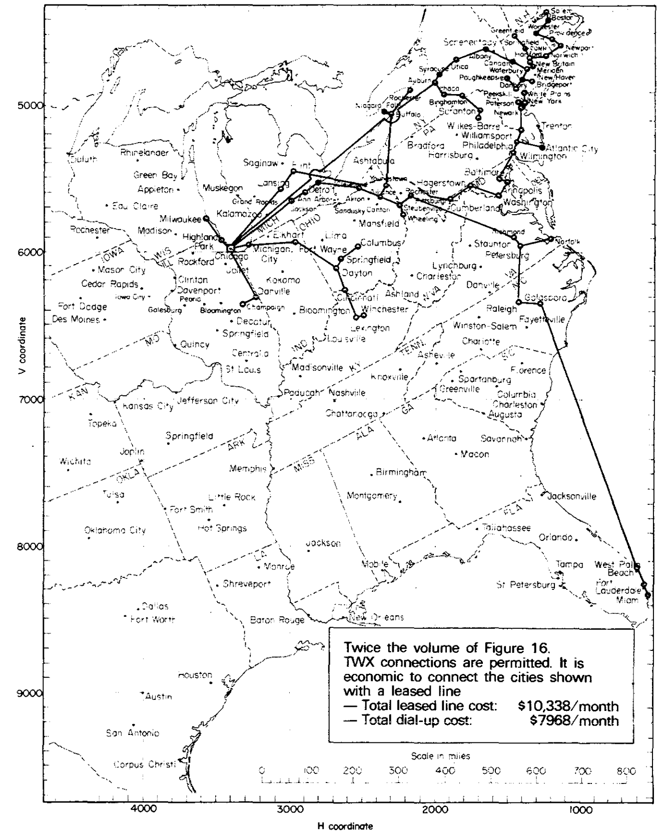


Figure 17. The lowest-cost network with traffic volumes twice those of Figure 16

```
    IF D> 430 THEN DO;DCOST=.95 +.30*CEIL(TIME);GO TO B2;END;
    IF D> 354 THEN DO;DCOST=.85 +.25*CEIL(TIME);GO TO B2;END;
    IF D> 292 THEN DO;DCOST=.80 +.25*CEIL(TIME);GO TO B2;END;
    IF D> 244 THEN DO;DCOST=0.70+.25*CEIL(TIME);GO TO B2;END;
    IF D> 196 THEN DO;DCOST=0.70+.20*CEIL(TIME);GO TO B2;END;
    IF D> 148 THEN DO;DCOST=0.70+.20*CEIL(TIME);GO TO B2;END;
    IF D> 124 THEN DO;DCOST=0.65+.20*CEIL(TIME);GO TO B2;END;
    IF D> 100 THEN DO;DCOST=0.60+.15*CEIL(TIME);GO TO B2;END;
    IF D>  85 THEN DO;DCOST=0.55+.15*CEIL(TIME);GO TO B2;END;
    IF D>  70 THEN DO;DCOST=0.50+.15*CEIL(TIME);GO TO B2;END;
    IF D>  55 THEN DO;DCOST=0.45+.15*CEIL(TIME);GO TO B2;END;
    IF D>  40 THEN DO;DCOST=0.40+.10*CEIL(TIME);GO TO B2;END;
    IF D>  30 THEN DO;DCOST=0.35+.10*CEIL(TIME);GO TO B2;END;
    IF D>  22 THEN DO;DCOST=0.30+.10*CEIL(TIME);GO TO B2;END;
    IF D>  16 THEN DO;DCOST=0.25+.05*CEIL(TIME);GO TO B2;END;
    IF D>  10 THEN DO;DCOST=0.20+.05*CEIL(TIME);GO TO B2;END;
                      DCOST=0.15+.05*CEIL(TIME);
B2:
```

IN FORTRAN:

```
 DIMENSION DISTB(21),PRICEB(21),PRCNTB(21)
 DATA DISTB /1910.,1360.,925.,675.,430.,354.,292.,244.,196.,148.,
1           124.,100., 85., 70., 55., 40., 30., 22., 16., 10.,
2            0./,
3      PRICEB/1.35,1.25,1.15,1.05, .95, .85, .80, .70, .70, .70,
4            .65, .60, .55, .50, .45, .40, .35, .30, .25, .20,
5            .15/,
6      PRCNTB/ .45, .40, .35, .35, .30, 3*.25, 3*.20, 4*.15, 3*.10,
7            3*.10/,

C
      DO 50 I=1,23
      IF (D.LE.DISTB(I)) GO TO 50
      DCOST=PRICEB(I)+PRCNTB(I)*(MINS-3)
      GO TO 100
   50 CONTINUE
  100 CONTINUE
```

For leased interstate telephone lines with conditioning:

```
    IF D>500 THEN DO;
    LCOST=745.0+CEIL(D-500)*0.75; GO TO B1; END;
    IF D>250 THEN DO;
    LCOST=482.5+CEIL(D-250)*1.05; GO TO B1; END;
    IF D>100 THEN DO;
    LCOST=257.5+CEIL(D-100)*1.50; GO TO B1; END;
    IF D>25  THEN DO;
    LCOST=100   +CEIL(D-25 )*2.10; GO TO B1; END;
    LCOST=25    +CEIL(D)*3;
B1:
```

In FORTRAN:

```
      DIMENSION DISTA(5),PRICEA(5),PRCNTA(5)
      DATA DISTA /500.0,250.0,100.0, 25.0,  0.0/,
     1     PRICEA/745.0,482.5,257.5,100.0,25.0/,
     2     PRCNTA/  .75, 1.05, 1.50, 2.10, 3.00/
C
      DO 20 I=1,5
      IF (D.LE.DISTA(I)) GO TO 20
      LCOST=PRICEA(I)+(D-DISTA(I))*PRCNTA(I)
      GO TO 30
   20 CONTINUE
   30 CONTINUE
```

The program that was used reads the cards prepared for the CNDP.

Its results indicate that it would pay to have 99 of the 187 terminal locations on TWX lines rather than on leased lines at current rates (Table 2).

The resulting network produced by CNDP for the terminals remaining on leased lines is as shown in Figure 17.

Planning Considerations for Dial-Up Versus Leased-Line Facilities

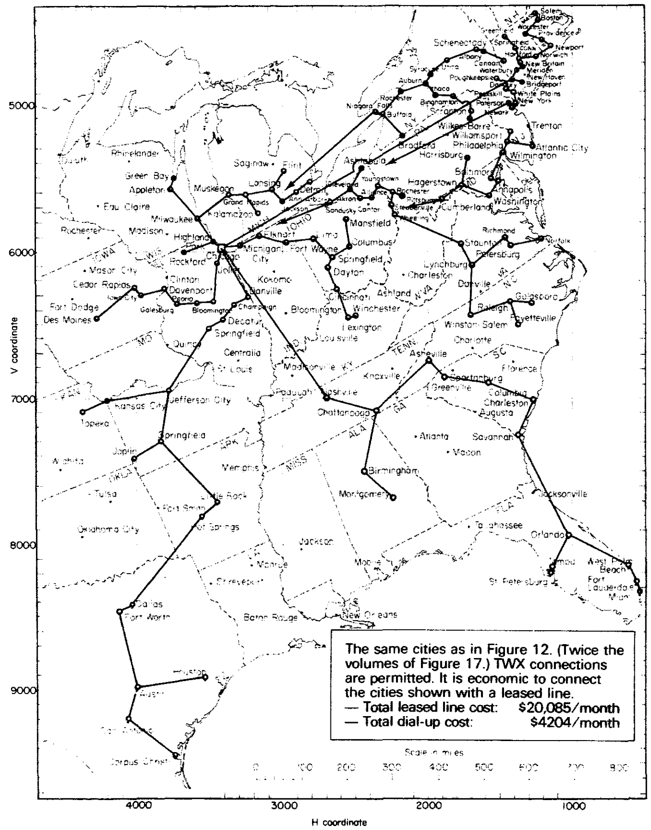


Figure 18. The lowest-cost network with traffic volumes twice those of Figure 17

The leased line costs are $10,338 per month, as opposed to $25,890 for the network in Figure 12. The total TWX cost is $7968 per month, so the saving is worthwhile.

If the traffic volumes were lower, then the cost calculation would favor dial-up lines more strongly. Figure 16 shows the result with half the traffic volumes. There are now only two lines. The leased-line cost has dropped to $5862 per month. The dial-up cost is $6938 per month.

The leased line in Figure 16 passed almost directly over some cities that are not connected to it, such as Richmond, Goldsborough, and West Palm Beach. If we could connect these cities to the line, then we could save the dial-up charges for these cities. Indeed, we can now take another look at any of the cities close to the leased-line path. We may not be able to connect them because the leased line may be fully loaded. Let us see.

There are 23 terminals on the line, including several at Chicago. Adding up the input and output load from all these terminals, we find that the total load on the line is 161.03 messages per hour. If we contemplate adding one of the terminals, our traffic rate table (Table 2) tells that the maximum load permissible for 24 terminals is 228.8 messages per hour. We should start with a location close to the line and as far from the computer counter as possible. West Palm Beach is the best candidate. This location has a traffic load of 7.19 messages per hour, input and output. It was allocated to a dial-up line because this load is relatively small. If we connect it to the leased line that happens to pass close to it, the total load on this line becomes 168.22, which is acceptable.

We can contemplate adding one more city—that will give the maximum number of terminals permissible on the line. Our traffic rate table says that the maximum load for 25 terminals is 225.1. Goldsborough has 5.13 messages per hour. Richmond has 9.45 messages per hour. The cost saving of dial-up calls would be greater if Richmond were added.

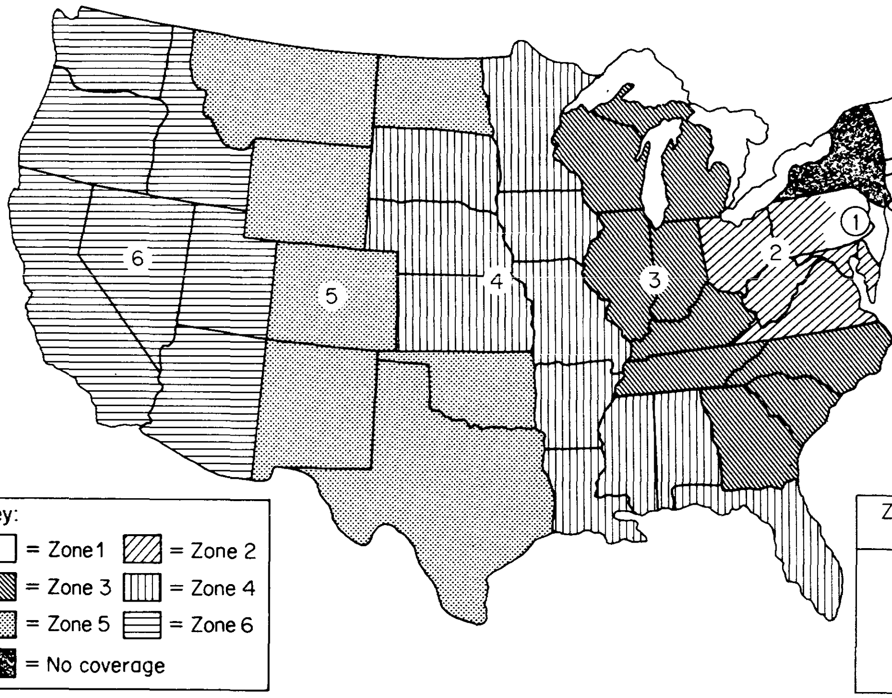
Similarly, we can make a minor manual adjustment to the network designed by the programs. When the traffic volume is doubled, the program allocates 143 terminal locations to leased lines with 44 using dial-up. The resulting line configuration is shown in Figure 18. The monthly cost of the leased-line configuration is now $20,085, as opposed to $27,023 when all the cities were on leased lines. The total monthly TWX line bill is $4204. Consequently, there has still been a worthwhile saving, but substantially less than in the previous cases.

When the traffic is eight times the volume listed, the program allocates all terminal locations except two to leased lines. Only Wichita, Kansas, and Jackson, Miss., are allocated to dial-up lines, both cities having very low traffic volumes and being fairly distant from any other city on a leased line. The saving to be made from having these two cities connected separately is slight, perhaps nonexistent, when the cost of a line adaptor for dial-up lines at the computer is considered. In this case, a network may well be used in which all terminals are on the leased lines. The line configuration and cost will then be as shown in Figure 14.

There are many variations on this theme. Some systems have no need for immediate processing of transactions, and transactions can be saved up at the terminals for periodic transmission. This factor could lower the TWX cost, or other dial-up line cost, because the first minute or three minutes must be paid for regardless of whether they are used or not.

Again, the dial-up lines may not go directly to the computer, as in the illustration in this section. They may go to concentration points connected to the computer on leased lines—possibly leased voice-grade lines with multiplexing.

Planning Considerations for Dial-Up Versus Leased-Line Facilities



Zone	Cost ($ per month)
1	500
2	1000
3	1300
4	1700
5	1900
6	2250

Figure 19. WATS zone configuration for southeast New York subscriber

WATS LINES

Considerations similar to those discussed in this report apply to WATS lines. As can be seen from Figure 11, a WATS line to a zone is usually slightly more expensive than a leased voice line. In a few areas, at large distances, the WATS line is cheaper than the leased line.

WATS calculations are complicated by the fact that the WATS line can serve many different locations. The designer must be concerned with the grade of service that he requires, or the probability of failing to obtain a connection.

The question then arises: "Should the zone 6 lines, for example zone 6 in Figure 19, serve zone 5 also, or should zone 5 have its own lines?" If zone 5 has its own lines, they are lower in cost than if the zone used zone 6 lines—in Figure 19 zone 5 lines are $1900 per month, compared to $2250 for zone 6. However, if zone 5 and zone 6 are combined, fewer lines will be needed to achieve a given grade of service because of the larger grouping. Which, then, is cheaper? It depends on the traffic volumes in zones 5 and 6, and only a detailed calculation can provide an answer.

If we now consider zone 4 also, there are five different possible arrangements:

1. Handle all three zones with the same zone 6 line group.

2. Handle zones 4 and 6 with the zone 6 line group and zone 5 with a zone 5 line group.

3. Handle zones 5 and 6 with the zone 6 line group and zone 4 with a zone 4 line group.

4. Handle zones 4 and 5 with a zone 5 line group and zone 6 with its own line group.

5. Handle each zone with it own line group.

Again, only a detailed calculation with the traffic volumes for each zone will reveal which of these alternatives gives the required grade of service at the lowest cost.

If we consider all six zones there are many different combinations possible. A computer program is needed which takes the traffic volumes in erlangs for each zone and does a calculation for each possible combination. The combination with the lowest resulting cost would be used.

In performing such a calculation, the resulting cost will be substantially lower if the grade of service (probability of line busy signal) is set at, say, 0.05 rather than 0.005 because more lines will be needed. However, a grade of service in which 5 per cent of the callers receive line busy signals would not be acceptable for most systems. The possibility arises of designing the equipment so that it automatically dials

Planning Considerations for Dial-Up Versus Leased-Line Facilities

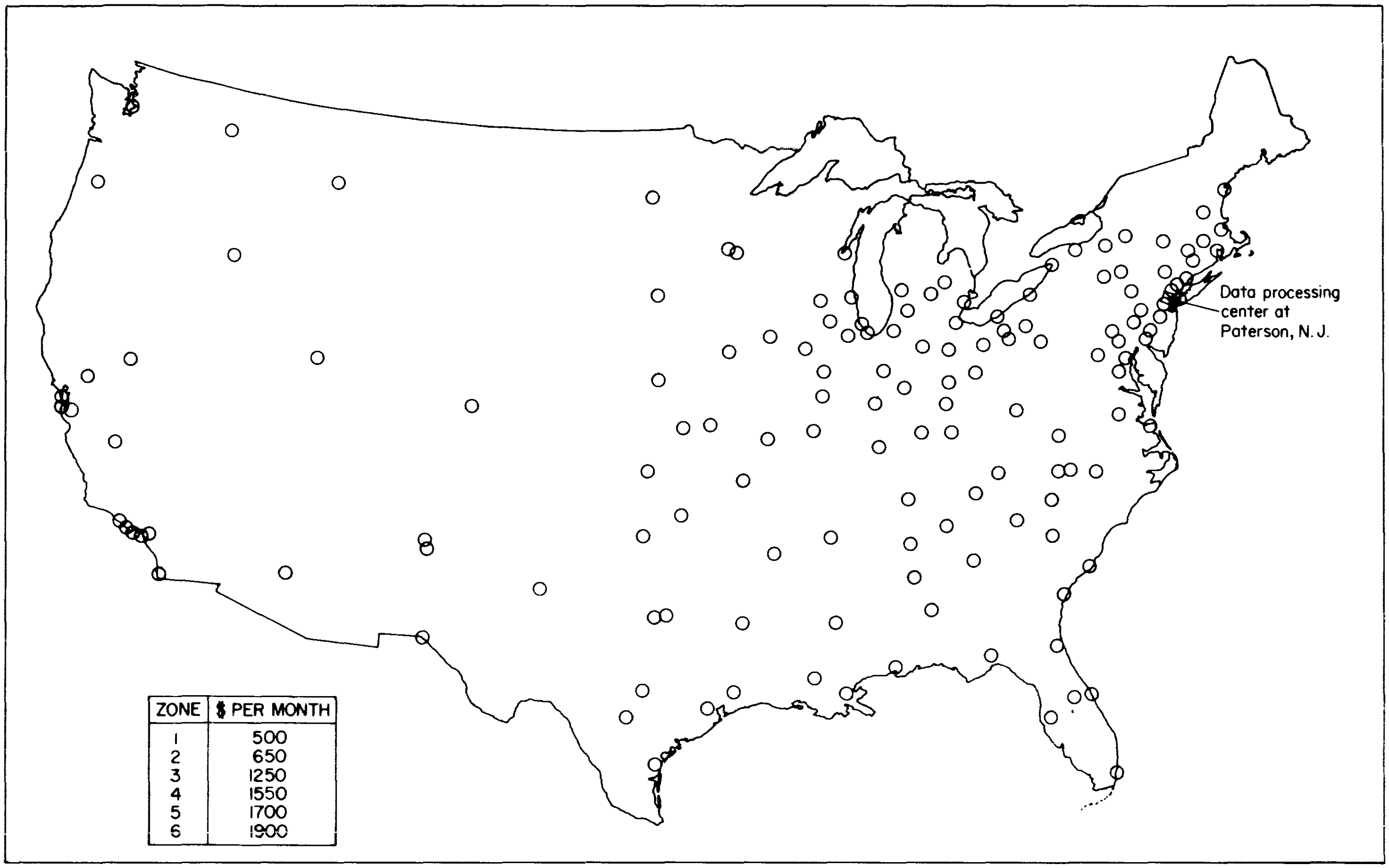


Figure 20. The 159 terminal locations marked are to be connected to the data processing center at Paterson, N.J., using the WATS tariff

on the public network when a busy signal is received from the WATS network. The cost of combined WATS and direct distance dialing would then have to be evaluated, and a grade of service for the WATS network that would minimize the combined cost determined.

Example 4. Design of a WATS Network. *Various locations throughout the United States must be given facilities to dial a location of Paterson, N.J. The calls may, for example, be telephone enquiries to a room of terminal operators. (The same calculation would apply to data traffic or any other form of traffic on voice-grade lines.) The traffic volumes in erlangs from the New Jersey WATS zones are estimated to be as follows:*

Zone 1:	*7.29 erlangs*
Zone 2:	*3.58* erlangs
Zone 3:	*8.73 erlangs*
Zone 4:	*5.77 erlangs*
Zone 5:	*3.88 erlangs*
Zone 6:	*5.95 erlangs*
Intrastate traffic:	*1.22 erlangs*

The grade of service required is 0.05. What combination of dial facilities will give the lowest cost?

The monthly costs of the WATS lines from New Jersey are as follows:

Zone 1: $500
Zone 2: $650
Zone 3: $1250
Zone 4: $1550
Zone 5: $1700
Zone 6: $1900

Because the points of origination of the calls are not stated, the cost of telephoning on the direct distance dialing network cannot be calculated exactly. An approximate calculation indicates that the cost will be between $200,000 and $300,000. This is much higher than the cost of using WATS.

Let us first calculate the cost of using a separate group of WATS lines to each zone. We will use the equation

$$P_B = \frac{\frac{(M\rho)^M}{M!}}{\sum_{N=0}^{M} \frac{(M\rho)^N}{N!}} \qquad \text{Eq. 3}$$

Where: $Mp = E(n)E(t_S)$ = number of erlangs of traffic
M = the number of lines, and

Planning Considerations for Dial-Up Versus Leased-Line Facilities

P_B = the grade of service (probability that all lines are busy) = 0.05.

Solving this equation, we find that the number of lines required to each zone are as follows:

Zone	Number of Lines	Cost ($ per month)
1	12	6,000
2	7	4,550
3	13	16,250
4	10	15,500
5	8	13,600
6	10	19,000

Four intrastate WATS lines are also needed to handle the traffic within New Jersey at a cost of $1300. The total cost therefore adds up to $76,200.

Now let us calculate the cost with a zone 6 line group handling all of the zones. This group will handle a total of 35.2 erlangs of traffic. From the above equation we can calculate that this needs 41 zone 6 lines, costing $77,900. Adding the intrastate line cost of $13,000, we have a total of $79,200.

There are many possible combinations of WATS zones. A computer program is needed to perform the above type of calculation for the various combinations. Running such a program, the result indicated that the lowest cost configuration was to have a zone 6 line group handling the traffic from zones 4, 5, and 6, and the other zones being served by unique line groups. The cost was as follows:

	Number of Lines	Cost ($ per month)
Zone 1	12	6,000
Zone 2	7	4,550
Zone 3	13	16,250
Zone 4	0	0
Zone 5	0	0
Zone 6	21	39,900
New Jersey		1,300
Total		68,000

The total cost, not including the cost of modems, is thus $68,000 per month.

Example 5. Choice of WATS or Leased Network. *The 159 terminal locations shown in Figure 20 are to be connected to Paterson, N.J. This time the terminal locations are known, so dial or leased lines can be used. The time associated with dialing is acceptable. The terminals operate at less than 150 bits per second.*

This example is an actual commercial network now in operation. The costs were evaluated for leased and dial links, both voice-grade and subvoice-grade, as follows:

Type of Line	Cost/Month
Subvoice-grade:	
Leased: AT & T type 1006 lines, point-to-point	$104,472
Leased: AT & T type 1006 lines, multipoint	20,100
Dial: TWX-CE network	20,626
Voice-grade:	
Leased: AT & T type 3002 lines, point-to-point	$153,552
Leased: AT & T type 3002 lines, multipoint	29,700
Dial: Public telephone network	45,978
Dial: WATS group of zone 6 lines	17,750
Dial: WATS unique line group to each zone	30,000
Dial: Lowest-cost combination of WATS groups	17,750

It will be seen that the WATS tariff gives the lowest cost—lower than subvoice-grade lines. The best WATS combination is that in which a group of zone 6 lines serves all the terminals.□

An Office Automation Study

Problem:

Everyone is talking about automating the office but very few are prepared to define the automated office, simply because it can vary drastically from one company to another or even one department to another. Office Automation may not be the responsibility of the communications manager, but communications (internal and external) will be the hallmark of the automated office, and the communications manager will have to be conversant with automated office terminology in order to make an effective contribution, as part of the team charged with automating the office.

This report describes an office automation study (and a resulting prototype system) which was conducted by IBM Corporation in cooperation with one of its large customers. It is presented here not as an archetype of such a study and resulting system, but simply as a case history of an early and reasonably successful attempt at office automation.

Solution:

The term "office automation" generally refers to the machine-aided creation, communication, storage, retrieval, and control of messages and documents handled by professional, clerical, and secretarial personnel in an office environment. Motivated by the conviction that "office automation demands that the new office machines be linked together to form integrated systems,"[1] an office study was begun in 1975 by IBM's Data Processing Division in conjunction with one of its customers. The purpose of the study was to investigate requirements for an integrated office communications system and to provide a framework for developing a prototype of such a system.

For this study, an office communications system was defined as a computer-based system that provides integrated facilities for the processing of business communications more efficiently and economically with little or no use of paper records. An objective of the study and prototype was to find ways to increase the return on the investment in office personnel by handling a broad class of functions through terminal work stations, with increased labor productivity and improvement in the quality of activities performed.

In this report, the study is first described. Included are the major factors involved and the resulting requirements that led to development of the prototype. In the discussion of the prototype which then follows, the objectives of the system, its setup and installation, and its operational characteristics are covered. Finally, some of the results derived from the prototype operation are presented.

"An Office Communications System" by G.H. Engel, J. Groppuso, R.A. Lowenstein, and W.G. Traub. Reprinted by permission from IBM Systems Journal, Volume 18, Number 3. © 1979 by International Business Machines Corporation.

[1]"The office of the future," *Business Week,* 48-83 (June 30, 1975).

Activities	Average percent of time*			
	Level 1	Level 2	Level 3	All
Writing	9.8	17.2	17.8	15.6
Mail handling	6.1	5.0	2.7	4.4
Proofreading	1.8	2.5	2.4	2.3
Searching	3.0	6.4	6.4	5.6
Reading	8.7	7.4	6.3	7.3
Filing	1.1	2.0	2.5	2.0
Retrieving filed information	1.8	3.7	4.3	3.6
Dictating to secretary	4.9	1.7	0.4	1.9
Dictating to a machine	1.0	0.9	0.0	0.6
Telephone	13.8	12.3	11.3	12.3
Calculating	2.3	5.8	9.6	6.6
Conferring with secretary	2.9	2.1	1.0	1.8
Scheduled meetings	13.1	6.7	3.8	7.0
Unscheduled meetings	8.5	5.7	3.4	5.4
Planning or scheduling	4.7	5.5	2.9	4.3
Traveling outside HQ	13.1	6.6	2.2	6.4
Copying	0.1	0.6	1.4	0.9
Using equipment	0.1	1.3	9.9	4.4
Other	3.1	6.7	11.4	7.7
	100	100	100	100
Total number of principals	76	123	130	329

*Level 1 represents upper management.
*Level 2 represents other managers and management-equivalent personnel.
*Level 3 represents nonmanagerial personnel.

Table 1. Principal activity summary

The office study

Our study partner was a multinational corporation with a consumer and industrial product line. The study site was the corporate and divisional headquarters for this company, where over 1,700 people were employed in a traditional decentralized administrative environment.

In the early stages of the study, it became evident that if we were to define a system that would help the office, it was important to know what people did with their time. Thus, the activities of three groups of employees were examined: secretaries, clerical workers, and principals (which included both managerial and professional exempt personnel). People in each group were given questionnaires that asked them, among other things, to estimate the amount of time they spent in various activities. These estimates are summarized in the tables that are included. A word about each may be useful in order to understand some of the requirements that would be derived for this system.

Activities among principals generally appeared to be consistent with the levels of the people responding (see Table 1). For example, scheduled meetings, unscheduled meetings, and travel, with 13.1, 8.5, and 13.1 percent, respectively, were very prominent activities among upper-management respondents. In contrast, more administratively oriented activities, such as filing, searching, and retrieving, were more evident among nonmanagement people (13.2 percent).

Upper-level management appeared to avoid administrative work and concentrate on communications-oriented activities. Their ability to avoid administrative work appeared to depend on the degree of support they received from subordinates and on the percentage of their work delegated to these people.

Even so, there was evidence that even more work could have been delegated if the proper people or systems were available. When asked if there were tasks that they do now that others could do for them, 51 percent of the principals indicated that they had one or more such delegable tasks. At the time this study was made, principals were spending 14 percent of their work month doing tasks that, in their opinion, others could do for them. Our analysis of these tasks showed that many of them could be done by less highly compensated people and involved such activities as copying, calculating, assembling data, typing, covering telephones, and filing and retrieving documents. In many cases (45 percent of the time), the tasks would have to be done by other principals. However, trained secretaries or clerical workers could do 55 percent of the delegable tasks.

This led to an analysis of secretarial and clerical activities to determine what these people were doing and where savings might be made.

Among the secretaries, typing was by far the number one activity, but it varied with the number of principals supported by a secretary (see Table 2). Thus, private secretaries supporting a single professional estimated that they spent only 26 percent of their time typing. By contrast, secretaries who supported more than four principals estimated that 45 percent of their time was devoted to typing.

With the extra time available to them, it appears that private secretaries do more administrative work, such as conferring with their manager, keeping calendars, taking shorthand, and handling mail.

On the other hand, secretaries with heavy typing workloads had time for little else. To the extent that their typing productivity is improved, they will have more time available for administrative support of principals or more time for typing—if that's what's needed.

An Office Automation Study

Activities	Average percent of time
Writing	3.5
Mail handling	8.1
Bulk envelope stuffing	1.4
Collating/sorting	2.6
Proofreading	3.9
Reading	1.7
Typing	37.0
Telephone	10.5
Copying or duplication	6.2
Conferring with principals	4.3
Taking shorthand	5.5
Filing	4.6
Pulling files	2.8
Keeping calendars	2.6
Pick-up or delivery	2.2
Using equipment	1.3
Other	2.0
	100
Total number of secretaries	123

Table 2. Secretarial activity summary

Activities	Average percent of time
Filling out forms*	8.3
Writing*	7.3
Typing*	7.8
Collating/sorting*	5.2
Checking documents*	10.4
Reading*	2.9
Filing†	5.9
Looking for information†	10.2
Telephone	9.2
Copying or duplicating	3.9
Calculating	10.3
Meetings	1.9
Pick-up or delivery in HQ	0.8
Scheduling or dispatching	1.2
Using a terminal	6.3
Other	8.4
Total	100
Total number of clerical personnel	115

*Primary paper-handling activities (41.9 percent)
†Secondary paper-handling activities (cumulative total = 58 percent)

Table 3. Clerical activity summary

Clerical activities did not fit any single pattern, other than to show that at least 41.9 percent and as much as 58 percent of the time is spent in paper handling (see Table 3).

Information Flow

In addition to understanding what people did with their time, a considerable amount of study activity was devoted to understanding the paperwork process within the company. For example, conventional business correspondence—letters and memos—were sampled to determine where they were coming from and where they were going. The results are shown in Table 4. The pattern that emerged, on both the incoming and outgoing side, was that a substantial amount of the paper stayed within the company: 75 percent of the incoming letters and memos originated within the company and 81 percent of the outgoing documents remained within the company.

A lot of time was spent in copying, and we found that for each original, six copies were made on average. Most of these were machine copies and the time devoted to making them was very much disliked by the secretaries. It was also partly unproductive time in that it involved traveling to and from the copiers, waiting for them to become available, and not infrequently finding them to be out of service.

Although the data collected provided an interesting statistical picture of the company studied, a large part of our understanding of the organization, and the direction in which it wanted to move, came from interviews with employees, who ranged from clerical personnel and secretaries to corporate executives.

Study Objectives

From the management team, in particular, we were able to derive a set of company office system requirements or objectives that were to guide us throughout the study. These objectives were:

1. Increase professional and managerial productivity.

Copies of outgoing documents	First two combined	First three combined	First four combined
24% to dept.files			
19% routed in dept.	43%		
24% other HQ dept.		67%	
14% other company locations			81%
19% outside company			
Incoming letters and memos			
3% from same dept.			
14% other HQ depts.	17%		
58% other company locations		75%	
25% outside company			100%

Table 4. Document flow

An Office Automation Study

2. Grow in stages.

3. Fit within the existing organization.

4. Tie in to data processing applications.

In deriving the first objective, we found that the exempt population at the customer site was twice as large as the nonexempt population, which was composed mainly of secretaries and clerical workers. The customer was at a point where the exempt staff output was not increasing at the rate that other areas in the company were increasing. Something was needed—equipment, procedures, other motivators—that would allow these people to become more productive.

This reason was the primary motivation for this customer's interest in office communication systems. Not only were there twice as many exempt employees as others, but their total cost was more than four times the nonexempt labor cost. Thus, even small productivity gains might have had high value to the customer.

This is not to say that the customers were not interested in improving secretarial productivity; they were, but not as an end in itself. Thus, they accepted the idea that to increase principal productivity, it might be necessary to first improve that of the secretaries so that they could provide better support to principals by either handling an increased typing workload or accepting more administrative work.

For the second objective, we found that to achieve maximum value from an office communications system, almost all company locations would need access to it, and most employees in those locations would have to become users. Great risk would thus be involved. Not only might such a system be potentially expensive, but the application itself was untried and untested and the question of user acceptance of an office system had not been answered.

The customer was committed to minimizing these risks and felt that an office system would have to grow in stages, starting in departments with potentially high value before moving to other departments and locations. In this way, the financial risk would not only be minimized, but as the state of the art advanced, the company would be able to take advantage of new technological breakthroughs as they occurred, provided the communications interface was defined.

Concurrent with the concept of growing in stages, the customer was also anxious to avoid disruption of their established organization by introduction of office communications system concepts. As a matter of philosophy, they wanted the system to fit the users and not the reverse, thus providing the third objective. Practically speaking, they wished to avoid clustering the secretaries into groupings simply because of the physical limitations of the machines (such as cable lengths) or because the economics of the system required grouping in order to reduce work station costs.

This desire did not mean that the customer would not accept a reorganization in order to achieve efficiency of operation; however, it was clearly indicated that such a reorganization would be a by-product of an office communications system, rather than a prerequisite to such an installation.

In deriving the fourth objective, we observed that this customer was aware, as are many others, that employees who were not working in data processing were becoming more and more familiar with computer concepts. Many of these employees were in touch with data processing daily. At the very least, they were providing input to or receiving output from computers.

The real revolution, however, was occurring among a smaller group of noncomputer professionals who were doing terminal data entry or inquiry. They were the forerunners of the on-line office system users.

They, and their managers, realized that office communications, to be meaningful, would involve not only access to text documents, but also integration with data processing applications as well. The more prescient users could foresee a single user interface to all computer applications. At the very least, each user would have a single physical terminal connected to all systems.

With the management requirements as a framework, the study team proceeded to define application requirements for end users of an office communications system. These requirements would become the basis for system design and development for a prototype office communications system. The major functional areas are now described.

Document capture—So-called image documents—material originating outside of the system—must be converted into system documents via scanning devices. Incoming mail, magazine articles, photographs, handwritten notes, charts and graphs—anything that a user wishes to preserve in noncoded form—should be entered using scanning devices.

Document creation—Entry, edit, correction of text, and limited (line and character) graphics are required for the preparation of correspondence, reports,

An Office Automation Study

forms, and other basic business documents. This operation could be done either interactively (via a display) or off line via a text-editing unit (for example, a magnetic card device) with batched input to the system.

Forms, such as check requisitions, expense accounts, personnel change authorizations, etc., should be stored internally so an authorized person could display them and fill in the blanks. If the form is for intraheadquarters use, such as a requisition, it should be transmitted to the proper receiving department upon request without hard-copy output at the originating station. A system facility for the definition of such forms is needed.

Distribution and receipt—This facility should provide for the distribution of correspondence to an electronic "mail box" of designated recipients. Documents may be those created via the system or captured by scanning. Functions would include logging of mail, control of the status of work-in-process (if a task is interrupted or passed from one user to another) and disposition. Security would be provided for control over document access, modification, filing, duplication, destruction, etc. It would include audit trails of access, modification, receipt, and duplication. Documents may be "distributed" for formatting as hard-copy output for destinations external to the system.

File, search, and retrieval—Electronic filing of documents during active use should include capabilities for creating multiple personal files (folders), establishing descriptor indexing schemes (e.g., date, originator), or full text automatic indexing.

The system would actually retain only one copy of a completed document. Documents could be moved automatically to lower-cost archival storage when current need was ended.

Provision for identification of desired documents could be done via indexes, or keywords, alone or in combination, through search queries. The document could then be accessed and reviewed.

Format and output—Where hard copy is desired, formatting may be done. Formats should contain headings, footings, prestored units (paragraphs, addresses, etc.) and computed fields. Output may be directed to any appropriate printer—local, subsystem, or system, proof quality or finished copy, character or image. Multiple copies may be produced including copies of external documents that were captured via scanning.

Personal services—Such services include the following:

- Follow-up files. Follow-up (or action files) may be kept by any user. Entries into such a file are typically due to receiving or distributing a document. Part of the filing may be any entry into the follow-up file with the system automatically supplying an action due date or the user keying an action due date. The user can also create entries in the follow-up file that are not associated with documents (e.g., meeting dates, appointments, etc.). The user may then, at any time, cause an automatic display or printing of the action file by due date, all future dates, this date, etc.

- Instruction and prompting. The system must support the self-training of users via computer-assisted techniques. Prompting facilities must be available to permit predefinition of procedures required to perform any well-defined tasks. Aids for helping—error diagnosis, tutorials—must also be an integral part of the system interface to the user.

- Calculation. The system should include calculator capabilities to allow definition by the user of such items as number of memories, more complex functions (such as square root, interest compounding, percentages, etc.). The most common functions will be predefined within the system. This capability should also be used with sets of numbers entered as text, in order to add or subtract them or verify previously entered totals.

- Automated correspondence. The system must allow for the easy creation of form letters, letters composed of standard paragraphs, and letters automatically generated when a specified event occurs (e.g., a new employee hire).

- List creation. The system requires facilities to allow the user to define, maintain, and use various lists (such as mailing lists and telephone directories).

System services—These services include the following:

- Security. A security scheme is required to protect the system from access by unauthorized persons. Because the system will store sensitive information (personnel data, profit figures, contract backup), additional protection will be needed to control access to this information by legitimate system users.

- Accounting. Extensive facilities are needed to collect data about system use so that costs can be determined and properly allocated among users.

An Office Automation Study

- Data base interface. A capability is required to permit authorized users to extract structured data from data processing files, to pass updated and new information from the office system to data processing files, and to execute data processing programs.

An office communications system prototype

Our ability to develop a system that met the requirements in detail as specified in the results of the joint study varied considerably. For some requirements, the state of the art had just not advanced to the point where a practical system solution was possible. For other requirements, the technology was available, but only at prohibitively high costs.

As extensive as our study had been, we also began to find that there was much more to be investigated. We were at a point where we could either do more studying or do something else, and the choice was for something else—a prototype system.

Our primary objective was to develop, based on some subset of the requirements derived from the joint study, an office communications system that could be used directly by and be of benefit to principals.

Not only did we want to give principals a system that they could use, we also wanted to give them something they would value. Our joint study had shown us that productivity of principals was the real interest of management. We also knew that we could increase secretarial productivity via improved typing systems and administrative aids. The question was whether we could do the same for principals.

This question and others needed answering, and a prototype system seemed to be an ideal way to help find the answers. Thus, a prototype would serve as an experimental learning system. It would allow us to validate existing requirements, develop new requirements, test human factors, and develop and evaluate tools and techniques for assessing value and usability to principals.

We did not know what effect a prototype system would have on an organization. We decided to make our own organization—headquarters of the IBM Data Processing Division—the site for our prototype test.

Test Site Selection

Our search for a suitable department for the test site was based on the following criteria:

1. A strong desire by the management of the department to participate in the test, including a commitment to cooperate by all department staff members.

2. A department composed largely of managers and professionals, supported by a secretarial staff.

3. A reasonable amount of communication among department managers and professional staff. This criterion was necessary in order to better evaluate the distribution functions of our prototype.

4. An ability to easily create document data bases that would later be accessed for search and retrieval.

We examined several departments at the headquarters location and finally determined that the Account Marketing Department would be the best site. The department consisted of approximately 50 people, headed by a director with five managers reporting to him. Secretarial support for the director, managers, and professionals was provided by eight secretaries. Account Marketing was then surveyed in more detail to better understand their "current system."

Principals were asked to fill in questionnaires and telephone logs. They were also interviewed to get a more in-depth understanding of their jobs and their expectations of an "office system."

Both at IBM and at the customer site, many of the activities of the principals were similar. Use of the telephone, attendance at scheduled meetings, and travel, which ranked highest for upper management at the customer, also were the highest-ranked activities among Account Marketing management. Other activities between the two groups also compared favorably, e.g., mail handling: customer, 4.4 percent, IBM, 5.6 percent; filing: customer, 2.0 percent, IBM, 1.7 percent, etc.

The secretaries were also surveyed. They were asked to estimate how they spent their time during the week, and many of the results closely matched those from the customer. For example, typing was the number one activity of both groups, 37.0 percent for the customer and 41.6 percent for IBM. The telephone ranked second with 10.5 percent for the customer and 11.1 percent for IBM. Ranking third with both was mail handling, which was 8.1 percent for the customer and 8.8 percent for IBM.

Installing the prototype

Upon completion of the survey, a specific group of people was selected to use the prototype. Figure 1 shows the Account Marketing configuration as originally installed.

An Office Automation Study

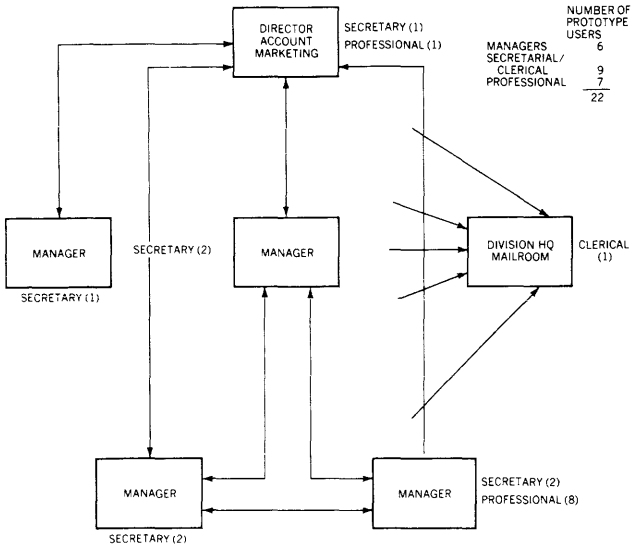


Figure 1. Initial Account Marketing installation configuration

The initial task was equipment procurement, which began in September 1977. The second task was modification of the facility. Thirty-eight coaxial cables were installed, running from the computer room to the Account Marketing Department. While the computer was being installed, the user's work area was redone, and work station equipment was installed. Communication features were added to existing IBM magnetic card typewriters, tables were procured on which to place IBM 3277 Display Stations; extension cords and electrical adpaters were added to accommodate the extra equipment. Extra telephones and their couplers were installed so that an IBM Communicating Mag Card/Selectric Typewriter (CMS/ST) could link to the host CPU.

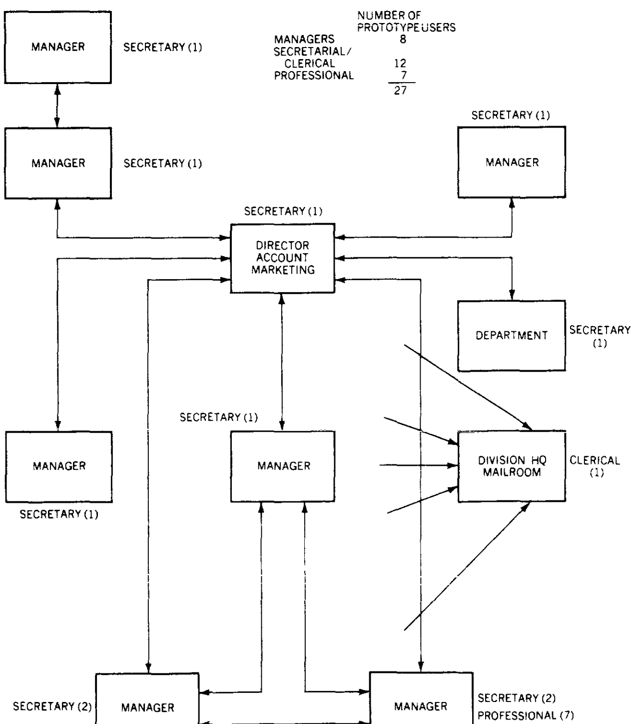


Figure 2. Final Account Marketing installation configuration

In retrospect, the time anticipated for the installation period was inadequate and problems were unforeseen. Possibly a distributed system would have been less subject to facility modifications, and the attendant cable stringing that the CPU installation required could have been reduced. Whatever the design, detailed plans must cover all aspects of the installation from CPU to end user to ensure a smooth evolution into an automated office.

Training

Training for the prototype users was conducted on an individual basis. Two two-hour hands-on sessions were conducted with the secretaries to familiarize each one with CMC/ST functions. Another two-hour session was used for instruction on the display station that complemented each secretarial work station.

The principals were treated less formally. Once the secretaries were trained and were using the system, the principals were given a one-hour hands-on session in their own offices.

All users were provided with user guides for reference.

Although procedures were discussed during the training period and an operational procedures manual was prepared, unanticipated problems occurred. Throughout the succeeding three months, sessions were held with the secretaries to discuss these operational and procedural problems. One combined meeting of principals and secretaries was also held to gather feedback from all users.

Prior to the operational period, all secretaries were requested to save the magnetic cards for the documents they created. These cards were loaded into the prototype and provided a base of six months of information with which to begin the test period.

On April 17, 1978 the system became operational, two weeks after it was first available. It remained in Account Marketing until December 1978.

As the test period progressed, many changes occurred in the Accounting Marketing Department. A new director was named, the organization grew from five managers to ten, the secretarial support increased from eight to twelve and had a turnover of 60 percent, and the original eight professionals had a turnover of 63 percent. Figure 2 represents the final Account Marketing configuration at the end of the test period.

An Office Automation Study

This turnover caused many unexpected operational problems but also provided an awareness of the dynamics of the office. Cables had to be restrung, new phones were installed, and electrical outlets were added to the area. These problems and the ones associated with training and education should be taken into account when planning an automated office.

As with any computer system, questions concerning the operation of the office system must be answered. What are the available hours? Who operates the system? To which organizational structure does it report? How long is information retained on line and off line?

The interview responses to the questions on availability had mostly to do with the hours of system availability rather than the inoperable time of the prototype. The users saw a need for an office system to be available from early in the morning to late in the evening. This time frame would not only accommodate the early and late workers but also make the system available for people in other time zones and for people who may travel or work at home.

Another area of operational concern is the protection of the information in the system. The prototype has the capability of setting parameters relative to (disk) space and time (days, months, years) that control the amount of information retained in the active (on-line) system. Once those thresholds of space and time are exceeded, documents (information) are rolled out of the active system to an archive (off-line tape storage).

A company-wide office system would have to take the information protection a step further. Information required by law to be retained would have to be designated and protected accordingly. Information necessary to reconstruct the business in case of disaster would also have to be distinguished and protected. Our prototype only provided the archive-level of protection described above. The whole area of records management and protection of corporate assets must be accommodated in any automated office.

As the prototype period progressed, it became evident that an office system does not operate on its own. An operations staff is required to perform the following duties:

1. Daily startup of the system.
2. System backup on a scheduled basis.
3. Recovery operations are required.
4. Operating system maintenance.
5. Office system maintenance.
6. Application code enhancements and testing.
7. User education and liaison.
8. Equipment coordination.

Because we were operating as a prototype installation and had a minimal operations staff, the above operational requirements became evident very quickly. Also at question was the organizational position of an office system operation. For example, in our divisional headquarters, the logical position might be in the organization that supports the branch offices. This organization has responsibility for all computers, a majority of the programmers and analysts, and the operations staff to support the division from an administrative perspective. Since they already provide a support function for many departments, they could be the logical choice to have the office systems responsibility.

Functional description of prototype system

The office communications system prototype is an internal program designed to provide managers and nonmanagerial professionals with an easy, fast, and direct method for handling their business communications. More specifically, it was to give the ability: to look at their mail, to search for and retrieve documents (memos, reports, messages, etc.), and to print, file, suspend, circulate, or pass documents stored in electronic or paper files by date/date ranges, originator, addressee, and keywords by using a display station through a channel connected to a System/370. Figure 3 illustrates the prototype configuration.

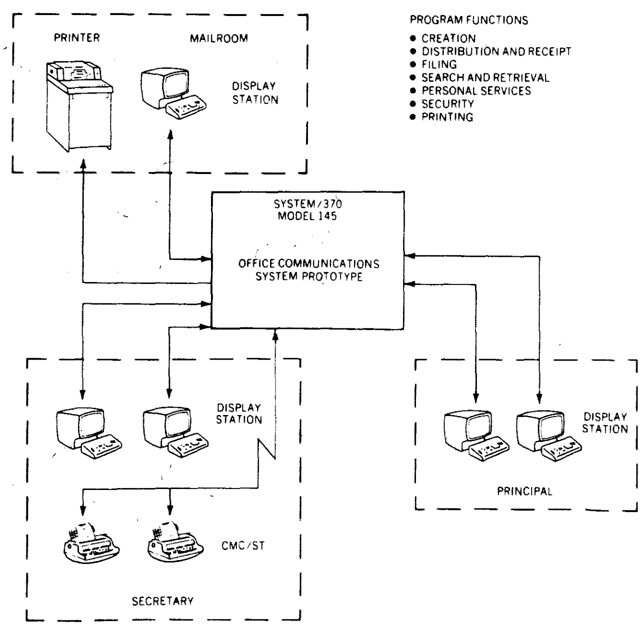


Figure 3. Prototype configuration

An Office Automation Study

```
                         INITIAL DISPLAY          OCTOBER 3, 1977
                                                  2 USERS AT 9:45
                                                  1 IN A REQUEST

DEPRESS FUNCTION KEY TO SELECT NEXT ACTION:

    KEY 1 - FILE SEARCH (RETRIEVE)
        2 - ENTER INFORMATION (MEMO)
        3 - REVIEW SCHEDULE
        4 - DISPLAY SUSPENSE FILE (HOLD QUEUE)
        5 - REVIEW INCOMING COMMUNICATIONS (MAIL QUEUE)
        6 - REVIEW DATA PROCESSING REPORT
        7 - TERMINATE SESSION (SIGN OFF)

DEPRESS ENTER TO MOVE TO NEXT FRAME FOR ADDITIONAL FUNCTIONS.
```

(A)

```
                 EXTENDED FUNCTION DISPLAY

DEPRESS FUNCTION KEY TO SELECT NEXT ACTION:

    KEY 1 -
        2 - INDEX EXTERNAL DOCUMENT (DCR)
        3 -
        4 -
        5 - RESET TO INITIAL DISPLAY
        6 -
        7 - TERMINATE SESSION (SIGN-OFF)
        8 -
        9 - DEFINE USERS, MAIL ADDRESSES, DISTRIBUTION LISTS, KEYS
       10 - UPDATE SCHEDULE
       11 - MAILROOM PROCESSING
```

(B)

Figure 4. Selection menu: (A) initial display (B) extended function display

With the display station, the managers or professionals can perform all functions by selecting the action desired from a list of appropriate options displayed on the initial screen (see Figure 4) or on the lower part of other screens. They can generally use the program function keys, or for some functions, key in the information to be communicated or retrieved. There are no complex commands to learn.

The prototype also provides secretaries with the capability of performing all of the above activities, as well as with additional "support" functions such as entering documents and messages into the system from magnetic cards for electronic storage and distribution, defining and maintaining user records, distribution and circulation lists, and calendars/schedules, and handling print and retrieval requests. These support functions can be done using a display station or a dial-up CMC/ST.

Each of the basic and supporting functions is discussed in greater detail below.

Creation and Entry of Documents

The creation and entry of documents into the prototype is done in the following manner. First, documents, such as letters or memos, originated by principals are typed by secretaries on a CMC/ST and follow the normal procedure that results in hard-copy forms as well as having the documents recorded on magnetic cards.

Using the CMC/ST, the secretaries sign on to the prototype. In response to a set of system prompts, they enter the magnetic cards. After the cards for each document have been read, the system prompts for confirmation of information scanned automatically from the document, such as date, subject, addressee, and those to whom copies are to be sent. The system also requests keywords and special handling instructions (e.g., confidential, receipt required). When this is done, the system prints out a document number.

Systems users can also create memos on line by using the display station. The "Memo Entry" option prompts the user to enter similar descriptive information, memo content, and addressees' names.

Distribution

At the completion of the memo entry process, the system automatically distributes the document to each recipient, whether individually named or on a distribution list. For recipients on the system, a descriptive entry is immediately posted in their "mail queues." This posting is the equivalent of almost immediate delivery of documents to the recipient's desk.

For the many recipients not on the system, the document is sent to the mailroom queue to permit the document and a matching address label to be printed by mailroom personnel, who place the document in an envelope with the matching label and deliver it in the conventional manner. See Figure 5 for the screen on mailroom processing.

Filing of Documents

Besides distributing documents, the prototype automatically stores a permanent record of each

```
                     MAIL ROOM PROCESSING

                        QUEUED (AS OF 13:45)
   TO                     FROM            SPECIAL HANDLING
 1 MR. P.C. JONES         ANYBODY              -
 2 MR. A.P. MOLER         QUIXOTE              -
 3 MR. J. SMYTHE          SOUTHERN             -
 4 MR. E. S. EAST         BROWN                -
 5 MR. N. N. WEST         COMPASS          CONFIDENTIAL
 6 MRS. E. NORTH          GROSS            PERSONAL
 7 MR. A. EQUIVALENT      HOSTILE              -

ENTER PRINTER UNIT:                     KEY 1 = PAGE FORWARD
OPTIONAL NUMBERS:                           2 = PAGE BACKWARD
  (BLANK MEANS PROCESS ALL DOCUMENTS)       3 = PRINT ENVELOPE (LABELS)
                                            4 = PRINT DOCUMENTS
                                            5 = RESET TO PREVIOUS DISPLAY
DEPRESS KEY                                 6 = INTERRRUPT PRINTER
 TO SELECT                                  7 = TERMINATE
NEXT ACTION                                 8 = RELEASE PRINTED DOCUMENTS
                                            9 = PRINT DOCUMENT LETTERHEADS
                                           10 = PRINT DOCUMENT (PAGES 2 THRU END)
```

Figure 5. Mailroom processing

An Office Automation Study

document. This record, which is indexed for later retrieval by the originator and recipients, becomes a reliable substitute for the traditional hard-copy files. The indexing of documents by the descriptive parameters keyed during entry provides the ability to retrieve documents by originator to produce the equivalent of an originator's chronological file, and by keywords for a "project" or "cross reference" file.

Hard-copy documents received from external sources, and documents that contain other than text (pictures, graphics, line drawings) are indexed the same as electronically created and filed documents and can be "retrieved" by the same search methods. The only difference is that the documents are filed in an ordinary file as hard copy in sequence by the document number, which is assigned by the system when the document is indexed.

Each user of the system has a "mail queue." Distribution of documents is reflected by an entry in the mail queue of each of the system addressees or those to receive copies. This posting of the mail queue is done in real time, and therefore, the mail queue is a dynamic facility.

The primary device for mail review and other administrative activities is the display station. The principals or their secretaries "sign-on" at the display station. The system response is a menu containing a selection of actions to perform (see Figures 4A and 4B) such as "Display Incoming Communications." Selection of this action displays the mail queue.

The mail queue consists of a series of single line entries arranged in time-of-receipt sequence (see Figure 6). Each line contains the name of the originator for the document or message, the originator's organization, the origination date, the subject line, and indicators showing if the user is an addressee, whether it is "personal" or "confidential," or that a "return receipt" has been requested.

By depressing a program function key, the user may initiate selection of a specific item from the queue (Figures 7, 8A, and 8B). The process is tutorial in that the actions (and corresponding keys) appropriate to the task are shown on the screen.

When the user has finished reviewing a document from the mail queue, he can select one of several options related to the disposition of the document (see Figures 8 and 9).

One option is to place the document in a suspense or "hold" status for action at a later date. If the user selects the hold queue function, a suspense date can be selected for later follow-up. The document

```
A. B. ANYBODY                        MAIL QUEUE                        MAY 1, 1977

 01 JONES          68J/DPD  04/27/77 REQUEST FOR TRANSFER - J. BROWN        AP
 02 BRANDT       --PASSED--04/27/77 MESSAGE ATTACHED, ACT BY 05/10/77     A
 03 ABERCROMBIE   B134/SCD  04/28/77 MACHINE AVAILABILITY - KINGSTON CENTE
 04 HARDLINE      C071/CHQ  04/30/77 MANAGEMENT PRACTICES - APPRAISALS        C
 05 CROMWELL       61R/DPD  04/29/77 #77-16503 MONTHLY APAR REPORT          RA
 06 SNOWFLAKE      703/DPD  04/25/77 REDUNDANT CORRESPONDENCE               *R
 07 SHUTTER        701/DPD  04/30/77 RELEASE OF PHOTOGRAPHS TO MEDIA
 08 BARRISTER      68K/DPD  04/30/77 PRODUCT ANNOUNCEMENT - 3999
 09 RETRIEVAL      REQUEST  05/01/77
 10 OVERHOLDT      802/DPD  04/28/77 FE MAINTENANCE RESPONSIBILITIES
 11 QUACKER        705/DPD  04/28/77 S & A TIME STATISTICS                    C
 12 RETRIEVAL      END      05/01/77

FLAGS:                      DEPRESS KEY TO SELECT NEXT ACTION
  A=ADDRESSEE
  C=CONFIDENTIAL    KEY 1 = PAGE FORWARD              6 = SELECT DOCUMENT
  P=PERSONAL            2 = PAGE BACKWARD             7 = TERMINATE
  O=OPENED              3 = DISPLAY SCHEDULE          8 = RETRIEVE
  #=OFFLINE #           4 = DISPLAY HOLD QUEUE        9 = ENTER MEMO
  R=RECEIPT REQ'D       5 = RESET TO PREVIOUS        10 = DISPLAY REPORT FILE
 *R=RECEIVED                PROCESS               ENTER = REFRESH MAIL QUEUE
```

Figure 6. Mail queue

```
A. B. ANYBODY                        MAIL QUEUE                        MAY 1, 1977

 01 JONES          68J/DPD  04/27/77 REQUEST FOR TRANSFER - J. BROWN       APO
 02 BRANDT       --PASSED--04/27/77 MESSAGE ATTACHED, ACT BY 05/01/77     A
 03 ABERCROMBIE   B134/SCD  04/28/77 MACHINE AVAILABILITY - KINGSTON CENTE
 04 HARDLINE      C071/CHQ  04/30/77 MANAGEMENT PRACTICES - APPRAISALS        C
 05 CROMWELL       61R/DPD  04/29/77 #77-16503 MONTHLY APAR REPORT          RA
 06 SNOWFLAKE      703/DPD  04/25/77 REDUNDANT CORRESPONDENCE               *R
 07 SHUTTER        701/DPD  04/30/77 RELEASE OF PHOTOGRAPHS TO MEDIA
 08 BARRISTER      68K/DPD  04/30/77 PRODUCT ANNOUNCEMENT - 3999
 09 RETRIEVAL      REQUEST  05/01/77
 10 OVERHOLDT      802/DPD  04/28/77 FE MAINTENANCE RESPONSIBILITIES
 11 QUACKER        705/DPD  04/28/77 S & A TIME STATISTICS                    C
 12 RETRIEVAL      END      05/01/77


 DEPRESS KEY CORRESPONDING TO DESIRED DOCUMENT.
```

Figure 7. Mail selection

```
MAIL DOC. NO.    1204   WP                        RECEIVED 14:34 MAY 1, 1977
FILE KEYS - PRODUCT ANNOUNCEMENT;3999;DELAY;

Mr. A. B. Anybody          DPH61R  White Plains, NY

April 30, 1977
J. J. Barrister
Systems Management Department
68K/DPD - 1133 Westchester Avenue
White Plains, N. Y.

Product Announcement - 3999

The product announcement of the 3999 will be delayed until the required release
by the corporate legal office has been received.  We have no projected date for
receipt of the release.  Since several of your plans are dependent upon the
timely release of this device, you may wish to review your schedules and
commitments.

DEPRESS KEY                KEY 1 = PAGE FORWARD       KEY 6 = MOVE TO HOLD QUEUE
TO SELECT                      2 = PAGE BACKWARD          7 = RELEASE (FILE)
NEXT ACTION                    3 = DISPLAY SCHEDULE       8 = HARD COPY REQR'D
                               4 = DISPLAY HOLD QUEUE     9 = PASS FOR ACTION
PAGE 1 OF 1                    5 = RESET TO MAIL QUEUE   10 = CIRCULATE
```

(A)

```
If you, through your channels, learn of other impacts in relation to this
announcement, I would appreciate being advised.

                                                   J. J. Barrister

JJB/CMH
cc:  Mr. J. Q. Benefactor
     Mr. R. L. Judge




DEPRESS KEY                KEY 1 = PAGE FORWARD       KEY 6 = MOVE TO HOLD QUEUE
TO SELECT                      2 = PAGE BACKWARD          7 = RELEASE (FILE)
NEXT ACTION                    3 = DISPLAY SCHEDULE       8 = HARD COPY REQR'D
                               4 = DISPLAY HOLD QUEUE     9 = PASS FOR ACTION
PAGE 1 OF 1                    5 = RESET TO MAIL QUEUE   10 = CIRCULATE
```

(B)

Figure 8. (A) mail display (B) continuation of mail display

An Office Automation Study

```
                    PASS FOR ACTION PROCESSING                    MAY 1, 1977


DEPRESS KEY CORRESPONDING TO DESIRED NAME
       KEY 1 = STAFF, I. M.                 61RDPH WP
           2 = PLANNER, I. M.               B028CD WP   <
           3 = CHASER, A. B.                81NDPH WP
           4 = FRIDAY, M. G.                62ADPH WP
           5 = DISTRIBUTION-A
           6 = DISTRIBUTION-B
           7 = DISTRIBUTION-ALL
           8 =
           9 =
          10 =
          11 =
          12 =

PRESS ENTER TO END NAME LIST, OR TYPE "cancel", OR A NAME AND PRESS ENTER.
```

Figure 9. Pass-for-action processing—initial display

displayed is removed from the mail queue and placed in the hold queue in suspense-date (if any) sequence. This queue is similar in function, format, and content to the mail queue. Hold queue documents whose suspense date is the current or passed date are highlighted to bring attention to action that is due. The user may change the suspense date for any document as necessary.

Next, the user may "file" or "release" the document, having familiarized himself with the content and having no further reason to see the document. The document is already a part of the system "file" and will be retained by the system in current and archival storage to the limit established by the using facility.

The user is also offered the option to request a printed copy of the document displayed. Selection of this option sends a message to the secretary's mail queue requesting printout of the specific document. The secretary can then use the CMC/ST to print the document, and deliver it to the user.

The last option displayed is "pass for action," which gives the user the ability to pass on a document to others. When this option is selected, the user is presented with a list of names of people with whom the most frequent communications occur. The user can select any of the names listed, or can key in other names at the bottom of the screen (see Figure 9). After the selections are made, the system displays a list for pass-for-action messages for each name selected such as "for your information," "please see me on this item," "give me your comments," etc. In addition to these prestored messages, the user can key in a unique message (see Figure 10).

The user can enter an optional suspense date which will be added to the message shown such as illustrated in Figure 10: "Give me your comments (by 10/14/77)." Each pass-for-action recipient can be sent a different message, or the same message, as desired. A reminder is automatically inserted into the sender's hold queue showing to whom he

```
                    PASS FOR ACTION PROCESSING                    MAY 1, 1977

DEPRESS KEY CORRESPONDING TO THE DESIRED MESSAGE FOR I.M. PLANNER
IF THERE IS A "DUE DATE" TYPE IT IN "(BY XX/XX/XX)" AND PRESS ENTER

     KEY 1 = FOR YOUR INFORMATION
         2 = FOR YOUR ACTION                        (BY         )
         3 = PLEASE SEE ME ON THIS ITEM
         4 = PREPARE A RESPONSE FOR MY SIGNATURE    (BY         )
         5 = GIVE ME YOUR COMMENTS                  (BY         )
         6 = I WILL ATTACH MY INSTRUCTIONS ON THIS MATTER.
                PLEASE FOLLOW THROUGH               (BY         )


 OR, DEPRESS KEY TO                KEY 7 = DISPLAY SCHEDULE
 SELECT NEXT ACTION                    8 = REPEAT LAST MESSAGE
```

Figure 10. Pass-for-action processing—action message development

passed the document, the suspense date (if any) and the message (see Figure 11).

Retrieval of Filed Documents

Filed documents can be retrieved in two ways: (1) by document number, if known, and (2) by search parameters. "Parametric search" is a search for documents based on parameters such as date, originator's name, organization, addressee, and "file keys" or keywords (see Figure 12). The user can key in whatever parameters are known, e.g., "I would like to see all documents originated between June 28, 1977 and July 6, 1977, addressed to Mr. J.J. Brown about 3705 EP, 3705 NCU." The system scans all indexes, selects those authorized or public

```
A. B. ANYBODY                      HOLD QUEUE                       MAY 5, 1977

 01 JONES         68J/DPD  04/27/77       REQUEST FOR TRANSFER - J. BROWN
 02 BRANDT        81N/DPD  04/27/77 05/10 ACRE USAGE STATISTICS
 03 ABERCROMBIE  B134/SCD  04/28/77 05/10 MACHINE AVAILABILITY- KINGSTON CENTE
 04 HARDLINE     C071/CHQ  04/30/77 05/11 MANAGEMEMT PRACTICES - APPRAISALS
 05 CROMWELL      61R/DPD  05/01/77 05/11 #77-1201 MONTHLY APAR REPORT
 06 SNOWFLAKE  -REMINDER-  04/01/77 05/12 GIVE ME YOUR COMMENTS BY 05/12/77
 07 SHUTTER       701/DPD  05/02/77 05/12 RELEASE OF PHOTOGRAPHS TO MEDIA


SUSPENSE DATE                DEPRESS KEY TO SELECT NEXT ACTION
 SHOWN WHERE
 APPLICABLE    KEY 1 = PAGE FORWARD          KEY 6 = SELECT DOCUMENT
                   2 = PAGE BACKWARD             7 = TERMINATE
HIGH INTENSITY     3 = DISPLAY SCHEDULE          8 = RETRIEVE
 FOR SUSPENSE      4 = NOT USED                  9 = ENTER MEMO
 EXPIRED           5 = RESET TO PREVIOUS        10 = DISPLAY REPORT FILE
                       PROCESS               ENTER = REFRESH HOLD QUEUE
```

Figure 11. Hold queue processing

```
                         RETRIEVAL PROCESSING

SPECIFIC REQUEST
  DOCUMENT NO.:                                   <== NUMBER OR YY-NUMBER LOC.

INDEXED SEARCH
  DATE RANGE:      04/01/77;05/01/77              <== ONE DATE, OR DATE;DATE
  ORIGINATOR:                                     <== LASTNAME AND INITIALS
    ORGANIZATION:                                 <== OPTIONAL
  ADDRESSEE:                                      <== LASTNAME AND INITIALS
  FILE KEYS (0-8)  :                  :
                   :                  :
                   :                  :
                   :                  :

ENTER DOCUMENT NUMBER FOR A SPECIFIC REQUEST, OR
ENTER PARAMETERS FOR AN INDEXED SEARCH.  THIS WILL RETRIEVE THE AUTHORIZED OR
PUBLIC DOCUMENTS IN THE DATE RANGE WHEN THE ORIGINATOR MATCHES AND THE ADDRESSEE
MATCHES AND ANY FILE KEY MATCHES.  DOCUMENTS ARE NOT TESTED FOR A PARAMETER IF
IT IS BLANK.  TAB PAST FIELDS THAT DO NOT HAVE TO BE ENTERED.

PRESS ENTER TO BEGIN RETRIEVAL, OR
  DEPRESS KEY TO
  SELECT NEXT ACTION                   KEY 5 = RESET TO PREVIOUS PROCESS
                                           7 = TERMINATE
```

Figure 12. Retrieval processing—initial display

An Office Automation Study

```
                        RETRIEVAL PROCESSING

YOUR RETRIEVAL REQUEST HAS RESULTED IN 4 DOCUMENTS FLAGGED FOR
    RETRIEVAL.

This retrieval was for authorized or public documents in the date range JUNE 28,
1977 through JULY 6, 1977 and written to J. J. BROWN and having one of these
file keys:  3705 BP; 3705 NCU.

DEPRESS KEY TO SELECT  KEY 1 = CANCEL, REPEAT       KEY 5 = CANCEL,RESET TO
  NEXT ACTION                  PARAMETER ENTRY              PREVIOUS PROCESS

                           2 = CANCEL, BEGIN            7 = TERMINATE
                               PARAMETER ENTRY

                          10 = RETRIEVE THESE DOCUMENTS TO THE MAIL QUEUE
```

Figure 13. Retrieval processing—intermediate results

documents that match the parameters, and displays the number of documents found and a description of the request on the next screen (see Figure 13).

If the number of documents found is not satisfactory, the user can return to the retrieval screen and either modify the parameters or enter new ones, and reinitiate the search as often as necessary.

When the results are satisfactory, the user can retrieve the documents. The retrieval request and a one-line description of each document is then entered into the mail queue. The user can select and look at each document, save the required information, then file or remove the rest of the retrieval results by filing the "Retrieval Request" line.

Since keywords can be very helpful in a retrieval, provision is made to add keywords to a document and to equate keywords (synonyms) to be able to relate documents to many different projects, subjects, etc. (Keywords defined by the originator are displayed with the document; added keywords or synonyms are not.)

Because the prototype is not available at every location, "dial-up" support is provided for the CMC/ST. This support enables users visiting a remote location to "dial-up" their own offices and use the facilities of the CMC/ST to communicate with the prototype.

This CMC/ST support does not provide the full "browse" capability of a display screen device; however, sufficient function is available at the remote site that most day-to-day operations can be performed. For example, documents from the mail queue can be printed on the CMC/ST. The contents of the mail queue may be listed. Retrieval functions may be entered with the results printed on the CMC/ST and selected items printed back in hard-copy form. The characteristics of the device limit the volume of material that can be produced in this manner; however, use of the facility for specific "hot" items is practical.

Privacy Protection—Data Integrity

The philosophy of the prototype system is that only users who originate a document or are on the addressee or copy list for a document are entitled to access the document. Simply stated, a user sees only a subset of the total document data base. Each user of the system is uniquely identified to the system. The sign-on procedures perform this unique identification. This process provides a logical linkage between users and the documents they are entitled to see.

The prototype assumes that the originator or any formal addressee or person receiving a copy of a document can pass the document to others. A record of this action is kept by the system, however, so that eventual distribution of documents within the system is trackable. As an example of this process, consider the "pass-for-action" facility previously described. A manager has "sent" a particular memo to a staff member for action. The original record contains an addressee list and a list of those to get copies. A third list is appended to the document in internal storage. This list is an extension of the copy list and contains the identification of the individual who passed the copy along to this new recipient. The third list also has a provision for the informal practice of making "blind" copies for internal distribution.

The prototype also operates on the philosophy that formal documents are a matter of permanent record; therefore, no facility to delete documents is provided. If a memo is transmitted in error, the memo is recreated correctly by the originator, and then reentered in the system as a new document.

Any information entered into and acknowledged by the system will be retained by the system. Power failure or other service disruptions simply cause the system to restart without data loss. In the event of catastrophic error where data is physically destroyed, sufficient facilities exist within the system to permit recovery of the data destroyed. A system of redundancy ensures protection against data loss, without significant overcommitment or auxiliary storage. Restart of the system is not affected by power loss, nor does it require manual positioning of archive volumes, etc.

Not uncommon in normal office environments is a requirement to receive confirmation that the addressee (or recipient of a copy) has received the document. The prototype provides a facility by which the sender can request this action. A reminder is put in the sender's hold queue for each receipt request. The item is flagged in the addressee's mail queue, indicating that a return receipt has been requested. When the addressee selects and displays the document, the prototype will automatically post an acknowledgement on the sender's mail queue. The acknowledgement returned to the originator carries the date/time of delivery, and is posted to the permanent record pertaining to this document.

Schedule and Appointment Calendar

The schedule is carried in system storage in a queue similar to that used for mail. Each entry in the queue corresponds to a day, with appropriate identification in the queue entry. To examine the schedule, the principal selects a day by depressing a key. The schedule for that day is displayed, and may be paged forward or backward by key depression just as documents being viewed can be paged. The format of the day's schedule is simple, containing the hours committed, the individuals involved, and a brief statement of the purpose of the meeting or appointment. Remaining "open periods" are also indicated. Since the "statement or purpose" portion of the schedule is free text, notations to remind the principal of departure and travel time, airline bookings, reservations, etc., may be included. No limitations on the span of days that may be carried by the system are imposed. Similarly, the "work day" may be defined to the system for the individual, permitting identification of "open" slots in the schedule that match the work habits of the individual.

Entries into the schedule are made interactively, with the system soliciting the required information, and the user responding with minimal key stroke action.

The schedule may be scanned by the system and a display of only the "open" slots presented. This display may be limited to a finite period or may extend indefinitely into the future.

In any office environment, appointments are canceled, and conflicts in making them will occur, necessitating the rescheduling of affected appointments. To assist in this process, the prototype will accept new appointments for already "booked" periods. The conflict is called to the user's attention, and provision is made to reschedule the original appointments to other open time periods without reentry of the original information.

Also, to assist in setting up meetings or appointments with multiple individuals, the prototype provides a "group" scheduling capability. This process will match the open time available for each person selected, present the best time when all will be available, and indicate the persons causing a conflict. Finally, the daily schedule allows the user to enter "reminders" of nonscheduled events or actions. These are entered in essentially free-text form. Each appears as an individual item on the display of the day's schedule. Since the schedule is modified in real time, the user's display can remain current.

Conclusions

The findings of the prototype provided an excellent basis for requirements evaluation. Some very critical answers were obtained, especially in the human interface area. The population that used the system gave insight into the acceptability of the prototype functions at the level of director, manager, professional, and secretary.

Interviews were conducted during the test period and were used to determine the requirements of the personnel involved, the human factors acceptability of the system, and value.

While the prototype was being run, statistics were collected regarding its use. Use of every function by every participant was recorded. This data was accumulated in a data base for further analysis to determine how an office system is used.

The often-asked question, "Will the principal use an electronic work station which involves a keyboard?" was addressed. We found the answer to be affirmative if individual benefits are perceived by the principal.

Table 5 shows which functions were most heavily used by principals (managers and professionals) and secretaries during the prototype test. The top five functions represent better than 80 percent of the total system usage by these three occupational groups.

In the case of the principals, their most frequently used functions were all individual in nature in that they helped them organize and control their work more efficiently.

Can and will the principals use soft copy on displays and not demand hard copy of all their memos and letters? The prototype empirically demonstrated that soft copy, in the vast majority of cases, would satisfy not only the informational needs of most persons but also their personal security feelings.

Of note for future planning were requests for additional functions that were not included in the prototype. Among them was the request to allow access to data electronically. Budget, personnel, customer files, and personal computing were among several items requested. Such requests indicate that the introduction of the work station

An Office Automation Study

Managers	Professionals	Secretaries
Mail queue	Mail queue	Mail queue
Schedule calendar	Schedule calendar	Define
Retrieval	Hold queue	Hold queue
Hold queue	Retrieval	Schedule calendar
Calendar update	Calendar update	Retrieval
84†	93†	87†

*Functions are listed in descending order of use.
†Percent of total system time used by these functions.

*Table 5. Most frequently used prototype functions by occupational group**

in the principal's office will create a greater demand for interactive information and have far-reaching impacts on system architecture at the host, distributed node, and work station.

Ease of interfacing to the system for the user is critical. We used the function keyboard and a heavily prompted full processing technique which was essentially self-instructing. As easy as we thought it was, it could have been more complete. Greater consistency of command, a help function, and phased learning that requires the user to know only what he wants to use are techniques that should be used in systems of this kind.

The interviews revealed a concern about the training received. Some felt it was adequate; others thought it should have been spread over time and reinforced periodically.

Human factors, another key objective of the prototype, was held to be as important as the system functions.

We found that principals are willing to use an office system terminal and that they favor the office system concept. The idea of off-loading tasks to secretaries and/or systems is acceptable to the principals, who viewed it as a positive way to gain productivity in the office.

Prototype category	Principals: Completely satisfied (percent)	Principals: Needs change (percent)	Secretaries: Completely satisfied (percent)	Secretaries: Needs change (percent)
Responsiveness	100	0	86	14
Availability	45	55	29	71
Human factors	55	45	57	43
Community of interest	9	91	14	86
Document file size	18	82	0	100
Training	18	82	29	71

Table 6. Rating of prototype

Secretaries readily accepted the prototype. They saw its potential for easing their workload and providing them with an extended career path. They also saw the prototype as promoting closer team work between the principals and their secretaries.

Although the mail-processing function (which allowed the mail clerk to print and distribute all non-system-user mail) provided a mechanism for installing and allowing growth of the office system in the headquarters location, there was a limit to what it could handle. In the mailroom, 40 percent of the volume comes from the payroll department in the form of checks, verification of changes in deduction, tax, and withholding forms, etc. These items will be difficult to include in electronic form until there is widespread availability of image scanners, displays, and printers and total use of an electronic office by all employees.

When asked to rate their satisfaction with the prototype on an overall basis in certain categories, principals and secretaries responded as shown in Table 6.

Quantifying the benefits of an office system for principals was one of the key objectives of the prototype. The methodology used to quantify those benefits was:

1. Determine in what activities principals spend their time.

2. Provide a system for principals to use that addresses these activities.

3. Determine if the principal uses the functions of the system and how frequently they are used.

4. Calculates the time saved per principal by using the system.

If it is assumed that a principal will only use a function repeatedly over a period of time if some benefit is perceived, then the prototype has quantified the benefits to a principal.

The key to this analysis is the premise that time is of value and time savings represent potential benefits to a company. The managers, professionals, and secretaries in Account Marketing were asked to estimate the way in which they spend their time. These profiles of information-handling activities were compared with self-estimates made at other studies conducted by IBM.

The next step was to determine the impact of the prototype on these profiles. In other studies, the impact estimates were made by the study teams.

An Office Automation Study

In the Account Marketing study, the users included in the post-installation interviews were asked to make the estimates. Twenty percent of the professionals and 43 percent of the secretaries were willing to make estimates. In their opinions, a system such as the office communications system prototype enhanced to meet their requirements could save: (a) 5 to 25 percent of a principal's time and (b) 15 to 35 percent of a secretary's time.

Another way of stating the potential time savings is to place a dollar value on the time. A conservative approach would be to use salary plus fringe benefits as a dollar value of the potential benefits due to time savings. Assume the salary of the principal to be $25,000 and the secretary's salary to be $10,000. In both cases, assume a fringe benefit amount of 35 percent of salary. The total cost per employee would be $33,750 per year or $2813 per month for the principal and $13,500 per year or $1125 per month for the secretary.

Applying the potential savings estimates yields these results: For the principal, a five percent savings would equal $141 per month and a 25 percent savings would equal $703 per month. For the secretary, a 15 percent savings would equal $169 per month and a 35 percent savings would equal $394 per month.

It is management's decision, quite likely at the executive level, as to the way in which these potential benefits are to be realized. Among the choices are the following methods:

1. Expanding mode: This method operates either by increasing the labor input, but at a rate less than the growth of output, or by holding the labor input constant, but increasing the output.

2. Steady-state mode: This method consists of either holding the labor input constant, but increasing the quality of the output, or reducing the labor input and holding the output constant.

3. Contracting mode: This method operates by reducing the labor input at a rate greater than the cutback in output.

The Account Marketing principals work in an environment having many interruptions. Much of their time is spent on the phone and in meetings, both scheduled and unscheduled. They also spend considerable time away from their offices—either elsewhere in the building, at other IBM locations, or at customer offices.

Because of the nature of their jobs (creating and supporting new marketing programs), work is often done at home or while in transit to remote locations. These individuals handle information for themselves and for others.

Conversely, secretaries in Account Marketing spend most of their time at their desks. When they are absent, their work is generally covered by other secretaries. They handle information primarily for others, do not travel, and are not expected to work at home. They do, however, still operate in the same interrupt-prone environment, where they must answer phones for many principals while still attempting to complete their other tasks.

Several ideas were expressed during the interviews with Account Marketing principals that related to "deliverables" (market support aids). Currently, Account Marketing is measured according to field personnel's estimates of the impact these deliverables have on IBM marketing programs. The ideas expressed to members of the study team related to improving those measurement ratings. Among those expressed were:

1. Use the time saved to be with field personnel helping them implement/understand the deliverables.

2. Get the deliverables out sooner.

3. Improve the quality of the deliverables.

Although qualitatively these alternatives appear to be worthwhile pursuits, they are not expressed in quantitative terms, nor were those interviewed able to assign a value. The findings in Account Marketing validated what the study team had learned in previous studies; that is, the requirements of the office, the selection of a benefits alternative, and the value of the alternative have to be decided by management, possibly at an executive level.

The other alternative for management is to allow an office system to change the way it does business. A suggestion came from the interviewing process in support of this alternative; e.g., use the electronic office system not only to create the deliverables, but also to distribute them directly to IBM field personnel. Some benefits of this alternative would be to:

1. Eliminate the reproduction of the deliverables package, which would (a) reduce production costs and (b) shorten the delivery schedule.

2. Reduce the distribution time required to get a deliverable to all end users.

3. Have the deliverable stored in the system for subsequent use after the initial presentation.

An Office Automation Study

Eliminate the possibility of a deliverable being out of stock or in a deteriorated form at the end-user location. Again, the value of this benefit was not quantified and could only be done by management at the appropriate level.

Regardless of the benefit analysis approach selected, the management of a company must transform the potential benefits into reality while taking into account the state of the economy, industry trends, corporate objectives, and managerial styles.

H. Anderson, "Consider the Bank of America," *Datamation* **24,** N0. 9, 153-155 (September 1978).

Automating Business Communications, Proceedings of the Conference, American Institute of Industrial Engineers, Santa Monica, CA (1978).

J.C. Burns, "The evolution of the office information system," *Datamation* **23,** No. 4, 60-64 (April 1977).

G.C. Edwards, *An Analysis of Usage and Related Perceptions of NLS—A Computer Based Text Processing and Communications System,* Bell Canada, Montreal, Quebec (October 1977).

P.C. Gardner, "An administrative automation application," *Proceedings of SHARE 51,* 96-123 (1978).

IBM 3730 Distributed Office Communication System, Introduction, GA33-3021, IBM Corporation; available through the local IBM Branch Office (October 1978).

IBM 3730 Distributed Office Communication System, System Description, GA33-3022, IBM Corporation; available through the local IBM Branch Office (October 1978).

R.A. Lowenstein, "Office system studies: Rationale, techniques, and value," *Proceedings of GUIDE 47,* 93-106 (1978).

H. Mintzberg, "The manager's job: Folklore and fact," *Harvard Business Review* **53,** No. 4 49-61 (July-August 1975).

H. Mintzberg, *The Nature of Managerial Work,* Harper and Row, New York (1973).

R.B. White, "A prototype for the automated office," *Datamation* **23,** No. 4 83-90 (April 1977).□

High-Velocity Money—Preparing for EFT

Problem:

The idea of Electronic Funds Transfer (EFT) is a disarmingly simple concept with very complex business and social ramifications. Just about every money dealing between two parties (people, institutions, governments. . .) eventually boils down to a debit/credit transaction. Even a barter can be thought of as a dual debit/credit transaction (debit: two cows from me, credit: two cows to you; debit: 20 bushels of pears from you, credit: 20 bushels of pears to me). In the olden days, a transaction meant the transfer of hard coin from my pocket to your pocket. Paper money ushered in the era of symbolic money backed up by hard coin, and checks further stretched the connection between symbol and reality. But the reality of gold or silver is no longer a significant component of money, and the current market in these two metals is built more on the psychology of old habits than on fact. The modern commodities of money support are intangibles like national productivity, growth rate, and balance of payments. So, modern transactions are refined to the ultimate symbolic simplicity of changing numbers in the buyer's and seller's accounts, and what could offer a better medium of connection between the two accounts than a data communications system?

EFT's major technical problem is security, which we feel will be solved ultimately by engineering techniques. EFT's potential impacts on society raise problems that are not so easily solvable. The easiest problem to solve is the business dislocation caused by being able to compress a transaction from, say, two weeks to a few minutes. There is very little time left to contemplate the deal between the time you agree and the time you pay your money. Also, you will no longer have that period of grace (called "float" in the trade) between the time you send the check and the time the funds are actually withdrawn from your account. People and businesses will adjust to these problems. The real issue in a cashless/checkless society is the total exposure and consequent vulnerability of every business transaction to the scrutiny and possible control of a large-scale EFT network by an unscrupulous power. More will be said about these issues later. For now, we offer this excellent report to acquaint you with the various types and applications of EFT systems and to help you fit them into your planning context.

High-Velocity Money—Preparing for EFT

Solution:

One of the most intriguing new applications of data networks is electronic fund transfer, EFT.

Over the next ten years the nature of the payments mechanism in some countries will swing from being predominantly paper-oriented to being in part electronic, with vast quantities of financial transactions traveling over data networks. Some of the world's largest data networks will be involved in this application. Some large banks are now planning private networks with tens of thousands of terminals. There are more than 1400 banks in the United States, and eventually they will be interlinked into nationwide networks for transferring money. Many institutions other than banks handle money, hold deposits, and offer credit. The financial data networks affect all such institutions and present sudden new opportunities that will generate fierce competition in the money-handling business. Eventually, many billions of messages per year will be passing over the financial networks in the United States.

MONEY IS INFORMATION

Before societies used money, trade used to be carried out by means of barter. Later man devised systems of exchange in which certain commodities became standards of value against which all others were measured. In early societies these standards had intrinsic usable value, such as corn, cattle, or wives. Later, gold became a standard. Gold was rare, divisible, unattacked by rust and lichen, and beautiful enough to inspire poets. For millennia gold has been the world's standard of value, fought over, traded, ornamented, stolen, and worshipped. Only in recent years has man gathered the effrontery to question the necessity of this "symbol of pre-eminance ordered by the celestial will."

Paper money was invented in 1694, and was at first regarded by many as a sinister banker's trick. Until this century paper money only seemed respectable if it was backed by an equivalent amount of gold in some banker's vault. Today the currency note is no longer convertible into gold. Currency bills used to say that a central bank "promised to pay the bearer on demand" the value represented by the bill. Today they merely say "In God We Trust."

In relatively recent years, checks came into common use by the man in the street, replacing the need to pay with currency. In the late 1960's check and currency usage began to give way to credit cards, removing the payments still further from the backing by gold or commodities. Today, what may be the ultimate payments mechanism is gaining momentum. The transactions are neither gold, nor currency, nor paperwork, but instead are electronic bits flowing between computers—electronic fund transfer.

EFT recognizes that money is merely a form of information. The dollar bills that pass from pocket to pocket have become merely a confirmation of man's ability to pay. If money is merely information, then that information can reside in computer storages, and payments can consist of data transfers between one computer and another. EFT enthusiasts began to talk of a cashless, checkless society.

In reality, society will be neither cashless nor checkless for the foreseeable future. Rather, what has worried bankers is that the paperwork associated with checks and credit cards is growing by leaps and bounds. One hundred million checks a day are written in the United States, and without automated fund transfer this number would double in the next ten years. While checks are expensive because of the paperwork costs, credit card transactions are even more so—their cost is approximately 50 cents per transaction and rising. Electronic fund transfer offers a way to slow and later reverse, the growth of paperwork.

The replacement of gold by paper money and of paper money by checks were each revolutionary in their day. Now we must become used to financial transfers occurring in the form of electronic pulses on a data link. The paperwork associated with the transaction will now merely inform us about the transaction, rather than represent the transaction itself. Nor will it have to be punched into cards and fed into a receiving computer. The eventual consequences of the simple idea of automatic credit transfer will be enormous. Vast random-access computer files in banks will hold full details of all accounts. As a transaction is entered into the system, transmitted data will cause the appropriate amount to be deducted from an account in one computer and added to an account in another. Eventually, the financial community will become one vast network of electronic files with data links carrying information between them.

Thomas J. Watson, Jr., then President of IBM, foresaw the revolution in banking as follows in 1965:

"In out lifetime we may see electronic transactions virtually eliminate the need for cash. Giant computers in banks, with massive memories, will contain

From *Future Developments in Telecommunications*, 2nd Edition, by James Martin. Chpt. 14, pp 239-257. © 1977 by Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

High-Velocity Money—Preparing for EFT

individual customer accounts. To draw from or add to his balance, the customer in store, office, or filing his balance, the customer in store, office, or filling station will do two things: insert an identification into the terminal located there; punch out the transaction figures on the terminal's keyboard. Instantaneously, the amount he punches out will move out of his account and enter another.

Consider this same process repeated thousands, hundreds of thousands, millions of times each day; billions upon billions of dollars changing hands without the use of one pen, one piece of paper, one check, or one green dollar bill. Finally, consider the extension of such a network of terminals and memories—an extension across city and state lines, spanning our whole country.[1]"

Perhaps payments of the future, instead of carrying the message "In God We Trust" should say "In Mother Bell We Trust."

FOUR TYPES OF EFT

There are four main types of electronic fund transfer representing successive steps towards an EFT society. The first involves transfer of money between banks, to carry out clearing operations. Second, there are transfers between the computers of other organizations and the bank computers. A corporation may pay its salaries, for example, by giving a tape or transmitting salary information to a bank clearing center, which distributes the money to the appropriate accounts. Third, the general public uses terminals to obtain banking services. These terminals include cash-dispensing machines in the streets. There are a variety of such terminals with different functions, and bankers refer to them as CBCTs (Customer Bank Communication Terminals). They threaten to play havoc with the traditional structure of banking, at least in the United States.

To operate the CBCTs, customers are equipped with machine-readable bank cards. These cards, which look like credit cards, make possible the fourth and ultimate phase of electronic fund transfer, in which customers pay for goods and services in restaurants and stores by using their bank cards or similar cards provided by American Express, large retail chains, petroleum companies, and other organizations. Today's credit card devices (which create paperwork) are replaced by inexpensive terminals that accept the new machine-readable cards. Thousands of such machines are already in use.

EFTS (Electronic Fund Transfer Systems) thus describes a wide variety of different computer systems, but in general EFTS has become synonymous with advanced new technical directions in banking.

The present payments mechanism is highly labor intensive. Credit cards have increased, not decreased, the quantity of paperwork and manual operations. Labor costs are rising, and it is becoming more difficult to obtain workers for dull, boring, but accuracy-sensitive tasks. It has been estimated that the overall cost of using credit cards exceeds 50 cents per transaction in the United States, and that the cost of equivalent EFT transactions could be dropped to 7 cents.

Furthermore, EFT can make cash available to bankers faster, and time is money, especially with today's high interest rates. The volume of checks alone in the United States is about $20 trillion per year. If the money from these could be available to banks one day earlier on average, because of faster processing and clearing, that represents a float of $54 billion per year. It is worth installing some expensive automation schemes to capture a portion of this float.

AUTOMATED CLEARING HOUSES

A bank clearing house takes checks drawn against many banks and allocates the funds appropriately. The automated clearing house (ACH) movement in the banking industry is an attempt to create an electronic infrastructure that can reduce the labor in check clearing. This will both lower the cost of check clearing and speed it up. It will also enable banks to offer new services to their customers. For example, computers in some corporations deliver the payroll in electronic form to a clearing house, and from there the money is moved into the banks of the employees. There is then no need to print and read payroll checks. Initially these electronic payrolls were delivered on magnetic tape. It would be quicker to transmit them by telecommunications.

In the United States the number of automated clearing houses is growing and may eventually reach 35 or so. Between these centers a telecommunications network will operate transmitting many millions of transactions per day by the early 1980s. The National Automated Clearing House Association (NACHA) coordinates the development of the clearing house facilities, which must remain neutral to the competitive banking industry. The automated clearing houses and their network will have a vital role to play as EFT systems spread to the consumer level.

PREAUTHORIZATIONS

Many of the payments made by check are repetitive payments; the same sum being paid at regular intervals, or at least a sum that can be calculated well in advance. Such payments include rents, mortgages, local taxes, society dues, interest payments, social security payments, salaries, installment credit pay-

High-Velocity Money—Preparing for EFT

ments, and so on. Much work can be saved if these payments are made by preauthorization (the term standing order is used in British parlance). An instruction to make the payment repetitively is given to a bank computer, and the payment is made without further paperwork.

The U.S. government handles some military payroll and many social security payments in this way. Some labor unions have discussed having workers paid daily by electronic means. The preauthorized payments may be handled by an automated clearing house or by a suitably prepared bank. Many bank customers would welcome a bank service that pays their rent, mortgages, society dues, and so on, without involving them in further paperwork.

The situation is slightly more complicated if the payments vary each time they are made. Dividends, like wages, can be paid automatically into customer accounts, and most customers, once they are used to it, welcome rather than resist this form of computer-to-computer payment. In a similar way, telephone companies and other utilities could send their bills directly to the bank clearing system. This, however, is a much more drastic step because money is being taken from the accounts of individuals rather than added to them. Many consumers feel that they should have the option of not paying their telephone bills! Nevertheless the majority would probably welcome an automatic bill-paying service.

Datapro Comment:

The right not to do something is just as important as the right to do something. The automatic, locked-in, no-option route is anathema to all civil libertarians, but part of the price we may have to pay for the glories of the Electronic Age is an abrogation of the right to be civilly disobedient.

If preauthorized payments of these types were fully used, the total volume in the United States would exceed five billion transactions per year. The only paperwork would be periodic statements informing the bank users what transactions had been made. The paperwork would be statements about the transactions, and would not be the transactions themselves. There would be no writing or processing of checks, or rooms of keyboard operators laboriously and sometimes erroneously entering details of transactions.

AUTOMATED CREDIT

If money is deducted from consumer accounts by electronic fund transfer, some of these accounts are likely to go into the red periodically. The consumer does not have quite the same control as when he can add up every payment he makes with his checkbook (though few today do so).

An essential aspect of EFT, therefore, is the ability for customers to have negative balances in their banks. The magnitude of the permissible negative balance would be set by a bank officer (or possibly set automatically). The customer would be automatically charged interest on his negative balance. There are various forms of automatic negative balance in operation today. Some of them seem designed to create customer resistance by charging exorbitant interest rates on an automatic loan. A floating negative balance with a reasonable interest rate is required.

Automated credit offers a constant temptation to overdraw. This might have great appeal to bankers who would make money on the interest charged. Customers, on the other hand, might resent losing their float or losing their ease of refusing to pay. A wide variety of incentives have been devised for making EFT appealing to customers including discounts, cash-dispensing terminals in corporate offices, EFT terminals in corporate cafeterias, lower bank charges, and general ease of obtaining money.

INTERNATIONAL FUNDS TRANSFER

Systems for transferring funds between banks electronically are coming into operation not only on a national scale, but also internationally. The first major international network is the SWIFT system.

SWIFT, Society for Worldwide Interbank Financial Transactions, is a nonprofit-making organization set up and wholly owned by banks in Europe, Canada, and the United States. SWIFT implemented and operates the network shown in Figure 1, the purpose of which is to send money, messages, and bank statements, at high speed between banks. The participating banks financed the system and a tariff structure charges for its use on a per-message basis plus a fixed connection charge and an annual charge based upon traffic volumes. The banks range from very small to banks with 2000 branches.

As in other telecommunication systems, SWIFT imposed standards for procedures and message formats on its users. These enable the banks to send and receive messages between countries in computer-readable form. Figure 2 shows typical SWIFT messages.

The SWIFT system is a message-switching network, which originally has two switching centers as shown

High-Velocity Money—Preparing for EFT

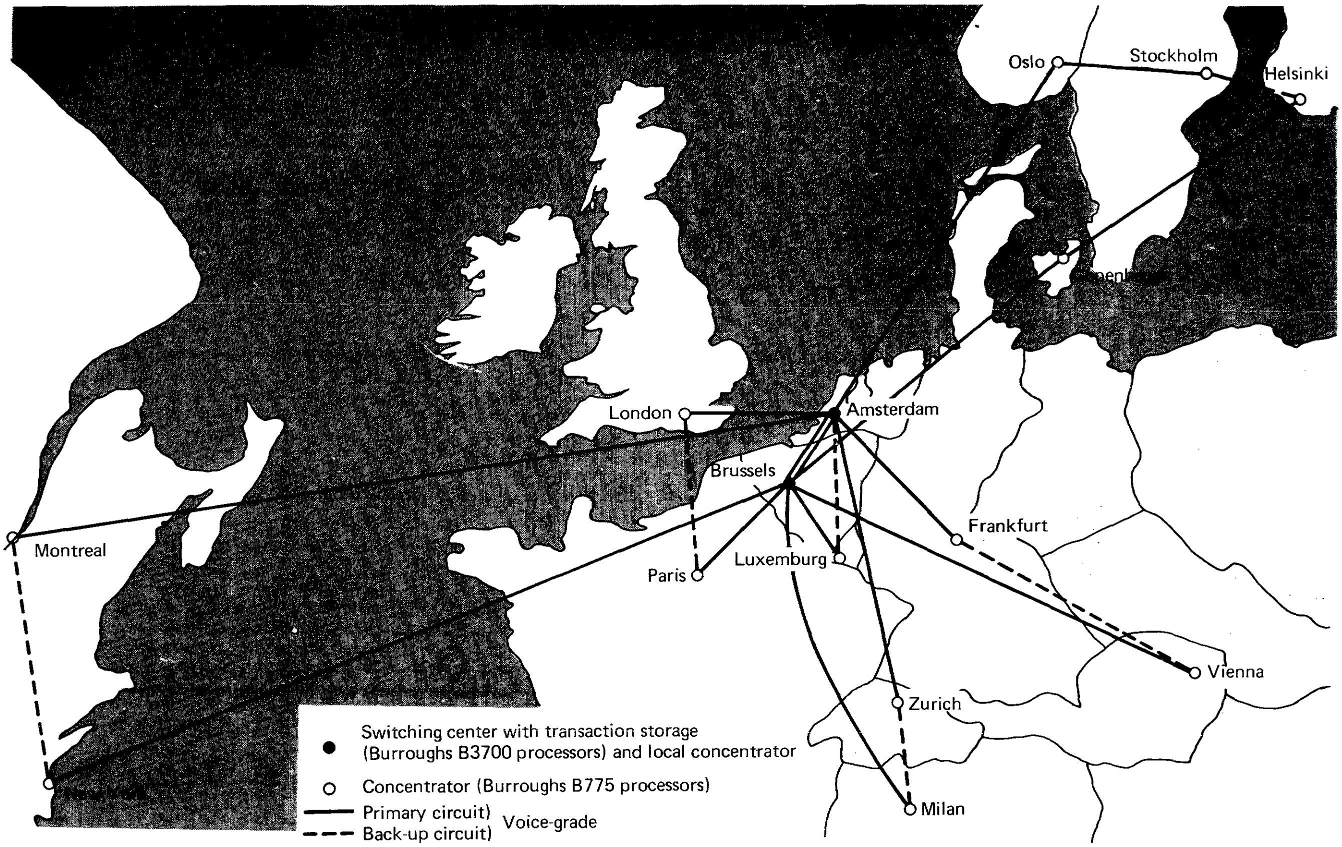


Figure 1. The SWIFT network for international fund transfer

in Figure 1. It can be expanded without functional redesign to have multiple centers. It uses voice-grade circuits, and most traffic is delivered in less than one minute. All traffic is stored at the switching centers for ten days and during that period can be retrieved if necessary. Transactions can be entered into the system regardless of whether the recipient bank's terminals are busy or not. The originator of an urgent message will automatically be informed by the system if there is a delay in delivering the message.

It is estimated that by the late 1970s the SWIFT traffic will be about one third of a million messages per day. The initial switches were each designed to handle 23 transactions per second. The system can accept either single messages or bulk traffic from computers or magnetic tape. Transactions can have priorities allocated to them.

BANK CARDS

In the mid-1970's, a new wave of banking automation swept across America triggered by the advent of machine-readable bank cards. These plastic cards have the size and appearance of a credit card, but unlike conventional credit cards, they carry invisible data that can be read by a terminal. On most such cards the data is encoded on two magnetic stripes. A third magnetic stripe is being added on which data could be written by the terminal as well as read. There are some exceptions to the magnetic stripe technology, notably New York's First National City Bank's cards, which contain a stripe read using ultraviolet light.

CUSTOMER-ACTIVATED TERMINALS

Using a bank card at an appropriate terminal, a bank customer can inquire about the status of his accounts. He can deposit or withdraw cash, borrow money if it is not in his account, or transfer money between different types of accounts. In fact, he can do virtually everything that he would previously have done by going to a branch of the bank, standing in a queue, and talking to a teller.The interesting question arises: if all of a bank's customers used bank cards and terminals, would the bank need tellers? Perhaps it would only need officers, who deal with situations needing human interaction and decisions. A bank could close some of its branches in expensive city streets and yet give its customers more convenient service because the automated teller terminals are becoming located in stores, shopping plazas, airports, factory cafeterias, and office buildings. Furthermore,

High-Velocity Money—Preparing for EFT

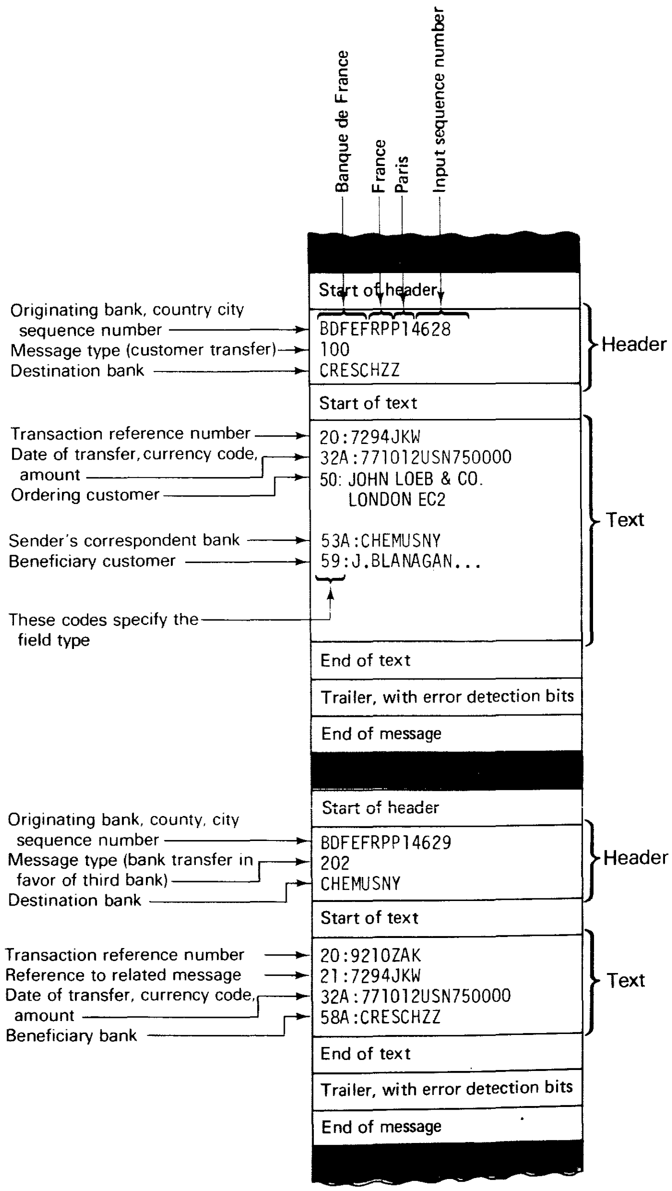


Figure 2. An example of a transaction being handled by the SWIFT network. A customer, John Loeb & Co., ask their bank in Paris, the Banque de France, to transfer $750,000 in U.S. currency to the account of the customer J. Blanagan in Swiss Credit Bank in Zurich. Because the currency is that of neither the sender nor receiver country, a third bank, the Chemical Bank in New York, is involved. Both the sender and receiver banks have accounts with this third bank which handles the reimbursement. The Banque de France first sends a message to the Swiss Credit Bank with details of the transaction. It then sends a related message to the Chemical Bank in New York asking it to debit the Banque de France's account with $750,000, and credit the Swiss bank's account.

the customers could obtain cash or other banking services when the bank was closed.

Some customers have an initial hostility to banking by machine, but once used to its convenience few want to go back to queuing in marble-pillared branches.

The prospect of doing away with the bank teller is revolutionary enough, but another implication of automated teller terminals threatens to play havoc with the entire structure of banking. Banking in the United States has traditionally been regulated by state and federal laws saying where a bank may have its branches. The McFadden Act of 1927 prohibits interstate branching and makes national banks conform to restrictive state laws. No bank can have branches in more than one state; some can operate only within a city; some can have no branches other than at the head office location. Persons living near state boundaries, for example near New York City, are sometimes flooded with advertising from banks they cannot use. In 1974 the Comptroller of the Currency, who regulates banking activity, made the ruling that a remote terminal which customers use does not constitute a bank "branch." Following this ruling, banks rapidly started to spread their tentacles into geographical areas from which they had earlier been excluded. The controversial ruling was then challenged in the courts and partially reversed, but nevertheless it seems certain that the structure of American banking will change fundamentally.

New York's First National City Bank has almost 300 branches in New York State and cannot have branches elsewhere. However, it has several thousand bank-card terminals in the New York area with some across the state boundary in New Jersey. A landmark court case ruled the bank terminals in the Hinky Dinky supermarkets in Nebraska were legal and did not constitute branches. In a Nebraska supermarket, an individual can elect to pay for $25 of groceries by switching the $25 from a savings account to his account with the supermarket. The transaction is completed without the use of currency or checks.

Automated teller terminals can be operated by organizations other than banks. Several savings and loan associations operate them. Consumer finance companies, large chain stores, gasoline companies, credit unions, and other organizations, collectively extend more consumer credit than the banks. Many of these operate credit checking terminals, and some have applied for permission to hold customer balances, in which case their terminals could have most of the functions of a banking terminal. The banks, in other words, could face electronic competition from a variety of other organizations.

High-Velocity Money—Preparing for EFT

The 1974 ruling that a banking terminal was not a branch opened up the possibility of large banks developing nationwide terminal networks. Later regulation may prohibit that because the competition could become too severe for small local banks. However, the bank customers who have become used to electronic banking certainly want to have its facilities nationwide. A New Yorker wants to be able to use his New York bank card across the river in New Jersey and across the country in California. He not only wants to be able to use the card in stores and restaurants, but he also wants to be able to obtain cash from cash-dispensing terminals.

Nationwide networks for checking a customer's credit already exist. Nationwide networks for transferring consumer funds and accepting the bank cards will be built. Hooking together many localized systems of banking terminals is a problem not unlike hooking together many localized telephone companies. There must be national standards for the terminals that are used. A nationwide network may develop that serves many different banks, as AT&T Long Lines serves many different telephone companies. A small-town bank, like a small-town telephone company, will be able to say to its customers, "We can hook you into the world."

TERMINAL SHARING

Although the simplest form of bank-card terminal can be cheap, a full-function automated teller terminal is expensive. It is desirable that banks should be able to share the services of these terminals, and the banking regulations appear to encourage such sharing. A big city store or restaurant would not want to install dozens of different terminals connected to different banks.

The International Standards Organization, ISO, has established standards for the cards that are used, and the coding that will be employed on their magnetic stripes.

When terminals are shared between banks, they relate to automated clearing house operations. Most customer-activated terminals are today installed by specific banks. In the view of some authorities (e.g., Reference 1) as electronic fund transfer matures the terminals and supporting networks will become more and more neutral, and competition will be based on the design of services not on control of the electronic delivery system.

Many small banks regard the spread of automated-teller terminals with great apprehension. They have neither the finances nor the expertise to establish such systems themselves, but fear that the terminals of larger banks will invade their territory and erode their customer bases. Organizations of small banks in the United States have been fighting the new regulations which permit the spread of customer-activated terminals. It is desirable that the regulations evolve in a manner which permits small banks to remain competitive and yet allows the desirable automation to spread as fast as possible.

In the long run the 1400 small banks in the United States cannot be prime producers of complex machine-based services, any more than the 1700 small telephone companies can be the prime producers of telephone equipment. Instead they will become service organizations which often provide a pleasanter service than the larger banks, and which enable their customers to use the nationwide EFT facilities. The small bankers will have an in-depth knowledge of their customer's needs and will be able to advise them how to use the increasingly complex banking services. They will act as franchisees of the big electronic systems and bank cards, and as with other franchise businesses, this can be both highly lucrative and competitive.

BANKING FROM HOME

The simplest customer-activated terminal is the telephone. Some banks offer services that enable customers to make payments by telephone. There have been experiments in which some banks have attempted to automate banking from the home by using a touchtone telephone and voice answerback systems.

One of the costliest operations in the use of computers is the preparation or keyboard entry of data. To reduce costs it is desirable to persuade customers to enter transactions themselves in an electronic form. This can be done at customer-activated terminals in banks or in the street. It could also very conveniently be done from the home on a 7-day, almost-24-hour basis. Paying one's bills at home with a touchtone telephone could be very convenient and appealing to many customers. Some banking authorities believe that banking from the home will become commonplace.[2]

Catalogue or mail order shopping could be done from the home telephone, using EFT. This is done today without EFT by one large store in Canada, using a voice answerback computer system. In the more distant future television, or still-frame CATV, selling may enable viewers to purchase goods and pay for them using a keyboard.

POINT-OF-SALE TERMINALS

A major amount of human drudgery will be saved when the payments made by customers in stores and restaurants are entered directly into the banking systems instead of being made by credit cards or checks. The rapidly spreading bank cards are the means for such transactions, and consumers have demonstrated in a few early systems that they like the convenience of paying with bank cards.

The term applied to extension of financial services to stores and restaurants is point-of-sale (POS). Eventually there are likely to be 50 million electronic point-of-sale transactions made per year in the United States. Much of the initial use of point-of-sale terminals is not for fund transfer, but for checking the consumer's bank balance so that checks can be cashed.

The networks needed for point-of-sale fund transfer are, as elsewhere in telecommunications, highly volume-sensitive. The initial systems may therefore be difficult to cost-justify, but once such systems exceed a certain volume they will become highly economic and will probably spread very rapidly, as did the use of credit cards.

One essential to the future of point-of-sale EFT is the availability of mass-produced inexpensive terminals. Such terminals ought to have some form of back-up, such as a tape cassette, for when the telephone link or computer system fails. Some terminals now in use are simple and low in cost. It is interesting to reflect that a bill-paying terminal for the home could be very inexpensive. Such a terminal could identify itself to the bank computer and have no need for a bank-card reader—merely a keyboard which sends signals like touchtone telephone beeps. The cost, manufactured in quantity, could be as low as that of pocket calculators.

Because of the volume-sensitivity, point-of-sale linkage to banks is coming from the very large banks, cooperative bank-card groups or from service bureau organizations. In many cases, it is not the commercial banks that are providing point-of-sale terminals. The banks face potential competition in this area from credit-card companies, credit unions, saving and loan associations, finance companies, and particularly from the large stores themselves. The larges stores see point-of-sale terminals as carrying out functions other than merely the cash transfer. They can provide better inventory control and sales analysis, tighter credit control, improved cash flow, shorter checkout time and hence less checkout staff.

A factor which will make the area highly competitive is the question of what organizations handle the sums of money, gigantic in total, that consumers deposit and which are extended as consumer credit. The battle for these funds, each running into hundreds of billions of dollars, will be fierce and is very much related to what credit cards or banks consumers use.

Figure 3 shows some examples of how individuals relate to EFT. There are many alternatives to the details shown in Figure 3.

CRIME

Eventually a high proportion of society's payments will be made with machine-readable cards. If appropriate security procedures are built into the systems, a criminal would be able to gain nothing by stealing a bank card. They could be made much safer than today's credit cards (and in some systems have been). One of the subsidiary benefits of an EFT society could be that street robberies are greatly reduced because pedestrians no longer carry much money or credit cards.

Major costs are loaded onto today's economy from robbery, theft, and fraud. It is reflected in the price of insurance, the reluctance of big-city police to investigate minor robberies, the fear of walking in city streets after dark, the inability to obtain some types of insurance. Insurance is impossible to obtain on many welfare check payments due to the risk, for example. Electronic payments systems with tight security controls could do much to lessen this burden of crime.

The EFT networks themselves offer new opportunities for ingenious computer crime, and tight security controls need to be built into the systems. The only way to make the transmissions safe from wire-tapping is to use cryptography. Some EFT terminals use cryptography having an extremely high level of safety. In the author's experience, on one network that did not use cryptography, funds were in fact stolen by an ingenious programmer. Like most such crimes no mention was made of it outside the organization in which it occurred. The technology does exist for making EFT networks sufficiently secure, though on some systems it may not be used adequately.

Successful crime in a computerized society will probably require, like other activities, longer training and a higher IQ.

NEW CONTROL REQUIREMENTS

There may be some subtle problems associated with the fact that money can move at the speed of electricity, and these problems will require new control mechanisms.

First, it may be a cause for concern that international transfers can take place at an unlimited rate. One can imagine a stampede to move funds out of a particular country, or sudden planned currency moves by power groups, adding to world financial instability.

High-Velocity Money—Preparing for EFT

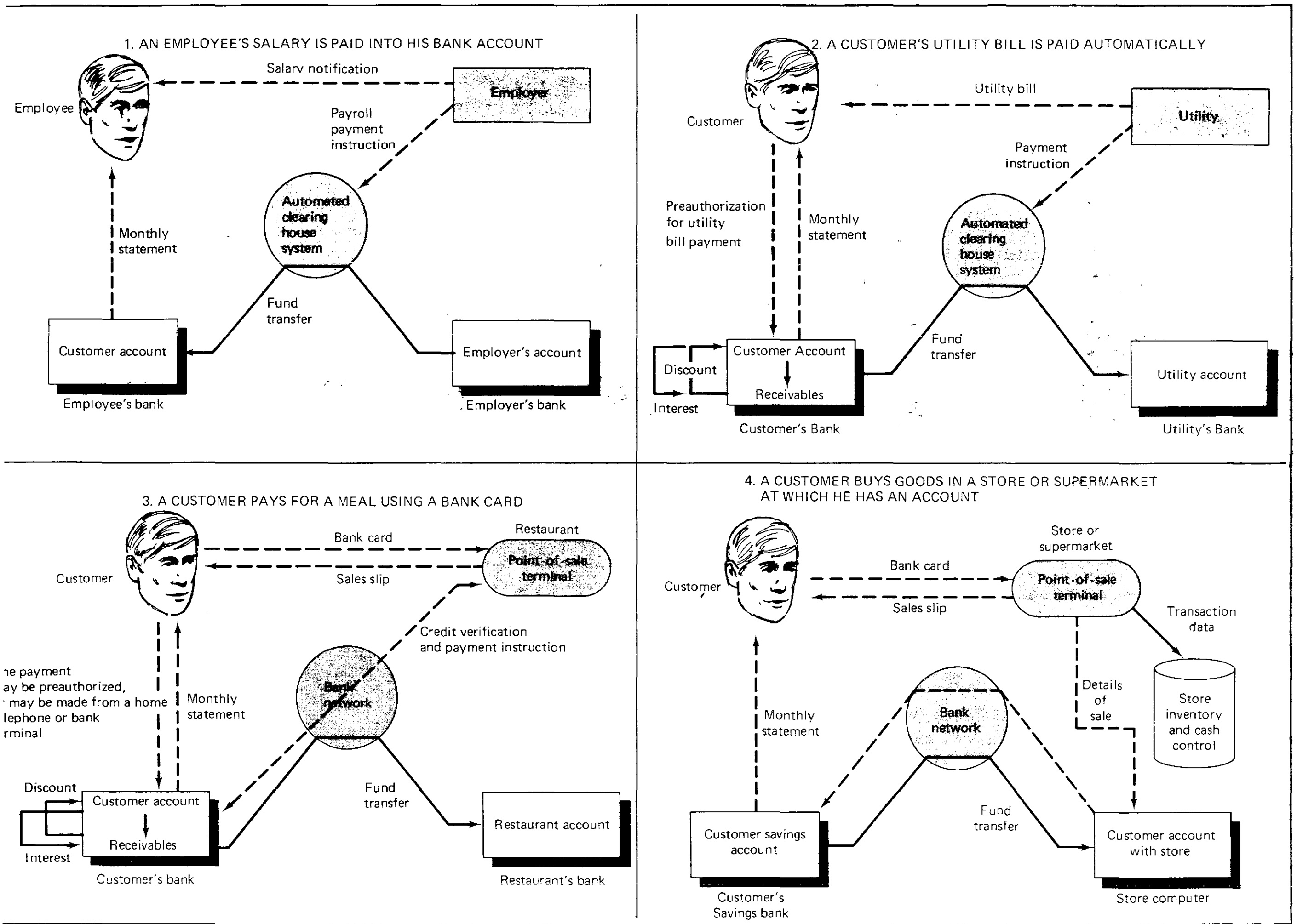


Figure 3. Examples of electronic fund transfer as it affects an individual. The solid lines represent payment instructions or fund transfers. The dashed lines represent information external to the system for the benefit of the individual. There are many variations on the operations shown here.

An extensive use of EFT will reduce the "float" money within a country, like taking up the slack in an anchor rope. It will make the float available for people or corporations to spend. Economists use the phrase "velocity of money' to refer to the rate at which money turns over. Today one dollar issued by the government might be spent 17 times in the course of the year. With EFT, the velocity would be higher, perhaps much higher; the same dollar issued would be spent many more times. Unless compensating controls were devised, this increase in the velocity of money would probably have an inflationary effect. As with a vehicle, the faster the velocity the tighter the controls must be. Today's mechanisms would provide inadequate control.

Furthermore, the nature of monetary controls will be complicated by the spread of EFT to institutions other than banks. The Federal Reserve authorities today control what is referred to as M_1—currency in circulation plus the demand deposits in banks. With the electronic systems now developing, that will become a steadily smaller proportion of the money that is actually available for spending. Savings banks, savings and loan associations, and credit unions are talking of making their deposits electronically available, and have been told that they will have direct access to the electronic clearing houses rather than going via regulated banks as they do with checks today. These insitutions are likely to resist any attempt to regulate them like commercial banks. Even some large chain stores are considering accepting customers' deposits as well as making credit available to customers.

Individuals, as EFT spreads, will have multiple lines of credit available to them via their EFT cards. They will have credit available from stores, from travel and expense organizations like American Express, from negative balance privileges in bank computers, and other sources. Using such facilities, individuals will be able to offset to some extent federal efforts to tighten the money supply.

While today's controls seem inadequate for an EFT society, the electronic mechanisms lend themselves to new and far more intricate control techniques. A comprehensive set of codes could be appended to the transactions so that computers could recognize many different categories of transaction. The transactions in different categories could be summarized and the summaries digested by federal computers. The government could be given a precise picture of how the electronic money was being used, and regulators would obtain rapid warning of any sudden movements or changes in trends which are exceptional enough to warrant attention. Legislators, or the public, could be given a detailed economic balance sheet each week, and the greater the proportion of electronic money the more accurate would the statement be. The summaries produced by the electronic systems could form input to econometric models.

Using EFT it would, in effect, be possible to have more than one type of money. For example, money for individual consumption, money for individual liquid savings, money for individual investment as in bonds and stocks, industrial capital, and government money. The government attempts to regulate the proportion of individual wealth that is used in consumption and saving or investment, but the regulating mechanisms are very crude. If the users of money were categorized it would be possible to regulate the flow from one category to another in an attempt to influence the economy. This might be done by tax mechanisms, the controlling of interest rates, possibly by more direct mechanisms such as price and wage controls. Money might be taxed, for example, when it passes from the investment category to the consumption category. If it passes from one investment to another within the investment category it would not be taxed. Other categorizations might be used to effect the balance of payments.

Such regulators could be made much more sensitive than today, and could provide a fine-tuning mechanism assuming that their effect was modelled sufficiently accurately in econometric models. It is possible to visualize the economy, perhaps by the end of the 1980s being much more precisely controlled than today, with all bank transfers being electronically summarized and fed on-line to the highly intricate econometric models which form a basis for regulation and argument concerning the economy. Compare the sensitivity of such mechanisms with that of the Ford administration doing a once-only mailing of checks to consumers to give them a tax rebate.

PRIVACY

A concern with electronic fund transfer, as in other advanced uses of teleprocessing, is that individual privacy may be eroded. An individual's financial history might be laid bare to government authorities. Some doctors rebelled against using on-line services for billing patients, presumably because they revealed too much to the IRS.

The on-line terminal services, while accepted with delight by some, are regarded with distrust by others. There is still a fear, if not always consciously stated, that the remote machine will lose information, make mistakes, or print out one's financial details for other people to see. The computers can be made trustworthy, superbly accurate, and secure. The computer files can be made private, but it will take time for these facts to be understood by the public, and it may require new privacy legislation.□

Planning Guidelines for an EFT Audit and Control System

Problem:

The major operational problem in an EFT system is fairly straight forward. The messages that flow back and forth in the network represent money, and money can be stolen from the network in the same (albeit subtler) way as it can from any other money-handling enterprise. The subtlety of the theft makes it very difficult both to block and detect unless some very rigid safeguards are built into the system from its inception. This report presents some proven guidelines for just such safeguards and explains how they can be put into effect to reduce the vulnerability of your EFT system.

Solution:

Losses from accidental and intentional acts involving computers and data communications in financial institutions are growing. The current estimate of losses from credit card fraud alone is $500 million and could rise to $6 billion or $10 billion by 1986. Of greater concern than this problem is the growing potential for single instances of massive losses as EFTS grows and as participants become highly dependent on continuously available computer services where most of their assets are stored in electronic form.

There is obviously a need to apply the principles of secure systems to the emerging development of EFTS, and much work has already been completed and published (See References 1 through 9). Most of this work has been focused primarily on the technology of security, such as: Personal Identification Numbers, Plastic Card Security, Cryptography, Physical Security, and Operating System Integrity.

On the other hand, there is also a growing awareness that EFTS security is a "people" problem. The risks, threats, and vulnerabilities to EDP systems derive primarily from the activities of individuals, either accidental or deliberate.[7] [10] In the EFTS environment, individuals may be employees of the financial institutions, merchants, telephone company, hardware and software vendors, maintenance personnel, computer service bureaus, security personnel, and auditors.

"Systems Auditability and Control in an EFTS Environment" by Russell Dewey, SRI International. From the 1978 PROCEEDINGS of the National Computer Conference. Reprinted by permission of AFIPS.

Between the two orientations—technology and occupation—a body of information is emerging regarding audit techniques and system application controls that, if implemented, could help reduce the potential for significant loss.[5] The purpose of this report is to review and summarize the state of that art within the context of the EFTS environment.

THE EFTS ENVIRONMENT

The number of financial terminals being installed in remote locations to automate all or part of the transfer of credits and debits is increasing. By 1980 there may be over 100,000 terminals providing a variety of EFTS services, including:

- Deposits
- Withdrawals
- Transfers of balances between accounts

Planning Guidelines for an EFT Audit and Control System

- Direct Debits for purchases
- Balance inquiries
- Check authorization and guarantee
- Credit card authorization and data capture
- Corporate cash management
- Funds concentration
- Corporate-to-corporate wire transfers

During the last few years, components of EFTS technology have begun to be implemented in a variety of different configurations, including shared access networks. A shared access network is one that allows for the switching of EFTS transactions to more than one possible destination, regardless of ownership considerations. Since this environment is more complex than single institutions dedicated to EFTS, we will focus on shared access networks for this report.

The likely evolution of such shared access networks may be described by their principal components:

1. Remote Terminal—The terminal may be operated entirely by one person, such as an Automated Teller Machine (ATM), check guarantee terminal, or cash management terminal. They may also be operated by an intermediary, such as a financial institution teller or merchant sales person.

2. Communications—The terminal may be connected to a computer located at a merchant, corporation, or (potentially) a government facility. In this case, the originating computer must be connected to the destination computer through intermediate computers and telephone communications facilities. Depending on the complexity of the local environment, the number of institutions participating, and economic considerations, the terminal may optionally be connected directly to a local financial institution, directly to a joint venture shared EFTS switch, or through common carrier facilities directly to the destination.

3. Files—EFTS files take several forms and may be found in several discrete locations:

- At the remote operator (merchant, corporation, government agency) there may be audit trails of EFTS transactions passing through or decision parameters to control the processing of transactions when the remainder of the network is down, but there are probably no balance, account data, or financial institution programs.
- At the acquiring financial institution there may also be audit trails, but there are also balance and account data programs for its merchants and corporate customers and for transactions that do not need to be switched, such as an "on-us" debit. The acquiring financial institution may also have decision parameters to control the processing of "not-on-us" transactions when the remainder of the network is down.
- At the EFTS switch there may be audit trails and decision parameters for any destination facility that may be down. There are probably no financial files other than reconciliation totals and settlement amounts between institutions.
- At the destination financial institution there may be audit trails, memo-post master file balances, transaction files with backup, and off-line master files with back-up.

4. Communication Equipment—Each computer site will have specialized hardware to interface the EDP system to the external communication lines.

5. EFTS Software (Program Logic)—Each computer expected to participate in such an EFTS network must have specialized programs developed. These include:

- Terminal protocols
- Message format conversions
- Switching and routing logic
- Interface logic/protocol to other computers
- Interface to existing financial software, such as—
  - —Demand deposit accounting
  - —Savings account accounting
  - —Customer information data bases
  - —New account processing
- Specialized audit techniques and application controls

EFTS APPLICATION AUDIT TOOLS AND TECHNIQUES

The transition from traditional financial institution record-keeping to today's EFTS application systems—characterized by large on-line master files, lack of manual intervention, and remote terminal entry—has brought with it design concepts that cannot rely

Planning Guidelines for an EFT Audit and Control System

on traditional manual control procedures. The accuracy and reliability of EFTS application processing are becoming more dependent on the incorporation of automated application controls (discussed in the next section). The purpose of the EFTS audit tools and techniques is to evaluate those controls, to verify processing accuracy, and to assure continued compliance with processing procedures.

Some of the key audit tools and techniques that apply to the EFTS environment and that verify the correctness of processing logic and controls include the following:

- Base Case System Evaluation—Execute application programs against test transaction data, entered through a "test mode" EFTS terminal, and compare results against pre-determined test results.
- Parallel operations—Execute new or revised application programs and existing application programs and compare results.
- Integrated Test Facility (ITF)—Enter test transaction data through live terminals commingled with live transaction data.
- Parallel Simulation—Live transactions are copied and processed against auditor programs that simulate processing logic. The simulation results are then compared with the live results.
- Transaction Selection—Systematically screening and selecting transaction samples entered through EFTS terminals for subsequent manual verification.
- Embedded Audit Data Collection (Sometimes known as System Control Audit Review File: SCARF)—Audit subroutines are embedded in the application program to screen and select internal generated messages between EFTS logic modules that result from the original terminal transaction.
- Terminal Audit Software—Direct access to inspect live files during actual operation of the EFTS. The auditor will want to examine terminal polling lists, routing tables, merchant codes, floor limits, and default authorization tables.
- Snapshot and Trace—Secure documentary evidence of logic paths, control conditions, and processing sequence of a specific transaction. This is done by continuously recording transaction status for some selected transaction as it passes through the system.
- Job Accounting Data Analysis—Select, extract, and display job accounting data to monitor access to sensitive data files and on-line libraries.
- Code Comparison—Use of an off-line program to compare two versions of an application program to identify differences in coding.

EFTS APPLICATION SYSTEM CONTROLS

The purpose of the EFTS Application System Control are to assure the accuracy and completeness of the processing results, the security of the environment in which the EFTS transaction is effected, and the effectiveness of the overall computer design and operations.

The controls described in the following sections are some of those likely to be beneficial in a remote terminal EFTS environment. They are grouped according to five phases:

- Transaction Origination and Data Entry
- Data Communications
- Computer Processing
- Data Storage and Retrieval
- Output Processing

Transaction Origination and Data Entry Controls

In developing controls within this phase, an organization may wish to implement:

- Special-Purpose Forms—At merchant-operated or teller-operated terminal locations. Proper form design encourages completeness and accuracy of data. In the cardholder-operated terminal environment, Video Display Units (VDU) may be used for pre-formatted and/or interactive input.
- Transaction Identification Cross-Reference and Source Document Numbers—This control calls for sequentially assigned source document numbers (or sequential screen numbers) that are transmitted to the EFT processor as part of the transaction identification.
- Sequence Log—An internal log of which sequence numbers have been assigned to which merchant/teller locations or cardholder-operated terminal location.
- Signatures—In the EFTS environment, this control consists of appending a series of endorsements to each transaction as an audit trail of each node that handles it. It also serves as a "return destination" for undeliverable messages. The first endorsement is the terminal I.D.

Planning Guidelines for an EFT Audit and Control System

- User Identification—A unique identification by class of user (cardholder, merchant, teller, bank supervisor) used to restrict the transaction to be processed as well as files that may be accessed.

- Batch Serial Numbers—Batches are identified by serial number to provide accountability of data and to assist in the isolation of errors when an EFTS terminal cannot successfully reconcile at end of day (or other cut-off time).

- Limit the Number of Transactions in a Batch—This is done to simplify the reconciliation process.

- Turn-Around Documents—The source documents contain pre-recorded data in machine readable format. This will simplify processing in the event the on-line system is down and paper documents are used as back-up.

- Retention Dates on All Transaction Logs—This is based on legal requirements and management policy.

- Source Documents Maintained at Origin—In the EFTS interchange environment this means two things: (1) The paper documents (credit card vouchers, debit card vouchers, deposit slips, withdrawal slips) are maintained as closely as possible to the acquiring bank, to prevent unnecessary risk of loss through transportation, and (2) the electronic version of the EFTS transaction is also maintained at the acquiring facility until balanced at the end of the day. Subsequent settlement between banks and actual movement of value data for posting can occur in batch mode at lower cost and higher reliability.

- Error Logging—Maintained by the acquiring facility to record and monitor the types of errors occurring at the terminal. Suspicious patterns of such errors could imply an attempt by an outsider to breech system security.

- Verification of Re-Entered Data—The data fields on resubmitted transactions are subjected to the same verification procedures as the original transaction.

- Security of EFTS Terminals—EFTS data entry terminals should be physically secure by placement in a lockable room, putting a keylock on the terminal itself, or by placing a lockable cover over the terminal device when not in use.

- Terminal Logs—Journals in the terminals to record all transactions.

- Terminal Control Logs—Journals in the terminal controllers to identify imposter terminals on the network (i.e. controller total of transactions should equal the sum of the legal terminals.)

- Built-In Terminal and Terminal Controller I.D.'s—These devices are provided with electronic identification that can be queried by the computer; used to validate proper terminal authorization.

- Editing and Validating Routines—These may be partially performed in the terminal itself, given anticipated hardware capabilities.

- Passwords—Used to verify that the input is being received from an authorized source. Likely to be the Personal Identification Number (PIN), although other techniques are being actively explored.

- Unauthorized Access Attempts—An immediate report is produced of unauthorized attempts to access the system. After a threshold of repeated attempts the system shuts down the terminal in question and allows access from that terminal only after intervention by security personnel.

Data Communication Controls

In developing controls within this phase, an organization may wish to implement:

- Secure Phone Equipment Rooms—Locks and alarms are used to control access.

- Network Configuration Polling Table—No open addresses for unauthorized terminals to gain access.

- Communication System Control Log—To detect any unauthorized changes to the network done through network supervisor terminals.

- Communication Line Routing—Data communication lines are not put through public switchboards (PBX).

- Local Loop Security—Between the terminal, or terminal controller, or computer, and the telephone company branch office.

- Encryption Techniques—Such as the National Bureau of Standards (NBS) Algorithm. For an excellent discussion of encryption techniques, see Kaufman and Auerbach.[1]

- Forward Error Correction (FEC) and Automatic Request for Retransmission (ARQ)—Common types of link control; to control transmission errors

Planning Guidelines for an EFT Audit and Control System

on each segment of a transaction path through the EFTS network to the cardholder data base and back to the terminal.

- Message Sequence Number—Each transaction at an EFTS terminal may generate several messages (terminal-to-acquiring CPU, acquiring CPU-to-switch, switch-to-cardholder CPU, etc.). This provides a traceable log to match inquiry with response, detect lost messages, or detect imposter messages.

Computer Processing Controls

In developing controls within this phase, an organization may wish to implement:

- Monitoring of Internal-Generated Messages—Internal generated messages (such as an authorization request to the issuer on an interchange withdrawal transaction) should be uniquely identified and cross referenced to the external transaction that spawned them.
- Control Totals—Reasonableness checks and internal control totals between program modules.
- Default Option—Each level in the system hierarchy may need to make decisions in default when the rest of the system is down, or when it is uneconomical to do so (e.g., the merchant bank may authorize a $3.00 transaction against a negative file instead of requesting authorization from the issuer).
- Dual Fields—All entries should be carried as a credit against one account and a corresponding debit against another, throughout the life of the transaction.
- Arithmetic Accuracy—Techniques such as double arithmetic and arithmetic overflow checks are placed in critical points in the application.
- Destructive Update—Debit and credit types entries are used to correct error conditions, not delete or erase commands.

Data Storage and Retrieval Controls

In developing controls within this phase, an organization may wish to implement:

- File Classification—Each file is classified by security level, and access is restricted by level.
- Data Base Control Table—No data base access is allowed unless it comes from an authorized program module.
- Program Linkage Control Table—These tables control the authorized module linkage between programs.
- Dormant Files/Accounts—The system reports on dormant files/accounts that suddenly have activity on them. Many attempts at EFTS fraud try to make use of dormant accounts.
- Excessive Activity—The system reports on records and data fields that have excess activity over a certain threshold. It could mean that a lost or stolen card is being used.

Output Processing Controls

In developing controls within this phase, an organization may wish to implement:

- Reconciliation—The response to an EFTS request is matched up and reconciled with the original request before completing the transaction.
- Transaction Log—The central transaction log is periodically matched against the journal tape in the EFTS terminal. These totals are also verified against individual application control totals.
- Output Activity Review—Real-time statistics such as the number of terminals on-line, transaction quantities, and circuit traffic are reviewed by the appropriate management. System generated subjective judgments, such as default authorization, are reviewed by user departments.
- Device Verification—EFTS devices whose action such as imprinting a card or dispensing cash is controlled by the application programs will send a status report back to the host computer to verify completion of the requested mechanical operation.

CONCLUSION

Any remote terminal-based interactive system faces a variety of threats. EFTS, in particular, because of the electronic movement of funds and value, is particularly vulnerable. Fortunately, a number of system audit techniques and application controls are available to the system designer and more are under development. Depending on the particular system design, careful implementation of these audit techniques and application controls is likely to diminish the vulnerability of EFT systems.

Planning Guidelines for an EFT Audit and Control System

REFERENCES

[1]Kaufman, D. and K. Auerbach, "A Secure National System for Electronic Funds Transfer," *Proceedings 1976 NCC,* pp. 129-138.
[2]Mazzetti, Joseph P., "Design Considerations for Electronic Funds Transfer Switch System Development," *Proceedings 1976 NCC,* pp. 139-146.
[3]Backman, Frank, "Are Computers Ready for the Checkless Society?" *Proceedings 1976 NCC,* pp. 147-156.
[4]"The Analysis of Certain Threats to EFT System Sanctity," Kranzley & Co., Cherry Hill, N.J.—A study done for the Electronic Industries Foundation; January 1977.
[5]"Systems Auditability and Control Study," SRI International, January 1977. (Available from the Institute of Internal Auditors, Orlando, Florida).
[6]"Introduction to EFT Security," prepared by the Division of Management Systems and Economic Analysis, FDIC, Washington, D.C., August 1976.
[7]Parker, Donn, *Crime By Computer,* Charles Scribner's Sons, New York, 1976.
[8]Martin, James, *Security, Accuracy and Privacy in Computer Systems,* Prentice-Hall, Englewood Cliffs, New Jersey, 1973.
[9]Hoffman, Lance J. *Security and Privacy in Computer Systems,* Melville Publishing Company, Los Angeles, 1973.
[10]"A Guide to EDP and EFTS Security Based on Occupations," A report prepared for the FDIC by SRI International, Menlo Park, California, 1977.□

The Economics and Other Issues of EFT Systems

Problem:

Among technologists, forecasters, investors, bankers, retailers, and many others there is wide agreement that electronic funds transfer systems are coming. Many important commitments to the new systems are being made. But, however sure a thing this "brave new world" of technology is in the long term, it is a sure investment for no one in the near term. It seems likely that the "I can't afford not to" fad will claim a variety of victims in the corporate community, and it is possible also that some of the worst fears of users will come true. Authoritative studies indicate great uncertainty about the costs of EFTS, important consumer questions must be resolved, and significant legal and regulatory issues have yet to be handled. The aim of this report is to help business decision makers see the main problems, frame the right questions, and grasp the most relevant facts and findings about EFT systems.

If we seem to be taking potshots at EFT throughout this segment, we are—but only because we feel that the implications of EFT can have a far greater impact on Everyman's Quality of Life than any other upcoming application of easy communications facilities. Widespread EFT is inevitable because it is convenient and potentially cheaper than present cash-transaction methods. Some erosion of personal freedom is equally inevitable—all we can hope for is that it won't be very much.

Solution:

Since 1975 the financial community has been buzzing over the prospect of an electronic system that may substitute for much of the paper-based money transfer system in this country. Already widely discussed in European nations, the electronic funds transfer system (EFTS) holds out vast promise as a means of ensuring the adequacy of payments in America's high-voltage commercial world. EFTS offers a solution to the mounting problems of a money transfer system increasingly choked with checks.

But EFTS also contains pitfalls concealed in the underbrush of capital investment and of operating expense projections. The new technology already has claimed its share of overly eager entrepreneurs.

Nevertheless, EFTS is coming, in one form or another, to one degree or another, and it behooves all prudent business people to educate themselves about this new technology and to watch developments with care. The

Reprinted by permission of Harvard Business Review, "Electronic Funds Transfer: Pitfalls and Payoffs" by John B. Benton, July-August 1977.

fortunes of business—literally—may ride on EFTS in the near future. For instance:

- Companies will transmit payroll on magnetic tape to depository institutions for automatic crediting to employees' accounts.
- Retailers will accept customer cards for virtually all merchandise, debiting buyers' accounts electronically.
- Insurance companies and holders of mortgages or loans will receive automatic installment payments through their customers' banks.
- Utility companies will encourage customers to give telephone authorization to their banks to transfer payment funds automatically.

EFTS already has become significant to bankers, managers of thrift institutions, food merchandisers, and federal and state regulators of depository institutions. Lawyers, judges, and legislators, as well, are concerned with developments in this field. Also, the new technology is a subject of intense interest in the electronics industry, which annually will install a half billion dollars' worth of terminals and communications hardware in retail establishments.[1]

Congress, too, has shown interest in EFTS. Hearings and debates on the subject have been held in both the House and the Senate in recent years, and significant changes in federal law would appear to be near certainties.

To help deal with the social ramifications of EFTS, Congress established the National Commission on Electronic Fund Transfers and directed it to study developments and to explore the steps required to ensure that consumers are protected from unnecessary inconveniences, invasions of privacy, and reduced competition.

Unfortunately, business people in pursuit of profit will receive no protection from legislators. Companies will be as vulnerable as they were in the past when new technologies burst on the scene.

These situations lead to my basic warning: Business decisions about EFTS are going to depend, in a larger measure than policy makers may wish, on sound instincts. Business decisions are going to depend, to a smaller degree than might be wished, on an ability to collect information, to evaluate costs and benefits, and to choose on the basis of economic rationality.

The heart of the matter is: How much EFTS and when? This is a question top executives soon may have to answer for their businesses. To do so, they might examine EFTS from four standpoints: economics, technology, consumer wants, and the legal and regulatory climate. Let us examine each of these topics.

ECONOMIC TECHNOLOGICAL HURDLES

One reason for the initial growth of EFTS was the notion of cost displacement. In the late 1960s, the American Bankers Association, the Federal Reserve, and independent organizations sponsored studies that claimed that the existing paper-based payment system was becoming too expensive to be justifiable. The authors argued for a conversion to electronics, thereby substituting an inflation-hedged, machine-intensive delivery system for one that was labor-intensive and relatively unprotected from inflation.

These studies were the initial theoretical basis for EFTS. Two programs soon emerged—automated clearing houses and retail point-of-sale systems.

Throughout the past 10 years, it has become clear that nobody is sure that these cost saving arguments can be supported. In fact, I am not aware that any empirical research on the economics of EFTS has been completed and distributed in the public domain. A study prepared by the staff of the National Commission found: "The published estimates of the costs of implementing and operating (electronic funds transfer) systems are universally based on feasibility studies that have not been tested in actual practice."[2]

Nonetheless, during the 1960s, the growth of EFTS remained tied to the cost effectiveness argument. That growth was modest and disturbed no one. During the 1970s, however, new considerations began to alter the picture. Many marketing people in banking argued that terminal-based services brought increased customer convenience, a wider variety of services, and eventually more deposits. Attention turned away from the automated clearing houses—although steady progress continued in spite of little publicity—and shifted toward the potential for retail point-of-sale systems.

Still in experimental stage: But the shift in focus to retailing did not lift EFTS out of the experimental

---

[1]This number is based on estimated growth figures for general merchandise point-of-sale terminals, automated teller machines, and bank point-of-sale terminals over the next five years and is drawn from 1977 data published by Frost & Sullivan (New York), Creative Strategies [California], and Payment Systems, Inc. [Georgia].

[2]National Commission on Electronic Fund Transfers, A Conceptual and Methodological Framework for the Economic Analysis of Electronic Funds, Transfer Systems, Internal Working Document 27 [Springfield, Va.: National Technical Information Service, U.S. Dept. of Commerce, November 1976], p. 2.

The Economics and Other Issues of EFT Systems

stage. Many problems remain to be solved and, as we shall see later, experience to date has raised real questions about the practicality of EFTS.

At the same time business was experimenting with point-of-sale systems, it was also experimenting with automated teller machines. I call this experimentation because the machines certainly could not be justified during the 1971-to-1975 period on the basis of economic return. Consider the evidence:

- Conservatively speaking, there was an average installed base for the four years through 1975 of 3,000 automated teller machines in about 1,000 banks—with less than 10% of them on-line. Easily $200 million worth of terminal and communications equipment, research and development, systems maintenance, telephone tariffs, and operating funds were poured into these programs.

- According to industry statistics, an average total of 1,500 machines were in operation. Each of these machines processed about 1,000 transactions per month. Overall, the cost per transaction for the entire installed base must have averaged at least $2.50.

- But because only a small amount of the check processing cost was displaced by the systems, the economic justification had to come almost exclusively from the revenue earned from increased deposits or new fees. Few managers of financial institutions would argue even today that market shares have moved enough to provide increased earnings sufficient to offset the high levels of costs.

- A 1976 updated attempt to look at the economics of the same types of off-premise programs has been attempted by the U.S. Savings and Loan League. The league estimated the cost per transaction of all installed automated teller machines and point-of-sale programs at more than $1.50. (The drop from the earlier cost figure can be attributed almost entirely to increased operating efficiency and a large change in transaction volume per machine—from an average of 1,100 transactions per month before 1975 to almost 1,800 per month by the end of 1976.) In contrast, the highest estimates of the cost for conventional check processing are less than 50% of this number, and some giant banks claim publicly that their explicitly identifiable share of check-processing costs amounts to less than a nickel per transaction.

The lessons from such statistics are obvious: EFTS is still experimental. Only large volumes of business will ever make it economically feasible.

What explains the inability to get out of the experimental stage? One reason is that many of the pre-1976 experiments in the depository institutions failed to meet even the research and development goals. Much of what was learned soon became irrelevant. The most uncertain aspect of EFTS is the complexity of the communications systems development work involving many different types of computers, networks controllers, data records, systems interchange, and communications protocols. Most of the early systems eschewed such complexity and installed off-line machinery. New lessons are being learned in today's on-line, off-premise banking environment.

In 1977, the trends for both automated clearing house and point-of-sale programs are becoming clearer, the experiments more fruitful. The automated clearing house programs, in particular, have grown steadily behind the scenes, with cost displacement still acting as the driving motivator. For example, 32 automated clearing house associations have established 26 on-going projects, with 6 more planned by 1978. As of December 1976, the federal government was pumping almost 6 million transactions per month through these projects, and commercial payments were comprising another half million. The government expects to reach 20 million transactions per month by 1981. Yet adding some comparative expansion estimates for commercial payments and even bumping the figures a nod, this is still less than 1% of the 40 billion checks that are expected to be drawn yearly by 1980.

Point-of-sale systems have also grown, although not as quickly as some proponents claim. Through the end of 1976, almost 2,900 depository institutions had installed about 10,000 off-premise terminals as part of approximately 200 operational projects. (Most of these are the less expensive point-of-sale terminals that today sell for about $1,500. Less than 500 of the installed base of off-premise devices are automated teller machines.) Some 150 more are on the drawing boards. But this is a modest infrastructure compared with the approximately 225,000 on-line terminals now installed in general merchandise stores throughout the United States. (About 155,000 of these are full-scale general merchandise point-of-sale terminals, and about 75,000 are credit authorization terminals.)

Although approximately 20 million transactions per year are being pumped through the systems installed by depository institutions, most of these are not EFTS. They are much simpler authorization/verification-type transactions. However, the infrastructure is essentially the same as that which will be required for EFTS.

High expenses up front: Let us reflect on some of the lessons that have been learned about data communications technology, and EFTS in particular.

The Economics and Other Issues of EFT Systems

Most important, the technology for data communications, whether it supports EFTS, credit authorization, or order entry, is expensive, much more so than was at first thought. Of the major zones of expense found in a business operation—research and development, capital investment, and annual operating—capital investment signals the scale of the operation and the downside risk. Comparatively speaking, EFTS requires more, rather than less, up-front investment than the average enterprise. For example:

- While a modest conventional credit authorization system, dedicated by a single department-store chain exclusively to the credit authorization function, may cost more than $1 million to build, the average point-of-sale system, with greatly expanded functional capability, may cost five to ten times as much. The cost of full funds transfer lies somewhere in between the cost of limited credit systems and general merchandise point-of-sale systems. The total operating cost for these kinds of systems averages from 25% to 50% of the investment value per annum.

- One typical check verification system today, at a bank with 350 terminals in a major metropolitan city, would cost $1.5 million to build and $350,000 per year to operate (including hardware maintenance, special software maintenance, telephone lines, systems support, and operation), or about 25% of the original investment per year. This particular system is proprietary, requires no message switching to any other bank's records, and is not even on-line to the main central processing computer.

But even in the seemingly simple systems, slight changes in the marketing strategy can cause expenses to spring out of nowhere. Suppose the bank desires to guarantee all checks passed through the system. The first thing that happens is that the data-processing cost center of the bank charges a few hundred thousand dollars against the program for what was believed to be modest batch-processing support.

Next, write-offs against bad check losses accumulate at a rate of from 0.05% to 0.07% of the gross dollar volume of checks cashed. That was planned. What was not planned was the cost of the check control program, of individual follow-ups, and of collection costs. In the examples used here, check write-offs, accumulating at 0.07%, exceeded $250,000 the first year. But it cost another $300,000 of personnel services expense to keep the number that low.

A funds transfer system, requiring more sophisticated terminals and interbank switches, would command at least another $1 million of hardware expense, create much greater management problems, and cause vastly expanded systems support expense.

What volume of activity would be required to create a 15% return on investment for a bank data communications system? To answer this, let us take the case of the bank just mentioned, assuming sensibly that:

1. The benefits derived from fee income are equal to $0.15 per transaction.

2. The displacement cost is uncertain, but it will eventually be positive (thus providing a hedge against uncertainty and making the modeled numbers even more reliable).

3. Increased income attributed solely to increased market share is not significant.

4. The model estimates cost and revenue over a five-year period.

Given such assumptions, the check verification system would require more than 600,000 transactions per month to be economically viable! Even if you doubled those transactions per month you would be doing well to break even if the checks were being guaranteed as well as being verified against a central file.

Two more points are worth noting. The first is that the up-front costs are so large that most depository institutions would have difficulty affording by themselves the expense of undertaking an EFTS. The second is that business's expectations about the time span required for next year's revenue to exceed next year's cost and for total revenue over the program's life to exceed total cost may have to be very different from the expectations that guide most business ventures. EFTS may require five years or more before it can show positive returns on even the last year's expenses.

One final point can be made. We have noted that these arguments about the positive effect of scale do not always hold firm. All cost elements are not amenable to scalar economics. If you are guaranteeing checks that average $45 per purchase in a supermarket, your bad check write-offs are probably averaging more than $0.02 for every check that you cash. And your check control costs for collection, phone calls, and other expenses are very significant and are sensitive to changes in volume.

Thus one of the key questions facing the financial manager is whether his board of directors and stockholders will tolerate capital intensive programs that take five to ten years to show net earnings on total systems cost.

The Economics and Other Issues of EFT Systems

Uncertainties in the cost picture: In my opinion, financial enterprises eventually will build EFTS. They sense that more than 50% of their new deposit dollars are being created by individuals. They sense, too, that the expanded services provided by data communications technology will appeal to consumers, and I think they are right. They will strive, therefore, to be first in this market.

But if you are one of those involved in this effort, beware of hidden costs. EFTS is expensive, and the people required to back up the terminals and support your program may cost you far more than you expect. There are other costs that have a way of remaining hidden until it is too late. For example:

- The hardware or electronic equipment initially may seem expensive to obtain, but after five years the maintenance expense may be even greater.
- The private lines leased from the telephone company are expensive and rising in cost. About 30% of your dollars may go to Ma Bell or one of her competitors.
- The staff specialists, planners, and others who play with computers will always find new ways to use them, and expanded use means greatly expanded expense.
- If you are top management, expect to be deeply involved. EFTS is going to be expensive in terms of your time as well as that of middle management.

In the case of automated clearing houses, the Federal Reserve has all but concluded that the facilities will eventually reduce back office clearing and settlement expenses. This conclusion has not yet been tested with empirical data, but I think it is accurate.

In the case of point-of-sale systems, the facts on the cost side of the equation appear to be these: depository institutions can even now establish point-of-sale check verification systems capable of producing transactions at costs ranging from $0.05 to $0.30, assuming sound management, five-year amortization, and a system size having hundreds (but under a thousand) of terminals that are each transmitting a couple of thousand transactions per month. Under the same assumptions, there are funds-transfer-type automated teller machines and point-of-sale systems producing transactions at costs ranging from $0.20 to $0.85. An average cost per transaction of about $0.45 for these fuller-service systems is not at all unrealistic.

I am convinced, based on my review of reasonably hard data, that future systems, made up of 5,000 to 10,000 more terminals and most transmitting about 2,500 to 5,000 transactions per month, will achieve costs per transaction of comfortably under a nickel.

However, there is also no question in my mind that during the next five years only a few systems will achieve these numbers. For some time to come, we can expect many more systems to cost a dollar or more per transaction. Obviously the question before the providers of EFTS is: How does an institution generate sufficient off-setting revenues from fees, reduced check or credit losses, reduced operating expenses, and revenues from new deposits that were attracted to the institution by EFTS-type systems?

A major caveat from all this is that the "walk first, run later" approach makes a great deal of sense. Entry through check verification is the most economical way to begin if one can avoid guaranteeing the checks. In this way, the customers of the financial institutions can become accustomed gradually to the systems that will support funds transfer, and the cost stream can be more properly aligned with the revenue derived from the benefits that are provided.

Will government follow or lead? In the preceding discussion I did not include estimates of the cost and operations of automated clearing houses. This is because the government seems to be playing by a different set of rules from those used by the private sector.

The government's analysis of EFTS benefits is not entirely guided by the same kinds of hard-cost numbers used in the private sector because of the nature of the public sector. Thus the government's support is not likely to vacillate as a function of a bad experience here and there. For example, both the Treasury and the Federal Reserve favor the expanded use of EFTS for government payments and transfers, and they can justify supporting automated clearing house developments on the basis of the benefits for the government alone.

Furthermore, the government, based on its evaluation of the public interest, supports the expansion of EFTS to point-of-sale facilities and may enter that arena as a provider of last resort if the private sector is unable or unwilling to build and operate the communications infrastructure required.

The National Commission has carefully considered what should be the government's proper role in all of this, particularly that of the Federal Reserve. It has concluded, at least in the context of today's world, that the Fed should keep its distance from point-of-sale programs. The National Commission strongly supports the use of the private sector as the primary deliverer of EFTS.

The Economics and Other Issues of EFT Systems

Suppose the private sector is not able to deliver? This could happen if stockholders and other providers of capital insist upon favorable earnings in the short run. Then government would become the prime mover—all signs suggest this. The private sector would default a large part of a new, and potentially viable, billion-dollar-per-year industry to the public sector.

Can industry bring the costs down? Perhaps the most important lesson to be learned from economic analysis of early EFTS experiments is that we are still focusing on the wrong systems configuration. Today we think in terms of retail systems, banking systems, supermarket systems, and so forth. Some of these systems are designed to support authorization/verification-type functions. Others control inventory. Still others do remote data processing. In the meantime, AT&T and numerous other businesses are creating competing and complementary terrestrial, microwave, or satellite-based data communications nets for prospective users. So what we have are many different systems performing seemingly different functions through different types of equipment when, in fact, the differences are more apparent than real.

In my judgment, EFTS will not become cost effective until all of these different kinds of systems are integrated into considerably more homogeneous data communications networks. There is room for regional and national approaches, different kinds of participants, and variations in style. But, more than likely, the banks will have to build future terminals and systems that can enable the data communications infrastructure established by the supermarkets, large and small retailers, gas stations, hotels, and airlines to become increasingly interdependent.

Only when these participants become interdependent will the cost per transaction for EFTS decline and become acceptable; and only in that way will the nation's relatively fixed communications resources be used effectively.

Cross currents in costs: The technology barrier has been crossed. The question is not whether the technology is available but when it will be cost effective.

Advances in the last few years—minicomputers, computers built from microprocessor chips, and general advances in semiconductor technology—have cut the cost of EFTS components. Only mechanical components like printers have been stubborn. Electronic technology is clearly experiencing a declining cost curve.

However, not everything is becoming cheaper. The trend toward "distributed intelligence" (that is, computers spreading from the back office to the selling floor) means increased attempts to design unique systems, expanding needs for software, and a sure boost in the needed number of computer programmers (whose salaries are considerably higher than those of the clerical help they replace). As a result, costs will not decline as fast as might be hoped.

There are many caveats for the unwary entrant into EFTS:

- Watch out for software and consider it an unstable expense category. Its cost can outstrip that of hardware.
- Watch out for uniqueness for its own sake. Many different designs may be adopted to EFTS. There is no single perfect design.
- Watch out for overloading computer intelligence. You do not need a steam roller to crack a walnut.
- Watch out for unrealistic expectations. Every system installation will take longer than you think, and therefore it will cost more.

SUPPORTS FROM CONSUMERS

Bank marketing people are betting that consumers will be so interested in EFTS services that they will shift their deposits to institutions offering the new systems, and perhaps will even be willing to pay additional fees for them. Although merchants are also a source of banking revenue, they will not go forward with EFTS until they are fairly sure of consumer acceptance. Consequently, a key question is whether consumers are ready for EFTS. Do they want it? Or is this market being created by a technology push?

Any survey would show that consumers lack knowledge of EFTS. But these surveys are misleading. Consumers are extensively and sometimes painfully aware of the peripheral activities that are inextricably related to EFTS—credit purchase, credit reporting, check cashing, check floating, plastic card usage, opportunity costs, and even postage. The successful marketing of EFTS will depend on the consumers' desires for increased conveniences. This is the major gamble, and in my judgment, it is a pretty good one to take.

At the moment, the consumer is less concerned about EFTS benefits than about bigness per se and any potential violation of his or her rights. The majority of letters and petitions received by the National Commission simply voiced opposition to EFTS without qualifications or further clarifications. The reasons and questions given are as varied as the technology is complex:

The Economics and Other Issues of EFT Systems

"EFTS will invade privacy."
"It is not as reliable as paper."
"How will errors be corrected?"
"Will my money be safe?"
"EFTS is another example of jobs lost to technology!"
"What are the real motives and intentions of the American financial community?"
"EFTS is a way for government to infringe on its citizens' rights and responsibilities."
"EFTS constitutes another technological interference with human interaction."

People have even likened EFTS to the Book of Revelation's prediction of a "Mark of the Beast," the anti-Christ who would lead the world to destruction immediately preceding the second coming of Christ. Clearly the consumers need to know more about the specifics.

A number of consumer concerns will require the business executive's attention. An important area is the sanctity of information describing an individual's financial affairs. Consumers will insist that their rights and responsibilities be at least as well protected and clarified with EFTS as they are with today's paper system.

The National Commission has gone to the side of consumer rights. In doing so, it has molded a partnership between government and business to anticipate and correct potential EFTS abuses of individual rights before the abuses can become realities.

In its report to President Carter and to Congress, the National Commission concludes that consumers benefit in a number of significant ways from EFTS.[3] Nonetheless, it strongly recommends that consumers be given a fair choice in selecting among available methods of payment. It further suggests that Congress enact a broad series of safeguards for consumers, including a procedure for contesting any government access to financial transaction information. It proposes that financial institutions be held liable for misuse of information concerning an individual's depository account as well as for fraud and computer errors that are clearly not the consumer's fault.

The National Commission's recommendations should probably be viewed as minimum protections for consumers. Consumer activists are likely to have more to say on this subject in the years ahead.

[3]NCEFT EFT and the Public Interest. Report to the President and the Congress (February 23, 1977).

Meanwhile consumer experience with EFTS has been rather good. Credit authorization in department stores, for example, is a tremendous success. Complaints, when they are heard, are most likely to come from those who have a history of abusing their credit.

Today's EFTS volume level is limited not by consumer acceptance but by consumer awareness. Where well advertised, as by the First National Bank of Atlanta, response has lived up to predictions. The bank dubbed the check verification program "Honest Face," installed terminals in local supermarkets, and invited everyone in Atlanta to participate. The bank is claiming transaction volumes in excess of $1 million per month, quite a success in terms of consumer acceptance.

J. C. Penney, Sears, Roebuck & Company, and Montgomery Ward could hardly function without their credit authorization programs. Even Security National Bank of Alaska is operating an off-premise automated teller machine system. Almost everyone seems to be getting into the act.

On the other hand, I know of a number of cases where several million dollars were invested in EFTS hardware and almost nothing in marketing. Without exception, these programs have failed. For the present, consumers are coming out way ahead.

In many regions of the country they are receiving substantially increased services from EFTS-type programs with little change in cost.

LEGISLATIVE OBSTACLES

Political and legal realities may pose the greatest obstacle to the rapid growth of EFTS.

On the surface, the key issue is whether existing laws and regulations should be stretched, convoluted, and molded around this embryonic industry, or whether new standards and regulations should be created. The National Commission has suggested the latter course. But creating a national policy for EFTS development, as the National Commission proposed in its report in February 1977, may be far easier than executing it.

The key stumbling block is the duality of government regulation. Who will be the EFTS "guru"—the federal government or the states? Sufficient guidance is not provided by the two laws that today guide the federal and state roles in regulating and guiding the development of banking—the McFadden Act of 1927 and the Banking Act of 1933. Senator Thomas J. McIntyre's Banking Subcommittee is heading up an inquiry to reexamine these laws and test their

The Economics and Other Issues of EFT Systems

applicability in today's world. Currently, the development of EFTS is plagued with the incongruity of having to depend on the McFadden Act as the primary source of public policy guidance.

The banking industry is in the throes of other struggles to maintain competitive balance among all types of depository institutions. The Senate has debated the issues extensively since the 1970 Hunt Commission report and has developed two legislative models for reform, the proposed Financial Institutions Acts of 1973 and 1975. The House counterpart is the proposed Financial Reform Act of 1975. Some of the recommendations in these acts could decide the future form of EFTS.

In my opionion, Congress and the state legislatures must eventually make the decisions that will determine the who, where, what, and when of EFTS. The National Commission believes that the barriers to EFTS development should be lowered—that they are artificial and violate the interest of the public. EFTS development should not be tied, for example, to the laws prescribing the conditions for approving a new bank branch. Such a condition is altogether too restrictive.

Left unanswered is the question of whether the legislators will act, either in the Congress or in the states, upon various recommendations made forthrightly by the National Commission. The commission has proposed clear guidelines that cut through the rhetoric of a number of vested interests supporting mutually competing points of view. But the National Commission has no power to legislate.

What are the implications of these realities? My view is that they will not stop the steady growth of EFTS. However, they are likely to keep EFTS from spreading rapidly.

In conclusion, let me offer these observations and questions for businesses pondering the future of EFTS:

- The cost of learning how to establish EFTS is high but, if carefully managed, worthwhile.
- In spite of unfavorable short-run economics, many new commitments to EFTS are being made. This is because the long-run models look good.
- Since depository institutions have never been immune to "go-go" urges, particularly during the past 15 years, their managements should beware of the "I can't afford not to" syndrome. Many of the world's great errors have been based on such thinking.
- President Carter, California's Governer Brown, and other leaders talk about limited resources and lowered expectations. Does this represent the America of the future? Is EFTS consistent with such an America?
- In a society with high unemployment and low energy resources, are the technological advances represented by EFTS a step forward?
- From technological forecasting to future scenarios of the Brave New World-type, there seems to be almost unanimous opinion that EFTS is coming. Although we may have no choice about this reality, we have other choices. We can mold the shape and growth of EFTS. We can't change the answer to the question, "Whether EFTS?" but we can influence the answer to the question, "How much EFTS and when?"[4]

---

[4]The National Commission on Electronic Fund Transfers, of which Mr. Benton is executive director, has produced a report entitled EFT and the Public Interest. The stock number of this report is 048-000-00296-0, and it may be obtained by writing to the U.S. Government Printing Office, Superintendent of Documents, North Capitol and G Streets, N.W., Washington, D.C. 20402.

How to Conduct a Distributed Processing Feasibility Study

Problem:

Distributed processing is not feasible without a good communications system. Period. The avowed purpose of this service is to guide you through the concepts, planning, design, implementation, and management of a good communications system, but the communications system is simply a means to an end. It is not an end in itself. The communications system is a medium for one or more applications, not the least of which is distributed processing. In reality, of course, you do not plan for and build a communications system and then cast about for an application that can use the system. You begin with the application and then devise a communications system that will make the application work easily, cheaply, and accurately. This report takes you back one step further in the planning sequence. It will help you to recognize if your existing processing system can benefit from distribution. If it can, then the decision to distribute your processing capabilities must become an early and important factor in your communications system planning.

Solution:

Any elaboration of distributed processing immediately exposes the issue of decentralization versus centralization. Initially, we shall examine this important aspect, with emphasis on decentralization of data processing tasks, i.e., distributed processing at the local and regional levels and centralization of functions that cannot be handled efficiently and economically at the lower operating levels.

Next, we shall discuss distributed processing design criteria that are very important for determining the feasibility or nonfeasibility of going distributed. If distributed processing appears to be a viable alternative for current DP operations, a feasibility study is initiated in which feasible system alternatives are developed along with their specific savings and costs and their intangible benefits. An analysis of this information forms the basis for selecting the best distributed processing system. An integral part of the feasibility study is the evaluation of the equipment to be utilized in the new DP environment. The equipment selection process concludes the feasibility study.

From *Distributed Processing Systems* by Robert J. Thierauf, Ph.D., C.P.A. Chpt. 3, pp 55-89. © by Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

DECENTRALIZATION VERSUS CENTRALIZATION IN A DISTRIBUTED PROCESSING ENVIRONMENT

Before discussing the pros and cons of decentralization versus centralization in a distributed processing environment, it would be helpful to define each. Fundamentally, centralization creates one functional unit within an organization. The unit has prime responsibility for providing information processing services for all of the operating units in the organization. On the other hand, decentralization

How to Conduct a Distributed Processing Feasibility Study

Centralized Processing

Advantages:

- Economies of scale in terms of hardware and software
- Improved efficiency in systems development and programming
- Better control of operations and standards
- Greater growth and expansion of CPU, I/O devices, and peripheral devices
- More efficient usage of magnetic tape storage, data bases, and data management software
- Improved total compatibility, thereby allowing for tighter control and security
- Capacity to process large, complex applications
- Reduction in duplication of effort

Disadvantages:

- Less flexibility for tailoring programs to meet end user's needs
- Higher total system costs for communication and transmission of data
- Lower reliability of total system
- More computing power and time allocated to overhead, i.e., operating system, for the computer
- Larger incremental expansion steps and costs
- Restricted and slower access to centralized files
- Higher cost of backup or redundancy

Decentralized Processing

Advantages:

- Capture of data input at the local level and placement of error correction where it belongs
- Tailoring output to place information in the hands of the user, particularly at the lower levels
- Lower total system communication costs
- Higher reliability for dispersed computer systems
- Less computer time allocated to overhead, i.e., operating system, for the computer
- Smaller incremental costs for expansion and off-loading from the main computer
- Easier and faster accessibility to local files
- Lower cost and more effective backup capability

Disadvantages:

- Higher costs due to duplication in terms of hardware, software, data, space, and people
- Duplication of input, operating procedures, and processing
- More difficult control of system development, programming, standards, and data bases
- Restricted growth, i.e., CPU power, storage capacity, and I/O device selection
- Possible incompatibilities among distributed equipment
- Application size and complexity restrictions

Figure 1. Advantages and disadvantages of centralized processing and decentralized processing

creates a functional unit within each operating unit, and it has the primary responsibility for servicing the information processing needs of the operating unit. The basic difference between centralization and decentralization, then, is the degree to which information processing decision making, authority, and responsibility are disseminated throughout an organization.

Many organizations in the past have centralized their computer operations for the purpose of reducing DP costs. The most frequent argument advanced in support of centralization is that it results in economies of scale—the first item in Figure 1. The reduced costs are the result of several factors: 1) decentralized small computers may have unused capacity; 2) individual small computers may be overloaded, generating pressure for upgrading equipment or purchasing service bureau time; 3) the costs in a single large installation for items such as floor space, electricity, air conditioning, and other facilities are less than in multiple small installations; 4) large installations need fewer support personnel than small installations; 5) large installations require fewer management and staff personnel; and 6) a large computer is more cost effective than a group of small computers. Likewise, large computers, with their higher internal speeds, greater primary storage, and higher channel capacity, may make certain applications practical that are not feasible on smaller equipment. Examples of this include scientific computation, sophisticated database management systems, and the maintenance of and access to hierarchically structured files for manufacturing systems.

How to Conduct a Distributed Processing Feasibility Study

Another reason for centralizing is to improve efficiency in systems development—the second item in Figure 1. Centralization permits the design and use of common databases as well as common standards for data entry and input validation. It can also facilitate the use of development and project techniques that result in specific benefits to the organization.

Larger installations seem to attract and retain highly qualified technical people. These individuals can provide management with a wide range of alternative solutions to problems. This fact reduces the cost of development, operation, and future maintenance of the systems. Additionally, other reasons can be given for centralizing data operations. These are noted in Figure 1. Likewise, the disadvantages of centralization are set forth in this illustration.

The reasons advanced for centralization are generally based on efficiency. In contrast, the arguments for decentralization deal with effectiveness. Until recently, no one could argue that decentralization offered anything but added cost. However, in the last few years, the minicomputer has held out the promise of substantial savings. A single-purpose mini, programmed for a specific application, is relatively inexpensive. If it is used as an office machine, it does not require a trained operator or the programming and technical support that a general-purpose computer does. Some minis can provide on-line inquiry, saving the cost of telecommunications. The high cost of communications, the overhead associated with large general-purpose computers, and the possibility that a large installation's capacity will not be fully utilized combine to weaken the case against decentralization.

In the area of applications, there are problems when an organization attempts to meet local needs from a central site. Applications developed for a centralized operation are often far more complex and costly than those developed for local needs. Also, maintaining the system for one operating unit could potentially affect all units. If the central computer becomes inoperative for any reason, all units are adversely affected. Not only are the risks increased, but centralization forces operating units into a common mold that may be inappropriate for their needs. These different needs could be satisfied by smaller installations at reduced cost and complexity.

Advocates of decentralization state that local analysts are more attuned to local needs and have a deeper understanding of local operations, managerial preferences, and organizational structures. This enables them to establish requirement specifications and to design systems that are best suited for the local user. The local analyst can also respond more quickly to emergencies and changes in priorities of local management. In contrast, a manager at the local level in a centralized environment has to battle with other users for the central system's development resources. In addition, the close association between the analyst and the user means that the user will be better informed about the benefits and limitations of data processing. The user can assume tighter control over DP personnel and the quality of their work.

Even though most centralized installations allocate their costs to users according to the resources used, the local manager has little responsibility for total data processing costs. Salaries paid to central personnel, overhead rates, choice of equipment, time spent on projects, and share of resources used all seem to be beyond the manager's control at the local level. As a consequence, the allocations are viewed as arbitrary. The manager's only objective is to obtain as much service as he or she can from the centralized installation. In the long run, this drives up costs. However, if the data processing resource is local, the manager has direct knowledge of all the elements of cost and an incentive to control them. Thus, the manager at the local level knows that there is better control over operations in a decentralized environment.

Having stated briefly the important benefits for both centralization and decentralization (refer to Figure 1), we can now raise the question, Which is the best approach for distributed processing? Typical business organizations would be well advised to combine the best features of centralization and decentralization for a distributed processing environment. To do this effectively, the feasibility study must evaluate what elements of both are best suited for the organization.

If the organization's data processing problems are in the area of service, cost, or effectiveness, it will be necessary for the feasibility study group to balance the requirements of the decentralized units against those of the organization as a whole. In this evaluation process, the DP manager should pay close attention to developing alternative organizational designs and comparative costs. In some cases, cost information may be all that top management needs to resolve the problem. However, if management is primarily concerned with service rather than cost, a different approach may be necessary. Thus, the feasibility study's objectives will determine what elements of centralization and decentralization will be found in the distributed processing system.

In some cases, certain functional activities for marketing, manufacturing, physical distribution, and accounting that can best be handled at the corporate level are left at that level. However, many data processing functions are off-loaded from the corporate

How to Conduct a Distributed Processing Feasibility Study

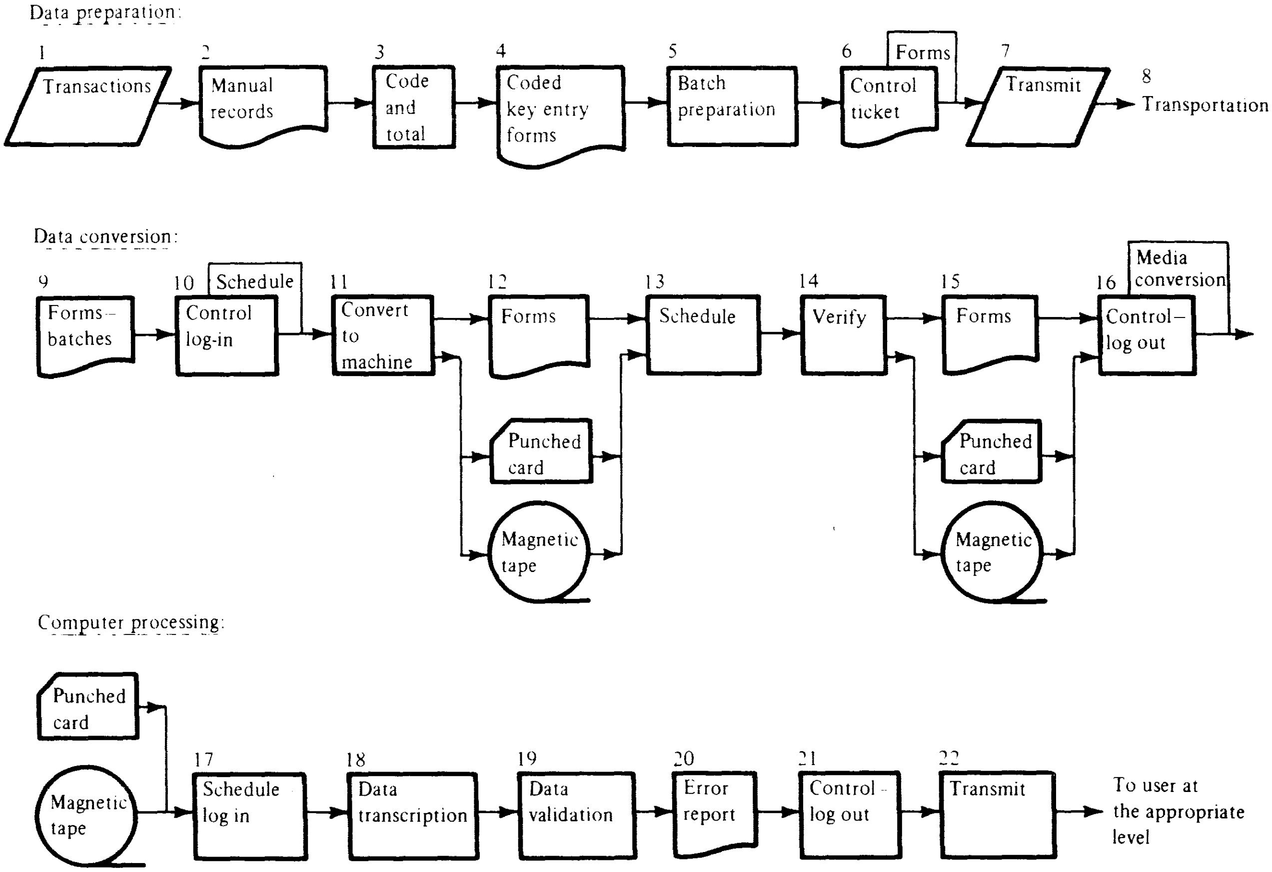


Figure 2. Typical DP system of personnel, equipment, procedures, and computer programs found in many organizations

level to the local level. In this manner, the benefits of centralization and decentralization are obtained by this combined approach to distributed processing.

CONSIDERATIONS FOR DESIGNING DISTRIBUTED PROCESSING SYSTEMS

Prior to discussing general and specific design criteria for a distributed processing environment, it is helpful to examine a typical data processing system, such as that illustrated in Figure 2. In such a system, there are many procedural steps involved in the processing of data. Fundamentally, eight steps are involved in the data preparation phase, and the same number of steps are needed for data conversion. In turn, a number of steps is required in the computer room for validity checking and control whether computer processing is performed at the local, regional, or headquarters level. In effect, a long and drawn-out structure of personnel, equipment, procedures, and programs has been built to convert organizational data to meaningful information. Looking at the structure another way, many of the data preparation and data conversion procedures are not equipment oriented but rather entail manual handling, procedures, and data control. In fact, studies have indicated that more than half of all information processing efforts concern just these functions.

In view of these realities of a typical DP installation, a distributed processing system should demonstrate improved productivity where it is needed most—in the labor-intensive areas. Likewise, it should fit into the existing scheme of operations. If a distributed processing system fails on these two important points, there is generally little or no incentive to make the move.

When evaluating the move to a distributed processing environment, many design criteria from a general standpoint must be kept in mind. Similarly, specific design criteria peculiar to these systems should be reviewed. General and specific design criteria, as set forth below, provide the systems designer with adequate guidelines for devising an effective distributed processing system.

GENERAL DESIGN CRITERIA FOR DISTRIBUTED PROCESSING

There are many important design criteria within a distributed processing system, or, for that matter, within any type of information system. Implied from a management viewpoint in the development of systems design, the distributed processing system must be designed to meet management goals and objectives, especially the degree of decentralization contemplated. The goals and objectives must be satisfied at all levels of the organization's management. Likewise, it must provide various levels of management within the organization with the necessary elements of information for planning, organizing, directing, and controlling present and future operations. This approach implies summary reviews for top levels of management and more detailed information concerning quantitative and qualitative data for management directly responsible for operations.

From a general design viewpoint, there are several important criteria, which are set forth in Figure 3. Although these are discussed below, any distributed processing system that is not acceptable to organizational personnel will generally fail even though it encompasses all of the other general design criteria found in a well-designed system.

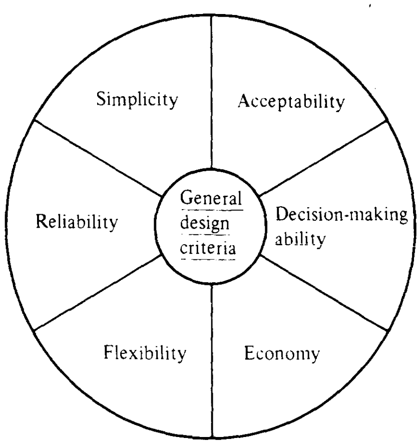


Figure 3. General design criteria for distributed processing systems

Acceptability

Success of a new distributed processing system pivots on its acceptance or rejection by company personnel. If operating personnel are convinced that it will not benefit them, that it is a poor system, or that it does not follow established company policies or have some other legitimate reason, the new system is in serious trouble. To overcome this resistance, their participation is essential, particularly during the design and implementation phases, for in reality they constitute the organization that must use and live with the newly designed system. In effect, the human element can "make or break" any distributed processing system.

Decision-Making Ability

An effective distributed processing system allows decisions to be made on a timely basis. Similarly, it must be able to utilize techniques that can allocate available resources most effectively. Hence, a distributed system is designed to make decisions as efficiently and automatically as possible. Decisions that cannot be made automatically as a result of the nature of the system must be relegated to the appropriate level of management or nonmanagement personnel, thereby making the firm's objectives obtainable.

Economy

For economic operations, data should be captured or created in machine language as near to the source as possible and allowed to flow through the distributed processing system automatically from that point on. Activities that must be performed in sequence should be located as closely together as possible, both organizationally and physically. Eliminating duplicate information files, reducing the provision for every possible contingency, and eliminating small empires in the firm's functional areas are other examples of economically improving DP operations.

The question of system cost versus the potential savings must be considered. Generally, no information or service should be produced that is not justified by its cost to the firm. It is often expensive to develop one functional area of the overall system that has greater capacity than its integrated and related parts. In essence, there is need for a proper balancing of the distributed systems and their related parts to effect economy of operation. For example, it makes no sense to develop fast-order processing methods and procedures for improving customer service if the company's plants are not geared on a comparable basis for providing prompt service.

In deciding whether to centralize or decentralize the performance of the firm's functions, the economy of specialized operations and the elimination of duplicate functions must be compared with the reduction of communications and paperwork costs under decentralized conditions. Also, the reduction in time cycles, the increased flexibility, and the greater unity of control and responsibility possible under decentralized operation must be considered.

Flexibility

To be effective, the distributed processing system must be flexible—capable of adapting to changing environmental conditions. There will always be variations of products, manufacturing processes, and accounting procedures, to name a few. Managers must be prepared to adjust their DP operations to changing conditions. Without this ability, the firm may lose customer goodwill as well as encounter problems with its own personnel. Thus, a well-designed distributed processing system should be able to withstand change by providing for ease of expansion and for additional production capacity.

Reliability

Reliability of the distributed processing system refers to consistency of operations. In other words, are data input, processing methods and procedures, and output information consistent over an extended period of operation? The degree of reliability can range from a constant and predictable mode of operation to a complete breakdown of the system. Although most distributed processing systems do not operate at these extremes continuously, they do operate somewhere between them. A high degree of reliability can be designed into the system by providing for good internal control—that is, numerous control points where variances from established norms and practices can be detected and corrected before processing continues. Control functions should be allocated to organizational units that are independent of the functions to be controlled. In all cases, controls should be an integral part of a distributed processing system.

Simplicity

The trademark of any effective distributed processing system is simplicity. Simplicity can be effected by providing a straight-line flow from one step to the next, avoiding needless backtracking. Input data should be recorded at the source, or as close to it as possible, to reduce or eliminate the need for recopying. Functions should be assigned to organizational units in a way that will reduce the need for coordination, communication, and paperwork. To state it another way, the distributed system should be easy to use. In addition, each organization group should have the authority and responsibility for its area and be held accountable for its performance to one superior only. Thus, the simpler the distributed processing system is, the fewer chances there are of having major problems and foul-ups.

SPECIFIC DESIGN CRITERIA FOR DISTRIBUTED PROCESSING

Going beyond the general design criteria explored above, there are several specific design criteria that must be considered in a distributed processing environment.[1] First, procedures are as important as technologies, perhaps even more important. The integrity of data depends more on the way it enters, moves, and is screened in a system than on any other factor; much of this processing, in turn, depends on the labor content. The conclusion to be drawn is that a distributed processing system should either displace or support labor content. In this manner, the system will be cost effective.

Second, for data handling to be accurate and efficient at the local or higher levels, it requires measurement. The more precisely the measurements can be made, the more accurate and efficient they will be. This means that any systems designer planning to convert to distributed processing equipment must insist that a vendor provide a clear methodology on how to measure performance at every level—from the operator interaction to central-site computation. In turn, such measurements are critical for determining cost allocations at the local and higher levels.

Third, from the viewpoint of network effectiveness, it must be a function of the system and not depend on any individual operator, group of operators, or departments. By definition, a distributed system is used by all kinds of persons in all types of environments—local, regional, and central; skilled and unskilled; rigidly controlled and uncontrolled. Thus, the control of quality, accuracy, and completeness of the entire system is paramount. Such an approach leads to overall optimization of the entire distributed processing environment versus letting one area exercise control over the new system requirements.

Last, systems productivity must be planned for, insisted upon, and expected from the system. It is not a natural byproduct of the prior criteria. The goal should be to improve performance on the part of each local operator, local site, and central component in the system. Otherwise, there can be no real economic justification for going to a distributed system.

DISTRIBUTED PROCESSING SYSTEM—FEASIBILITY STUDY

Having established the important design criteria for distributed processing, the DP manager and his (her) staff will find the process of evaluating alternative systems much easier. Fundamentally, evaluation of such systems is a process of matching the desired functions to be performed against the capabilities of available equipment. For best results, an in-depth feasibility study is necessary to determine what functions can be distributed and what equipment can be acquired to perform the functions. To do otherwise generally leads to less than optimum results.

To determine the feasibility of going distributed, the study team should accent the following items. First,

[1]Refer to William G. Moore, "Going Distributed," Mini-Micro Systems, March 1977, p. 44.

the feasibility study should determine whether the present system will benefit from distributing tasks for both present requirements and future needs. If the organization is diversified and the current system must allocate a large percentage of processing time to communications, it is likely that distributed processing could prove a valuable alternative.

Second, it is important to consider the corporate structure, the size and location of the organization, and its operating units. In a very large corporation, the concept of one totally centralized computer operation, serving everyone, is a difficult if not an impossible task. Hence, the need for distributed computing would be apparent. However, an organization which has many similar operating units, all requiring the same basic data, is generally not a prime candidate for distributed processing.

Datapro Comment:

This last point is extremely important and deserves to be highlighed because it is the fundamental go/no-go determinant of whether to consider distributed processing. For example, consider an accounting firm that works with a common pool of procedures relative to variable input data. In this case, even more operational centralization is desirable because everyone is doing essentially the same jobs with the same programs. The database may be distributed, especially if the firm's accounting services are partitioned into industry groupings, but even here a good security/access system could permit all customers' data to be relegated to a central communicating database. Distributed processing, for this example, would be a totally unnecessary complication.

Last, it is advisable to monitor current system activity through software monitors provided by the computer manufacturer, or software and hardware monitors offered by independent manufacturers, to determine the percentage of mainframe processing time being spent on communications housekeeping tasks that could be distributed.

Upon completion of these aspects of the feasibility study, a clear picture of the current data processing environment should emerge. If the overhead for the computer system is not too great and the projected increase in work load can be handled, the current computer system is probably adequate, and the low work load will make distributed processing unnecessary. In these cases, however, the desire to distribute processing may not be related to the computer's work load. Rather, it may be desirable because it offers better control over input/output functions of divisions or subsidiaries with specialized processing needs. In addition, the central processing center is too large to be economically or functionally responsive to user needs. An in-depth feasibility study of these factors, then, should provide sufficient information to determine the economic trade-offs involved in a distributed processing system versus a centralized processing system.

FEASIBLE DISTRIBUTED PROCESSING SYSTEM ALTERNATIVES

Before feasible distributed processing system alternatives can be developed by the study team, proposed system requirements must be clearly defined. They are determined from the desired objectives initially determined in the study. Likewise, consideration is given to the strengths and the shortcomings of the existing system. The distributed processing requirements that must be clearly defined and conform to the study's objectives are as follows:

1. Outputs to be produced, with emphasis on managerial reports that utilize the exception principle.

2. Data files (databases) to be maintained with on-line and off-line processing capabilities.

3. Types of input/output terminals to be utilized in the appropriate processing mode.

4. Input data from original source documents for processing by the system.

5. Methods and procedures that show the relationship of inputs and outputs to the data files (data bases).

6. Work volumes and timing considerations for present and future periods, including peak periods.

One starting point for compiling the above requirements is the outputs. After they have been determined, it is then possible to infer what inputs and on-line and off-line files are required and what methods and procedures must be employed. Although it is possible to start with the inputs, the output-to-input procedure is recommended because the outputs are related directly to the firm's objectives, the study's most important consideration. The future work loads must be defined for the inputs, the data files (databases), and the outputs in terms of average and peak loads, cycles, and trends.

Flexible System Requirements

The requirements of the new system may appear at first to be fixed. A closer examination, however, often reveals that these specifications have flexibility. For example, the objectives set forth in the study state that certain file data must be updated once a day. Perhaps the best solution is to incorporate the data into a database that is updated as actual transactions occur. This approach is within the constraints as initially set forth and introduces a new way of maintaining files. The important point is that alternative methods are available in data processing areas that may have the outward appearances of being fixed. With this approach in mind, it is possible to design a number of different systems with varying features, costs, and benefits. In many cases, more alternatives will be investigated and analyzed when flexible system requirements are considered.

Consultant's Role in Feasible System Alternatives

A clear understanding of the new system requirements is the starting point for developing feasible system alternatives. This phase is by far the most important and difficult undertaking of the study to date. An outside consultant's experience is of great value to the study group. His (her) knowledge of many installations can help immeasurably to reduce the number of possible solutions for the firm. Too often, a study group goes off on a tangent about a specific systems approach that should have been discarded as unfeasible. The outside consultant can make certain that time is not wasted on trivial matters. Also, he can point out the shortcomings of a certain approach that may have been strongly advocated by certain DP personnel. The consultant's objectivity in judging the merits and weaknesses of a new distributed processing system can enhance the firm's chances of selecting an optimum one. The key to developing promising system alternatives and selecting the optimum one is to employ fully the talents and experience of the DP feasibility study group.

SAVINGS, COST FACTORS, AND INTANGIBLE BENEFITS FOR EACH ALTERNATIVE

After developing feasible distributed processing alternatives, the next step is to determine the estimated savings and incremental costs for each alternative. Estimated savings (sometimes referred to as cost displacement) are enumerated in Figure 4. Incremental costs are segregated into two categories: one-time costs and additional operating costs. These are listed in Figure 5. The difference between the estimated savings and estimated one-time costs and additional operating costs represents the estimated net savings (losses) to the company before federal income taxes.

Accurate figures for a five-year period are of great importance, which indicates a need for accounting

- Reduction in the number of personnel, lower salaries and wages
- Lower payroll taxes and fringe benefits with fewer people
- Sale or elimination of some equipment—depreciation and/or rental—no longer applicable
- Reduction in repairs, maintenance, insurance, and personal property taxes
- Lower space rental and utilities
- Elimination or reduction in outside processing costs

Figure 4. Feasibility study—estimated savings.

Estimated one-time costs:

- Feasibility study
- Training of programming and operating personnel
- Documentation of all feasibility study applications
- Programming of these applications
- Program assembly and testing of programs for new system
- Data file (database) conversion
- Site preparation (includes construction costs, remodeling, air conditioning, and power requirements)
- Parallel operations (the old and the new system operate concurrently—duplication of personnel and equipment for a given time period)
- Conversion activities (from existing system to new system)
- Other equipment and supplies (includes forms-handling equipment, files, magnetic disks, and magnetic tapes)

Estimated additional operating costs:

- Data processing equipment (processors, computers, and related equipment)—monthly rental and/or depreciation
- Maintenance of equipment (if not included above)
- Program maintenance (programmers)
- Wages and salaries of data processing personnel (direct supervision, equipment operation, and other data processing jobs), payroll taxes, and fringe benefits
- Forms and supplies (for new data processing equipment)
- Miscellaneous additional costs (includes insurance, repairs, maintenance, and personal property taxes on equipment purchased; power costs; etc.)

Figure 5. Feasibility study—estimated one-time costs and estimated additional operating costs

How to Conduct a Distributed Processing Feasibility Study

	Years from Start of Systems Implementation					Five Years Total
	1	2	3	4	5	
Estimated Savings:						
Reduction in personnel (including payroll taxes and fringe benefits)	$120,200	$400,500	$440,300	$490,500	$540,500	$1,992,000
Sale of equipment	120,000					120,000
Rental (space) savings	25,000	51,000	54,500	58,000	61,800	250,300
Elimination of rental equipment	2,050	4,380	4,690	5,000	5,300	21,420
Other savings	3,000	3,060	3,210	3,370	3,540	16,180
Total Estimated Savings	$270,250	$458,940	$502,700	$556,870	$611 140	$2,399,900
Estimated One-Time Costs:						
Feasibility study (for this year and prior year)	$ 95,000					$ 95,000
Training	50,000					50,000
Systems and programming	255,500					255,500
Data base conversion	272,500					272,500
Other conversion activities	75,500					75,500
Site preparation	55,400					55,400
Other one-time costs	22,300					22,300
Total Estimated One-Time Costs	$826,200					$ 826,200
Estimated Additional Operating Costs:						
Data processing equipment rental (include maintenance)	$110,000	$120,800	$127,400	$134,100	$141,000	$ 633,300
Additional personnel for new system (includes payroll taxes and fringe benefits)	34,000	60,700	62,300	63,400	64,600	285,000
Program maintenance	20,000	30,700	32,200	33,800	36,000	152,700
Forms and supplies	10,000	21,500	23,000	24,500	26,000	105,000
Other additional operating costs	4,400	12,400	12,800	13,200	17,600	60,400
Total Estimated Additional Operating Costs	$178,400	$246,100	$257,700	$269,000	$285,200	$1,236,400
Net Savings (losses) before Federal Income Taxes	($734,350)	$212,840	$245,000	$287,870	$325,940	$ 337,300

Figure 6. Feasibility study—distributed processing system alternative 3 of net savings (losses) for a five-year period (rental basis)

department assistance. Often the best way to increase the accuracy of the figures compiled by the study group is to have the outside consultant assist the group and review the data. His (her) knowledge of current data processing equipment will save time in this phase of the study. His exposure to other similar cost studies will add creditability to the final figures, including the selection of a distributed processing system.

Projected Savings and Cost Factors

It is not desirable to base the estimates of savings and incremental costs on the present data processing work load. Rather, the trend of growth or cutback in the firm's work load should be analyzed and projected for the next five years. These data can then be utilized to project savings and costs, as shown per the analysis in Figure 6. In this feasibility study for alternative 3, consideration has been given to higher future costs. Salaries and wages are generally increased by five percent. Cost reduction through work simplification in the present system has been incorporated in the analysis.

Because the projected savings and costs factors in a feasibility study are for five years (starting with systems implementation), the difference between the two sums, after taking into account federal income taxes, should be discounted back to the present time. The purpose of the discounted cash flow is to bring the time value of money into the presentation. This is shown in Figure 7 for system alternative 3. Notice that the net savings after federal income taxes of $175,396 over the five-year period (anticipated life of the system), when discounted, shows a negative present value for this alternative of $27,229. With a discounted 20 percent estimated return on investment for this alternative, this one should not be chosen (the firm's cutoff point for capital investments is 20 percent). Even though the revised discounted rate of return is approximately 16 percent (based on present value factors), additional benefits should be considered.

Intangible Benefits

While the foregoing calculations have taken into account the projected savings and costs or quantitative

Year	*Net Savings (Losses) Before Federal Income Taxes (Per Figure 3-6)*	*Federal Income Tax @ 48% Rate*	*Net Savings (Losses) After Federal Income Taxes*	*At 20%* Present Value of $1	*At 20%* Present Value of Net Savings (Losses)
1	($734,350)	($352,488)	($381,862)	.833	($318,091)
2	212,840	102,163	110,677	.694	76,810
3	245,000	117,600	127,400	.579	73,765
4	287,870	138,178	149,692	.482	72,152
5	325,940	156,451	169,489	.402	68,135
Totals	$337,300	$161,904	$175,396		($ 27,229)

Figure 7. Feasibility study—distributed processing system alternative 3 discounted cash flow based on 20 percent return after federal income taxes (rental basis)

factors, a number of intangible factors or qualitative factors will be uncovered by studying the potential contributions of the new distributed processing system. A list of these factors is found in Figure 8. Even though qualitative factors are nonquantifiable initially, their ultimate impact is in quantitative terms, reflected in the financial statements.

An analysis of Figure 8 indicates that the intangible benefits of the system ultimately offer two major benefits: increased revenues and decreased costs. Better customer service and relations should enable

- Improved customer service through faster order processing and inquiry capabilities.
- Ability to handle more customers faster with custom-designed equipment.
- Needed information is more readily available in formats which are oriented toward the operating levels.
- Processed information is more accurate because data input error rates are reduced.
- Efficiency of the central processing facility rises as it is relieved of many time-consuming tasks.
- Better decision-making ability through more timely and informative reports.
- Closer control over capital investments and expenses through comparisons with budgets or forecasts.
- Improved scheduling and production control, resulting in more efficient employment of men and machines.
- Greater accuracy, speed, and reliability in information handling and data processing operations.
- Better control of credit through more frequent aging of accounts receivable and analysis of credit data.
- Greater ability to handle increased work loads at small additional costs

Figure 8. Feasibility study—intangible benefits from a distributed processing system

the organization to increase sales to its present customers and to many potential ones who are looking for these characteristics in its vendors. A distributed processing environment affects the organization not only externally but also internally in terms of faster and more frequent reporting of results. In addition to accuracy, speed, and flexibility, distributed processing equipment allows operating management at the local and regional levels to plan and organize activities and, in turn, to direct and control according to the original plans.

SELECTION OF THE BEST DISTRIBUTED PROCESSING SYSTEM

Once a thorough analysis of the important factors has been completed, the DP manager and the study group will be in a position to compare the system alternatives. Although there are several approaches to evaluating alternatives that can be employed for a definitive conclusion to the feasibility study, the decision table approach, shown in Figure 9, is used. A decision table is helpful in resolving a complex management decision because it assembles all factors for every feasible system alternative. A complete summary of all factors pertinent to making the important decision should be a part of the decision table for it to be a fully effective management tool.

The conditions in the upper part of the figure represent benefits (tangible and intangible), while the lower part shows the possible courses of action. Each distributed system alternative represents a set of actions corresponding to a certain set of conditions. In this case, system alternative 3 indicates the highest return or 16 percent for distributed computing, although system alternative 4 has the same answers but has a slightly lower return of 15 percent. Their intangible benefits must be reevaluated for a final decision. Thus, an examination of Figure 9 indicates that alternatives 3 and 4 are best.

Now the question is which one of these alternatives should be implemented. On the surface, both have about the same benefits, except that alternative 3 gives a higher return on investment. A closer inspection, however, reveals that only alternative 4 utilizes many more intelligent terminals than alternative 3. Conversion today will mean no or minimal conversion costs in the future for this type of equipment. With this added advantage, the DP manager feels the future cost savings justify accepting a lower return. Therefore, his recommendation has been finally reached. As with all feasibility studies, all information gathered must be documented not only for future reference but also to serve as a framework for selecting the appropriate equipment.

How to Conduct a Distributed Processing Feasibility Study

Decision Table	Table Name: Feasibility Study–Distributed Processing System Chart No: FS-ES-1	Prepared by: R. J. Thierauf	Page 1 of 1 Date: 7/25/7-

Condition	Rule Number 1	2	3	4	5	6	7	8	9	10	11	12
Tangible benefits:												
Meets return on investment criteria–20% after taxes*	N	N	N	N	N							
Lower order processing costs	Y	Y	Y	Y	Y							
Lower investment in inventory	Y	Y	Y	Y	Y							
Less future cash requirements	N	Y	Y	Y	N							
Intangible benefits:												
Improved customer service	N	Y	Y	Y	Y							
Improved promotional efforts	Y	Y	Y	Y	Y							
Ability to handle more customers faster	N	N	Y	Y	Y							
Better decision-making ability	Y	Y	Y	Y	Y							
More effective utilization of management's time	Y	Y	Y	Y	Y							
Improved scheduling and production control	Y	Y	Y	Y	Y							
Closer control over capital investments and expenses	Y	Y	Y	Y	Y							
Better control of credit	N	N	Y	Y	Y							
Ability to handle more volume at lower costs	Y	Y	Y	Y	Y							

Condition	Rule Number 1	2	3	4	5	6	7	8	9	10	11	12
More accuracy and reliability of data	Y	Y	Y	Y	Y							
Greater utilization of mathematical techniques	Y	Y	Y	Y	Y							
Action												
Utilizes a distributed processing system	X	X	X	X	X							
Utilizes remote batch processing	–	–	X	X	X							
Minor changes of inputs and outputs	X	–	–	–	–							
Substantial changes of inputs and outputs	–	X	X	X	X							
Need for distributed data bases	X	X	X	X	X							
Moderate revision of methods and procedures	X	X	–	–	–							
Complete revision of methods and procedures	–	–	X	X	X							
Employ an additional consultant for study	–	–	X	X	X							
Recruit new data processing personnel	–	X	X	X	X							
Great use of intelligent terminals	–	–	–	X	X							

**Other Information: 1-14%; 2-15%; 3-16%; 4-15%; 5-13%*

Figure 9. Feasibility study—decision table for appraising feasible distributed processing system alternatives

BASIC TYPES OF DISTRIBUTED PROCESSING EQUIPMENT

Currently, a second industrial revolution, based on machines that can add decision making, arithmetic, and memory to their usual functions, is evident in distributed processing systems. The brain of this revolution is the microprocessor. Even though there is no general agreement on a precise definition, the microprocessor is a tiny chip of silicon, the size of a pencil eraser, that provides the arithmetic and logic capabilities of yesterday's large computers. The equivalent of a network of wires, switches, and transistors has been fabricated on the chip's surface.

Basic types of distributed processing equipment utilizing new technology are set forth below:

- remote batch terminals
- smart and intelligent terminals
- communications controllers
- microprocessors and microcomputers
- minicomputers and small business computers

Within the discussion of each piece of hardware, the important component parts will be set forth as they pertain to distributed processing systems.

Remote Batch Terminals

Most traditional remote batch terminals offer very little for the user interested in a true distributed processing system. However, the low cost of microprocessor technology has allowed manufacturers to add more capability to their batch terminals and incorporate local processors to create small business systems. The growing acceptance of distributed processing has encouraged these advances to the point where some systems even provide interactive processing and batch processing. Such terminals not only perform local job processing and interactive data entry but also support peripheral magnetic tape drives, disk storage devices, and local interactive terminals. A few of these new terminals can even support remote interactive terminals. In essence, current hardware advances allow remote batch terminals to perform a wide range of distributed processing functions.

Smart and Intelligent Terminals

The user should be careful to distinguish between "smart" or microprogrammable terminals and truly "intelligent," user-programmable terminals (refer to Section CS15 for more information). Smart terminals do little to off-load the mainframe processing tasks. The only local data validation they can do is to restrict entry fields for alpha only or numerics only. And while some smart terminals have a buffer, they lack the necessary storage. Generally, microprogrammable terminals should not be considered vehicles for distributed processing. These terminals are not user-programmable. Changes to the logic must be done through the internal read-only memory, a task that must usually be performed by the manufacturer. Also, the architecture of most of these terminals limits the functions that can be changed to character generation and communications procedures, not data processing.

On the other hand, intelligent terminals are capable of performing more tasks than smart terminals, because they have some memory and control functions built in for programming. Within a programmed environment, they can stand alone or be configured in a cluster, sharing any or all of the computer power, storage, printers, and sometimes communications. CRT is the most common display technology for these types of terminals.

Generally, an intelligent terminal must have, as a minimum, the following characteristics:

- self-contained storage, random access memory
- user interaction with the terminal itself
- stored program capability
- processing capability at the terminal through a user-written program
- capability of on-line communications with another intelligent terminal
- human-oriented input, such as keyboard
- human-oriented output, such as a printer or a CRT

The trend today is for programmed intelligent terminals to perform multiple functions in data entry, data retrieval, inquiry/response, and monitoring and control. And this trend is expected to continue as the office becomes more automated. In response to this trend, terminal vendors are gearing up for the distributed processing era.

Communications Controllers

Functions that can be removed from the mainframe by communications controllers include error recovery, code conversion, polling, and network control. But like most intelligent remote batch terminals, intelligent communications controllers do no more than emulate their hardwired counterparts. When choosing a communications controller for a distributed system, the provision of adequate software support for these functions and implementation of these functions without major changes to the existing mainframe-resident communications software must be considered.

Microprocessors and Microcomputers

Microprocessors provide the underlying technology for distributed processing equipment. Similar to the central processing unit (CPU) of a computer, a microprocessor manipulates data by interpreting and executing coded program instructions. This general-purpose, data processing device is contained on large-scale integrated (LSI) circuits, which are produced by means of metal-oxide-semiconductor (MOS) technology. Fundamentally, a microprocessor consists of an accumulator, an arithmetic-logic unit, a scratch-pad, read/write memory, a register and decoder for instructions, a program counter and address register stack, a timing and control section, a parallel data and input/output bus, and a controller for data input and output.

Potential applications for microprocessors are wide open for distributed computing. Present equipment includes data entry devices, intelligent peripherals, dedicated processors, point-of-sale units, CRT terminals, printers, and a variety of other input/output units. An important advantage of microprocessor technology is that low-cost computing power is made available for equipment at the local processing level.

Going beyond the capabilities of the microprocessor and below the processing capabilities of minicomputers and small business computers is the microcomputer. Although the microcomputer is a smaller version of a minicomputer, there is a tendency to blend these two types of computers into a virtually indistinguishable product line. However, there are certain distinguishing characteristics. Microcomputers tend to differ from minicomputers by having smaller word size, slower memory cycle time, and more limited instruction sets; being lower in cost; using less power; and having custom-fitted controls for specific applications.

Based on these important features, a microcomputer can be defined as a microprocessor affixed with memory and input/output logic or circuits so that it can perform a useful function. To state it another way, when the microprocessor (CPU) is incorporated as a CPU in a working system along with a data storage memory, a program memory, and input/output circuitry, it is called a microcomputer. Some firms produce complete microcomputers—CPU, data storage memory, program memory, and input/output circuitry—on one or two MOS-LSI circuits. The tendency in such cases is to call the unprogrammed circuits a microprocessor system.

Due to the steady advance in circuitry sophistication and miniaturization, the logic and memory circuits of a microcomputer can be held in the palm of one's hand. It is possible to have tens of thousands of components on a single chip requiring only milliwatt power. Based on their size and capabilities, it is expected that microcomputers will have a decided impact on distributed processing. Similarly, they will affect our lives in much the same way electric motors have. A listing of their present and future applications for distributed processing and other areas would include store sales information systems, information processing systems, measuring systems, control systems, and education systems.

MINICOMPUTERS AND SMALL BUSINESS COMPUTERS

Minicomputers and small business computers are widely used for virtually all kinds of hierarchical distributed processing. Generally, they are located at multiple locations and are tied together via a data communications network. With this configuration, each small system functions as a combined data entry, computing, and printing system with the capability of performing simple processing tasks and transmitting the more complex processing to the larger system. This environment permits optimum use of the minicomputers and small business computers with the large computer in one integrated system.

Overall, the important distinguishing feature of minicomputers and small business computers is their level of communications support. Each machine should be able to support its own local and remote satellite terminals and perform highspeed communications with a host processor. Likewise, it should have a multitasking operating system—a system that is necessary for the machine to handle concurrent operations—and an extensive peripheral complement, especially a large disk subsystem to provide the necessary database storage.

In price and performance, minicomputers and small business computers span a wide range between conventional accounting machines and minicomputers at one extreme and medium-scale computer systems at the other. Though the current systems differ widely in their architecture, data formats, peripheral equipment, and software, today's minicomputer and small business computer systems typically consist of a keyboard/CRT for data entry (cards, floppy disks, or cassettes may also be used), a processor that starts with about 8K bytes of memory, a disk for file storage, and a serial printer with a speed of about 30 characters per second. From there, the only way to go is upward, i.e., more memory, additional peripheral devices, faster printers, etc.

The small business computer market is served by distinct types of vendors. The first type is the "Fortune 500" companies, such as Burroughs, Honeywell, IBM, Litton, and NCR, all of which have vast product lines and resources. For these firms, the small business computer is just one of a broad line of products (although in the cases of NCR and Burroughs, business minicomputers now account for a sizable portion of total corporate revenues). A second group consists of minicomputer manufacturers, such as Digital Equipment Corporation, Data General, Computer Automation, Harris, Hewlett-Packard, Interdata, Microdata, and others. This group has watched the small business computer marketplace mushroom in size and now wants a piece of the action. Their answer to this segment of the marketplace is a packaged configuration consisting of a minicomputer and associated peripherals from their current product line, usually accompanied by some applications software. Most minicomputer vendors also offer assemblers and compilers for the user who wants to do its own programming or to solve business problems that cannot be handled by packaged software.

Datapro Comment:

A third group of vendors who serve the small business user consists of relatively new firms like Jacquard, Basic Four, and Genesis. These firms are capitalizing on the fact that literally anyone can assemble a computing system from increasingly inexpensive off-the-shelf components. These firms specialize in producing very easy-to-use software and generally "aim" their products at a specific industry segment, like wholesale pharmaceuticals or retail lumber. Although they cannot offer the enormous support and product range of an IBM or DEC, they have been surprisingly successful because they approach a customer with a deep understanding of his exact needs and because they can afford to spend much more time with a $15-20,000 account than the larger companies. This willingness to "hold hands" from initial contact through to a smoothly running installation is not lost on the small, unsophisticated user.

The major applications software packages usually offered are accounts payable, accounts receivable, billing, inventory, payroll, and sales analysis. Some vendors offer a full library of applications programs, while others modify their software to the customer's needs for a negotiated price. It is important for the buyer to determine beforehand the kind and degree of software support being offered in a distributed processing environment.

EQUIPMENT CRITERIA FOR DISTRIBUTED PROCESSING

To implement the recommended distributed processing system (as determined in a previous section), certain equipment criteria must be considered. This checklist is extremely important since a small distributed system in one local or regional area of an organization may eventually be enlarged to a nationwide basis. Similarly, the initial system may take on new applications. The following criteria, then, to achieve such flexibility and expandability include equipment that is capable of:

- taking advantage of newer, lost-cost equipment as it can be economically justified
- improving throughput, i.e., the amount of data that can be processed within a specific time period.
- being used with current equipment and providing read/write media (magnetic disk, tape, etc.) that can be used by other devices in the system
- improving reliability for the user as well as greater accessibility for the user
- being configured as a single stand-alone piece of equipment or multi-terminal cluster or any combination
- performing local applications, working in a multipoint environment, or performing network control functions
- operating with no files, with small files, or with large disk configurations; offering a multiple of file access methods, from simple sequential to indexed to direct access
- facilitating the use of one industry-standard language or multiple languages
- being both fully programmable down to a single terminal level or not programmable at all
- supporting the intended applications via packaged software whether these packages operate on a dedicated basis or under the control of a multitask operating system
- fitting a single large application in dedicated mode or readily converting to handle multiple applications
- being acquired on various bases, i.e., rental, outright purchase, option to buy, and third-party leasing (lease-back arrangement)

Of the foregoing, perhaps the most important criterion is that the entire hardware family should be flexible to meet most any distributed processing system and be expandable in place. As low-cost devices become even lower in cost in the future, an organization should be able to take advantage of lower-cost system developments. Also, it should be noted that the expense to replace, change, and maintain a system will become greater than the value of the system hardware itself. Thus, a distributed processing system should be flexible and expandable to take advantage of newer, low-cost equipment and, at the same time, be sensitive to the benefits and costs of converting to this new equipment.

An integral part of the foregoing equipment criteria centers on the type of CRT display terminal. In addition to selecting the business type of terminal, there are the following other important considerations:

- Large-screen display features
- Operator aids
- Concurrent operations
- Automatic logging
- Security features
- Record insert/delete capabilities

Each of these will be discussed briefly.

When large-screen display features exist, a terminal can accommodate at least a full-page source document on a single-screen image. This implies a 1920-character display, formable into lines and columns, with the most typical specifications being 80 X 24. A smaller screen, which may be appropriate in a specific application, slows down data entry, adds to its complexity, and may cause operator confusion.

Every equipment manufacturer utilizes operator aids. Such features make it easier for an operator to perform the assigned tasks with precision and accuracy. They include screen prompting; programmable keys that enable an operator to call a routine, or execute a command, with a single keystroke; error display messages; upper-/lowercase and video highlighting features; cursor control that speeds up data insertion, deletion, and correction in fields or records; and mutiversion keyboards. These are all productivity aids that should be part of any large multiapplication system.

Concurrent operations enable operators to perform more than one task at the same time on an equipment configuration, such as a four-terminal local system that simultaneously executes data entry, file update, printing, and communications functions. The most often-stated advantage of concurrency is that equipment having the feature can perform either high- or low-volume work at the same time. A greater benefit, however, is that concurrency encourages and assures true distribution of equipment because it allows exceedingly cost-effective dispersal of linked equipment at a local, regional, or central site. Also, where a system has concurrency, it often causes serious degradation, with each task taking much longer to execute than it should. With a well-designed concurrent system, however, only extreme work-load queues should cause noticeable delays.

The next feature, namely, automatic logging, preserves data integrity at the point of initial entry, at the point of validation or audit, and at the point of recreation when necessary. It also results in other benefits. Productivity, for example, is knowing how much work an operator turns out and its quality compared to all other operators. Automatic logging counts and compiles statistics on operators, jobs, specific sites, multiapplication usage, and almost every other transaction handled by the equipment. Hence, logging benefits productivity and systems control and is highly recommended.

Security features are one of the industry's most sought-after capabilities. Although total systems security is not attainable at this time, a fully secure or semisecure installation can achieve a very high degree of data and systems integrity. Features that provide a practical form of security include password sign-ons, physical terminal locks, and software lockouts. But regardless of the form, good protection demands that the system prevent entry into individual record fields, individual records, single or multiple files, and individual terminals. Also, there should be no access to specific commands.

In any data entry system where the two most common mistakes are failure to enter a record and to duplicate the entry, the record insert/delete capability is a must. If a terminal had no provision to correct an oversight, the result would be a major problem. All sorts of totals would be wrong, computer time would be wasted, and tracking down the missing or duplicated entry would consume a lot of time. Some early key-to-disk systems did not have insert/delete of a record capability. In a distributed system, the implications are the same or even worse; data could be lost.

EQUIPMENT SELECTION—CONCLUDING PHASE OF FEASIBILITY STUDY

After the best distributed processing system has been selected from among the several feasible system alternatives, the appropriate equipment must be selected. Most organizations undertaking a distributed processing project have specific equipment under consideration. Because most of them have computer and related peripheral equipment salesmen calling on them at various times, they have had previous contact with most of the manufacturers. The representatives of the various equipment manufacturers should be contacted and invited to an orientation meeting on the proposed system. During the course of the meeting, they should be instructed about the applications to be covered, general problems that will be encountered, approximate volumes (present and future), and other pertinent data. Each manufacturer should indicate in writing whether it wishes to receive a bid invitation.

SUBMIT BID INVITATIONS TO MANUFACTURERS

Once letters of intent to bid are on file from equipment manufacturers, the company submits bid invitations to the interested equipment suppliers. The preferred approach when sending bid invitations is to mail the same set of data to all competing manufacturers. This permits bids to be placed on an unbiased basis and informs the manufacturers what requirements they must meet, keeps the number of questions to a minimum, and is a valid basis for comparison. Generally, the manufacturers will need additional information and assistance from the prospective customer as they progress with the preparation of their proposals.

Utilizing this approach, the respective manufacturers should have ample information to familiarize themselves with the company and its peculiar distributed processing problems. The recommendations made in their posposals should show clearly how the equipment will meet the customer's needs. Specifications lacking clear definition from the beginning will result in proposals with standard approaches that are applicable to any and all potential customers, making all the preliminary work a waste of time. It is of utmost importance that data submitted to manufacturers be as complete and self-explanatory as possible for the proposed distributed processing system.

How to Conduct a Distributed Processing Feasibility Study

Contents of Bid Invitation

Much of the material needed for the bid invitation can be taken directly from the data contained in the exploratory phase of selecting the best distributed processing system. The contents of the bid invitation include these areas:

1. General company information
2. Future distributed processing plans
3. List of new system requirements
4. New system flowcharts and decision tables
5. List of equipment specifications (in general terms)
6. Data to be forwarded by each manufacturer

In sections 1 and 2, the narrative should be brief so that attention can be focused on the remaining parts of the bid invitation. Data that are necessary for a thorough study are contained in sections 3, 4, and 5, forming the basis for the manufacturer's proposal. Section 3 is composed of five essential parts: planned inputs, methods and procedures for handling data, data files to be maintained (distributed data bases), output needs, and other requirements and considerations for the new system.

New system flowcharts and decision tables are contained in section 4. System flowcharts are needed for each function area under study and to show the interrelationships among the areas. Decision tables should be an integral part of the bid invitation. This will enable the manufacturer to have a complete understanding of the programming effort envisioned and help determine the hardware that is needed under the existing conditions. Finally, this section of the bid invitation should contain a flowchart that depicts the overall aspects of the new system. This allows the equipment manufacturer to obtain an overview of the system and its subsystems.

Section 5 contains a listing of equipment specifications in general terms. Competing manufacturers must have a basic understanding of the data communications network, I/O terminal devices, auxiliary storage devices, the central processing unit, and other peripherals. The inclusion of this section not only details present owned equipment that is compatible with a distributed processing environment but also helps assure greater compatibility of bids from each competing equipment manufacturer.

In this final section, data to be included in each manufacturer's proposal is listed. Specifying in advance what the proposals should contain ensures that comparable information for a final evaluation will be forthcoming.

Conferences with Manufacturers

Even though bid invitations specify the numerous details of the new system, legitimate questions will be raised by the various equipment firms. Many of the questions center around areas that may have need of modification, necessary to take advantage of the equipment's special features. The result may be favorable benefits to the firm in terms of cost savings. Conferences between the manufacturer and the potential customer, then, can prove beneficial to both parties. However, caution is necessary on the part of the study group during this period because salesmen may use this time to sell the firm and the final proposal, making the final evaluation of the manufacturers' proposals not objective but subjective.

Evaluate Manufacturers' Proposals

The distributed processing manufacturers should be given a reasonable amount of time to prepare their proposals. In most cases, approximately four to six weeks is adequate. When the proposals are completed, several copies are mailed to the customer for review and are then followed by an oral presentation by the manufacturer's representative(s). At this meeting, the salesman will stress the important points of the proposal and answer questions. After this procedure has been followed by all competing manufacturers, the DP manager and his staff should be prepared to evaluate the various proposals.

There are many criteria for evaluating a manufacturer's proposal. Among these are extent of automation proposed, evaluation of throughput performance—turnaround time, type of equipment, method of acquiring equipment, delivery of equipment, installation requirements, manufacturer's assistance, programming assistance, training schools, availability of reliable software, maintenance contracts, and other considerations. Finally, the proposals are evaluated in terms of how well they have complied with the bid invitation. Only after an intensive analysis of the facts can the DP group intelligently select the manufacturer(s) for distributed processing equipment.

Select Equipment Manufacturer(s)

Selection of the equipment manufacturer can be difficult task since computer manufacturers have different ways of viewing distributed processing. Currently, the main difference between IBM's approach to distributed processing and that of other major competitors is that IBM apparently believes in centralized host control over the network of

How to Conduct a Distributed Processing Feasibility Study

distributed systems, whereas Burroughs, Univac, CDC, and the minicomputer companies, such as Digital Equipment Corporation, believe in distributed control within networked systems or "netted" systems. On the other hand, Honeywell's current position appears close to IBM in this regard but may, in fact, provide an intelligent compromise between the two. Fundamentally, IBM clearly is against the "distribution of control" in that it would like to integrate the host computer, the front end, the network processors and multiplexers, concentrators, remote intelligent controllers, or satellite processors and terminals so each of these components and functions is ultimately dependent on the IBM central host(s) facility. This strategy frustrates the attempts of plug-compatible vendors to penetrate the IBM customer base and, at the same time, ensures greater total system or network control, improved data security and system integrity, and greater customer loyalty. In contrast, most other vendors cannot afford to be so restrictive due to their small market shares. Hence, this important factor should not be overlooked by the feasibility study committee.

The selection process is much easier if the equipment proposed is identical for all practical purposes. In such cases, the choice is normally based on the lowest-cost equipment. However, this approach is generally not followed because most manufacturers have certain unique equipment features, and this results in slightly different approaches to the customer's proposed system. To resolve this dilemma among the various competitors, several methods have been developed for evaluating and selecting equipment.

Method of Evaluation

One method of evaluation is utilization of a decision table, shown in Figure 10. A decision table for a final evaluation not only defines the important criteria in compact notation but also permits an objective evaluation because the values will have been determined before receipt of the manufacturer's proposals. In the illustration, the highest possible score is 100 points for each of the five distributed processing manufacturers. A value of 10 points is deducted for each no answer of a major criterion, while a value of 5 points if subtracted for each no answer of a minor criterion. The major criteria represent factors that have long-run effects on the firm in terms of profits and return on investment. Thus, the deduction of 10 points indicates greater importance attached to this particular criterion. Values for another firm might be different from those found in Figure 10. For the study, this is a realistic approach in making the final decision for the selected manufacturer, vendor 2.

Table Name:
Criteria to Select Distributed Processing Equipment Manufacturer

Decision Table | Chart No: FS-SM-1 | Prepared by: R. J. Thierauf | Page 1 of 1 | Date: 8/30/7-

	Rule Number											
Condition	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Major criteria:												
Flexible to meet present and future user needs	Y	Y	Y	Y	N							
Expandability of equipment at all operating levels	Y	Y	Y	N	Y							
Low-cost data entry	N	Y	Y	Y	N							
Monthly rental within budgeted amount	Y	Y	Y	N	Y							
Dependable and efficient software	N	Y	N	Y	Y							
Full service backup with proven record	Y	Y	Y	Y	Y							
Minor criteria:												
High degree of automation proposed	Y	Y	Y	N	Y							
Availability of equipment when needed	Y	Y	N	Y	Y							
Capable of meeting installation requirements	Y	Y	Y	Y	Y							
Adequate programming assistance available	N	N	N	Y	N							
	Rule Number											
Condition	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Good quality training offered	Y	Y	Y	Y	N							
Available equipment for compiling and testing program initially	N	Y	Y	Y	N							
Adequate personnel available	Y	N	Y	Y	Y							
Compliance with terms of bid invitation	Y	Y	Y	N	Y							
Action												
Subtract 10 points for each major criteria no (N) answer	X	–	X	X	X							
Subtract 5 points for each minor criteria no (N) answer	X	X	X	X	X							

Other Information:
Total points = 100 (6 major criteria X 10 pts + 8 minor criteria X 5 pts = 100)

Competitor's total points:
1,/70; 2, 90; 3, 80; 4, 70; 5, 65

Figure 10. Criteria to select equipment manufacturer in a distributed processing feasibility study

Signing of Equipment Contract(s)

The signing of the equipment contract by a top-level executive brings the study to a close. Generally, the feasibility study represents approximately one-third of the total time expended on a systems project. In the period just ahead—systems implementation—not only will more time be involved than in the feasibility study, but there will also be more involvement of organizational personnel. The problem of how to coordinate and control the activities during the coming period is a challenging task even for the most seasoned data processing manager and staff.

How to Conduct a Distributed Processing Feasibility Study

SUMMARY

The key to determining if distributed processing is feasible is to examine the DP operational aspects of an organization. Fundamentally, if the structure of the business is decentralized and a substantial amount of processing time is taken up with communications to headquarters, distributed processing can be a viable alternative. On the other hand, if the time spent for communication tasks is minimal now and in the future, distributed computing may be unnecessary. However, it may still be desirable to go distributed because of meeting specialized processing needs at the local and regional levels. Also, there may be better control over input/output functions at these levels.

The selection of the best distributed processing system is contingent upon performing a detailed feasibility study to determine the functions and tasks appropriate or desirable for distribution. Similarly, consideration must be given to the types of equipment and devices that are capable of performing these distributed functions and tasks. This information, in turn, forms the basis for forwarding bid invitations to equipment manufacturers. The receipt of the bids allows the DP manager and study group to select the appropriate equipment, thereby bringing the feasibility study of distributed processing to a formal close.

RELATED READINGS

Ainsworth, W., "The Primacy of the User," *Infosystems,* April 1977.

Bobick, S. A., Arman, E. J., and Yerkes, A. W., "Survey of Small Business Computers," *Datamation,* Oct. 1976.

Bowers, D. M., "Systems-On-A-Chip," *Mini-Micro Systems,* May 1976.

Brown, F., "Distributed Processing, There Ain't No Free Lunch," *Computer Decisions,* April 1977.

Burnett, G. J., and Nolan, R. L., "At Last, Major Roles for Minicomputers," *Harvard Business Review,* May-June 1975.

Business Week Special Report, "Minicomputers Challenge the Bit Machines," *Business Week,* April 26, 1976.

Business Week Special Report, "Glowing Prospects for Brainy Computer Terminals," *Business Week,* Oct. 25, 1976.

Colasanti, M., "Leasing and Distributed Processing," *Computer Decisions,* Nov. 1976.

Computerworld Special Report, "Distributed Processing," *Computerworld,* March 28, 1977.

Feidelman, L., "Distributed Computing: It's a Small World," *Infosystems,* April 1977.

Fortune Magazine, A Special Advertising Section on Distributed Processing/Data Communications," *Fortune,* March 1977.

Foss, W. G., "Guidelines for Computer Selection," *Journal of Systems Management,* March 1976.

Gilder, J. H., "All About Microcomputers," *Computer Decisions,* Dec. 1975.

Hannon, J., and Fried, L., "Should You Decentralize?," *Computer Decisions,* Feb. 1977.

Hansen, J. R., "Terminals Get Smarter So Do Users," *Infosystems,* March 1977.

___, "Minis Move to the User," *Infosystems,* June 1977.

Hebert, B., "Multiprogramming on Minis," *Computer Decisions,* June 1977.

Igersheim, R. H., "Managerial Response to an Information System," *AFIPS Conference Proceedings* (National Computer Conference), Vol. 45, 1976.

Jaffe, M. T., "There's No Business Like Small Business," *Computer Decisions,* Aug. 1976.

Johnson, J. R., "Mini or Mainframe? An On-Line Question," *Infosystems,* April 1977.

Jones, J. H., "Distributed Processing—Age of the Application Analyst," *Infosystems,* June 1977.

Keider, S. P., "Once Again—Centralize or Decentralize," *Infosystems,* Dec. 1976.

Kelly, N. D., "Cutting DP Costs," *Infosystems,* June 1977.

Knottek, N. E., "Selecting a Distributed Processing System," *Computer Decisions,* June 1976.

LaVoie, P., "Distributed Computing Systematically," *Computer Decisions,* March 1977.

Lecht, C. P., "The Waves of Change," *Computerworld,* June 6, 1977.

Luke, J. W., "Unraveling the Confusion of Distributed DP," *Infosystems,* Dec. 1976.

Lusa, J. M., "Selecting Terminals Intelligently," *Infosystems,* Sept. 1976.

___, "Distributed Computing, Alive and Well," *Infosystems,* Nov. 1976.

Monrad-Krohn, J., "The Micro vs. the Minicomputer," *Mini-Micro Systems,* Feb. 1977.

Moore, W. G., "Going Distributed," *Mini-Micro Systems,* March 1977.

Reynolds, C. H., "Issues in Centralization," *Datamation,* March 1977.

Stiefel, M. L., "What Is an 'Intelligent' Terminal?," *Mini-Micro Systems,* March 1977.

Thierauf, R. J., *Systems Analysis and Design of Real-Time Management Information Systems,* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1975.

Wagner, F. V., "Is Decentralization Inevitable?," *Datamation,* Nov. 1976.

White, R., Guzeman, D., Pezzolo, D., and Berry, T., "Impact of Microcomputers 1980," National Computer Conference Handout, 1977.

Yasaki, E. K., "The Mini: A Growing Alternative," *Datamation,* May 1976.□

Organizational Design for Distributed Processing

Problem:

The implementation of a distributed processing system has obvious ramifications in terms of hardware. The software ramifications are less obvious but we are beginning to appreciate the size of the task and can summarize it by saying "easier said than done".

Beyond hardware and software, distributed processing argues for changes in the data processing organization itself. The functions of systems analysis, programming, data control, data entry and operations, must of course, be performed, regardless of any formal organization.

In the distributed processing environment the traditional reporting lines are inappropriate and new ones must be devised. This report is a down-to-earth discussion of the traditional EDP organization and suggestions for its re-structuring to make it more effective within the distributed environment.

Solution:

The traditional data processing organization is centralized for some very valid reasons. In the earlier days of data processing, only about twenty years ago, computers operated in the batch mode. Then, too, the early computers were expensive and required special technical skills. It was not feasible to place an expensive computer and a well-trained technical staff in each division, functional area, or department.

Nor would it have been desirable, because coordination of systems would have been a problem. Systems development would have been difficult at best. This issue remains today, even with distributed processing.

Reprinted, by permission of the publisher, from *Distributed Processing Systems* by Judson Breslin and C. Bradley Tashenberg, pp. 103-121.

What is no longer an issue, however, is the need to return systems and data to the user.

The issue, from an organizational point of view, is one of centralizing versus decentralizing the data processing function. This issue can only be discussed with a full understanding of the makeup of the data processing function. For the data processing function is neither a single-purpose nor a single-minded organization. It is a three-headed entity that requires diverse skills and disciplines. The organization chart shown in Figure 1 illustrates the point. Our experience has taught us that certain basic elements should remain centralized, but that major processing improvements result from returning other skills to the user departments and organizations. We believe that the best of both worlds is attainable.

Organizational Design for Distributed Processing

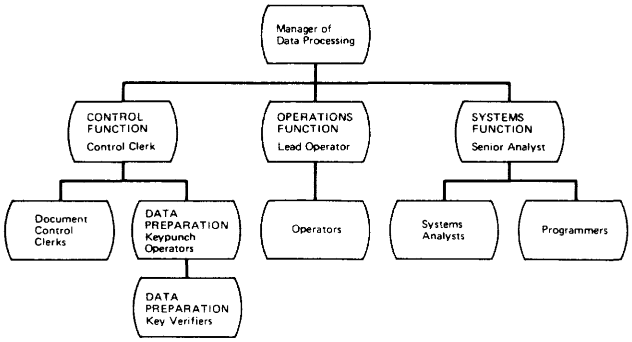


Figure 1. Organization chart for the department of data processing

MAJOR FUNCTIONS

Data Control

The centralized data processing organization must be scheduled. The average data processing center has many systems applications, which will be processed during the day, week, or month. These applications, which must be scheduled to run on the computer, include some of the following:

- The organizational payroll must be run semi-monthly or weekly. This requires collecting and preparing timecards and labor reporting information.

  At years end, W2 forms must be scheduled. Quarterly 941 forms must also be scheduled for the computer.

- Monthly accounting records and the general ledger must be scheduled. This means that the daily journal entries, or the government-oriented obligations and finding entries, must be prepaid for the computer.

- Manufacturing plans and inventory balances must be scheduled to provide manufacturing planners and purchasing agents with information on available inventory levels.

- Many miscellaneous applications, such as asset records, cash-flow reports, bank balances, and personnel reports, must be scheduled into the busy computer.

To coordinate this effort, the typical data processing centers normally have a data entry and data control function. Headed by a central control clerk, this function schedules the receipt of source documents from the individual departments or subsidiaries. The well-organized control function normally has a calendar of expected source documents with a special month-end closing schedule. Source documents are logged in and scheduled for preparation of data.

Data preparation is the process of preparing the source data and transposing it from the written word to some computer-readable format. The most common format in use today is the punched card invented by Mr. Hollerith before the turn of the century. Other methods that are becoming popular include keying the data directly to magnetic tape, called key-to-tape, and keying the data directly to disk storage, or key-to-disk. In any of these circumstances, a staff of keypunch operators or data entry clerks is required. Normally, these operators report to the central data processing manager through the data entry control function.

A related function is key verification. In this instance, the source data is re-keyed to verify the accuracy of the original keypunch effort. Generally, a red light flashes to indicate a discrepancy between the original keypunch card and the attempt at verification. By all rules of logic, verification is an expensive, time-consuming, and wasteful effort. But no one will deny that key verification is essential in a centralized data processing organization.

Operations

The second major function in a data processing function is operations, which differs drastically from the data entry function. Operations means "running" the computer. The people who perform this task are called operators. In a large organization, there may be many operators, and for those computer departments that have more than 8 hours of daily processing, there will be second- and third-shift operators.

The basic function of an operator is to run batch systems on a set, scheduled basis. To "run" a system, such as the weekly payroll, the computer must be turned on and loaded with the payroll programs, which tell the computer how to run the payroll. This program may be loaded from punched cards, magnetic tape, or the disk, the most conventional methods. Historical files must be mounted on tape drives, or disk drives, to be accessed by the computer programs. In this example, the permanent historical files would be the year-to-date payroll records filed by the employee. Now the system is almost ready.

The operator finally loads in the current data needed for preparing the payroll. In this case, the current data would be the hours worked during the week by employees. Normally, this is loaded through punched cards, or magnetic tape. The operator is responsible for putting the blank paychecks in the printer. Later

Organizational Design for Distributed Processing

he will be responsible for removing the checks and putting in the preprinted forms for the payroll journal.

After the payroll system has been run, all files that are now updated are stored for the next week's processing. Cards are "boxed" and stored, and the operator is ready to run the next scheduled system. In more sophisticated environments, the operator may "stream" multiple systems in a running sequence called a "job stream." The operations group is responsible for loading several systems with their files and new information while letting the computer call in each system, execute it, go on to the next system, and so on, until the job stream is complete. The operators in this environment control the processing of sequential systems through a job-control language technique referred to simply as the JCL.

The operations function has other functions that are important to the central data processing organization. One very important function is the control over all computer files. Extensive labeling of the tape and disk files is necessary in order to maintain the identity of the various systems files, including the general ledger and payroll. We will never forget one large chemicals company, which used a small minicomputer to control the tape files of its massive batch-system mainframe. Tapes were stored in vaults that encompassed room after room of tape files.

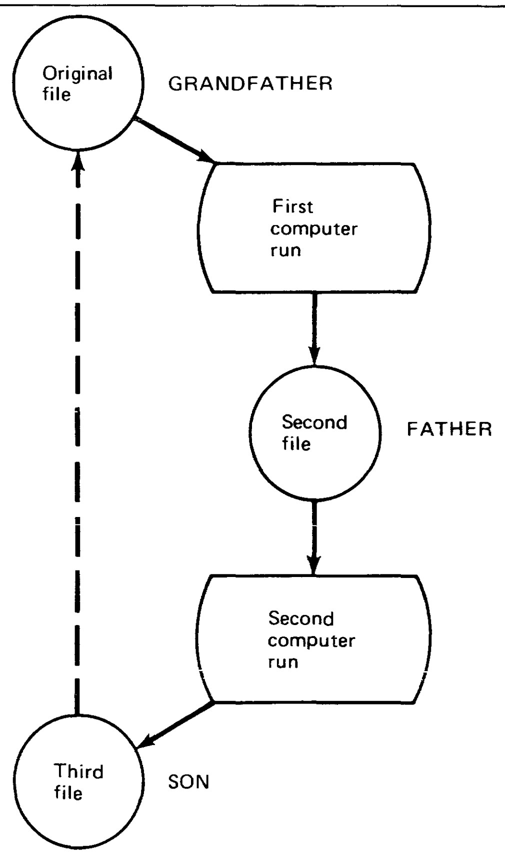


Figure 2. The grandfather concept

The operations function is normally responsible as well for "backup" files, kept as a protection measure should one of the current files be lost or destroyed. The most widely accepted method is the son—father—grandfather concept used in batch environments. The current file is updated by new data; the newly created file is called the son. The old file becomes the father. Somewhere in the file room is another, older file, which is the grandfather. It looks something like the scheme shown in Figure 2. This is a major responsibility and makes operations a vital function indeed.

Still other functions of operations include the scheduling and coordination of preventive maintenance by the manufacturers. Operations is responsible for the environment in which the computer functions, including the temperature and humidity. Finally, operations maintains a record or "log" of jobs run during the shift and the time required for processing. This log is later used to analyze scheduling and running times and can be a source of information for billing users for computer time used.

Systems

The last organizational component in this troika is the systems function. It is here that the impact of distributed processing will be felt most and will be contested hotly for years ahead. For it is here that the glamour and so-called professionalism of data processing rests.

The systems function is normally headed by a senior analyst or a manager of systems and programming. In larger organizations, the systems and the programming functions may be separate, which inevitably causes communications problems of some magnitude. Regardless, the team of systems analysts is made up of the professionals who design, program, and implement systems for the user departments. The systems analyst must be quite a remarkable person for he or she must understand the user's systems and design a system for the computer that will solve some obscure but ubiquitous problem, while at the same time improving the user's destiny. The systems analyst must then program that design for the computer and make it work. Then as we will discuss later, the same person must turn into a teacher and educate the user in this new system.

The demand for systems analysts in the job market remains strong. A talented systems analyst can make a major contribution to his or her organization, whereas a charlatan can raise havoc, and it is often difficult to distinguish between the two until it is too late. The problem is easy to understand. The systems analyst is a technician who understands the electronic computer, its capabilities and its limitations. A technician who can make the computer work through some technical programming languages called COBOL, FORTRAN, APL, PL-1, BASIC, or RPG, the systems analyst is not familiar with the user's needs and systems requirements. We place systems analysts in the untenable position of designing something to which they have had very limited exposure.

HOW IS PROFIT AFFECTED?

Theoreticians have argued for years for one or the other—that is, for a centralized data processing function versus a decentralized one. There are advantages and disadvantages to each. Furthermore, the real decision is heavily weighted by the particular circumstances of the individual organization. The real decision should be what works most effectively to improve the information processing and, in the profitmaking organization, what structure affects the bottom line most significantly. Too often, this is overlooked by managers who are not managing data processing. Let's look at some of the theoretical considerations in the issue of organizational structure, and then let us put them in perspective. Three issues in this decision are always the cost of the computer, the cost of software development, and responsiveness.

Hardware Costs

One computer is less expensive than two. Logically, the total costs of hardware seem to favor the centralized data processing structure. In fact, as we've pointed out previously, there is not much sense in placing multiple computers where one will handle the workload. We are proponents, remember, of "sharing."

The argument for centralization is also a historical one because of the high cost of computers and the technical support required. The current argument is for the "unit cost of computing." The proponents of centralized data processing contend that centralization results in a lower cost per transaction. This has led to the utilization of the computer 24 hours a day; it is the "fill it to capacity" syndrome.

One counterargument is vulnerability. If the entire corporation with all its divisions is dependent on one computer, the corporation would be seriously affected if its single computer were to become inoperable for an extended period of time. This, of course, depends on how heavily the corporation depends on its central computer.

Software Development Costs

No one should argue this point. A centralized or shared effort is essential, as it costs too much to develop similar systems for each of similar divisions in an organization. It is inefficient, too, and really not very smart. We remember a rather inept systems analyst in one multidivisional company who designed a fancy accounts payable system for a remote subsidiary (using a rather crude means to transmit the data), while the corporation processed its accounts payable transactions on unit record equipment. When the new enlightened management woke up, the corporate accounts payable system had to be completely redesigned, as the subsidiary's system had such limited capability.

In this case, just a little common sense, or a heavier dose of management, could have reduced the development effort by one-half as the larger corporate requirements would easily satisfy the remote subsidiary's requirements. Duplication of effort in systems development is expensive and must be dealt with if decentralization is to be effective. Can we share here too?

As a corollary to this argument, a centralized system allows for better control of the technical staff, which is critical in a data processing environment. As we pointed out earlier, good analysts are in demand, and this rather valuable resource is not infinite. Effective data processing and profit-line impact will depend on the best utilization of the systems function in the data processing environment. On the surface, this would clearly tend to favor a centralized function. Here analysts could work together on technical problems without recreating the wheel for each new system. In addition, more overall direction of the systems effort can be generated from a central corporate department, where goals and objectives can be translated into systems priorities.

Responsiveness

Few will argue this either. Responsiveness means time, and time is money. Responsiveness to user needs is most often enhanced by decentralization. Working at the user level, more effort will be generated directly to his needs rather than be dissipated through the multiple requests for multiple systems by multiple competing interests, these interests being either different departments, different functions, or different subsidiaries or divisions. The decentralized function also allows the user to deal more directly with the data processing entity rather than requiring that he sort through an organizational structure he does not understand, or deal with persons not totally familiar to him. This is especially true of the very large organizations and even the smaller ones that have a multidivisional structure.

Organizational Design for Distributed Processing

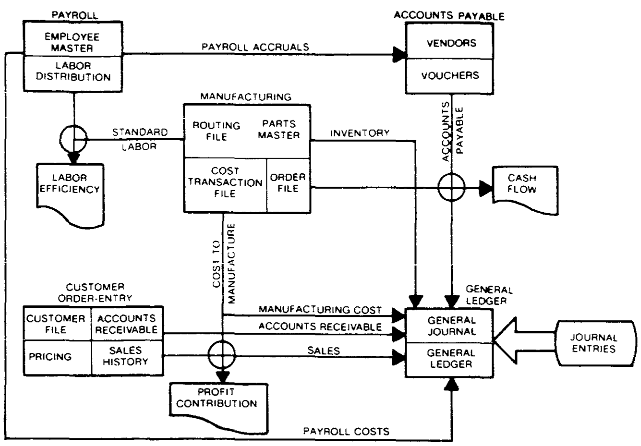


Figure 3. The data base

Responsiveness in a large central organization is almost always an issue. It is more than the time lag in submitting source data, keypunching it, verifying it, scheduling it, and then running it. It is more than the aggravation of error listings that must be corrected and resubmitted to the lengthy process. The argument is really that data processing technicians are trying to solve business problems which they simply do not understand, and time is lost. Responsiveness in a central data processing function is a deep-seated problem.

There are corollary issues here too. Frequently, animosity develops between the user departments and the central data processing department. In other cases, the cost of the service is a bone of contention because the central data processing function is expensive. For example, a manufacturing firm in Tennessee can process its requirements for one-half the cost quoted by the corporation's central data processing staff, that is, 6 cents per transaction, as opposed to 12 cents per transaction. The reason is the large main memory, disk capacity, tape drives, supplies, environmental equipment, and staff required to run the central data processing department. And it is the user who pays.

Animosity does not do much to improve the responsiveness of the central data processing entity. Setting users back in charge of their data processing destiny does improve responsiveness, however. The question here is how to do it. Then, if it can be accomplished, how will it affect our three cost issues—hardware, software, and responsiveness?

DISTRIBUTING THE FUNCTIONS

The answer may well be to distribute certain of the data processing functions, but not the entity per se. Data processing professionals maintain that you cannot set up separate departments throughout the organization, as it would be too costly and inefficient, and control over systems would be lost. Let us agree with them in theory that there should be control over the data processing function, so that everyone is not randomly doing his own thing to the detriment of the organization as a whole. The controlling factor should be top management. Let's call him the president or director of information systems, who reports directly to the president. Let's review the three functions of data processing ever so briefly.

- *Data entry and control:* Schedules and completes the data preparation by changing the written word to a computer-readable mode.
- *Operations:* Schedules and runs the various systems on the computer.
- *Systems:* Designs, programs, and implements user's systems on the computer.

Data Entry

In a distributed environment, data entry should be returned to the user. Source documents should be designed such that the user department, using on-line video display terminals, assumes full responsibility for putting all its information into the computer. With the responsibility for data inputting goes the responsibility for its timeliness and accuracy. This will mean taking the first important step toward on-line processing and an important step away from the old-generation batch mentality. It is a giant step.

The general ledger clerk will now enter her/his own journal entries and, as we discussed earlier, the computer will edit and verify the account number, the valid dollar limits, the date, and any other relevant information. The payroll clerk will now enter employee hours or labor distribution hours directly when the information becomes available. The manufacturing planner or the warehouse supervisor will now tell the computer that a shipment has arrived in a certain quantity against an established purchase order.

The middleman (the data processing entry clerk) is eliminated from the equation forever. This means savings in time and the total data processing budget. And this cost has not been transferred to the user, as you will see.

Operations

In a distributed environment, operations must remain with the computer, as in a centralized environment. Although the function changes dramatically, there is still a need for someone to maintain that piece of

Organizational Design for Distributed Processing

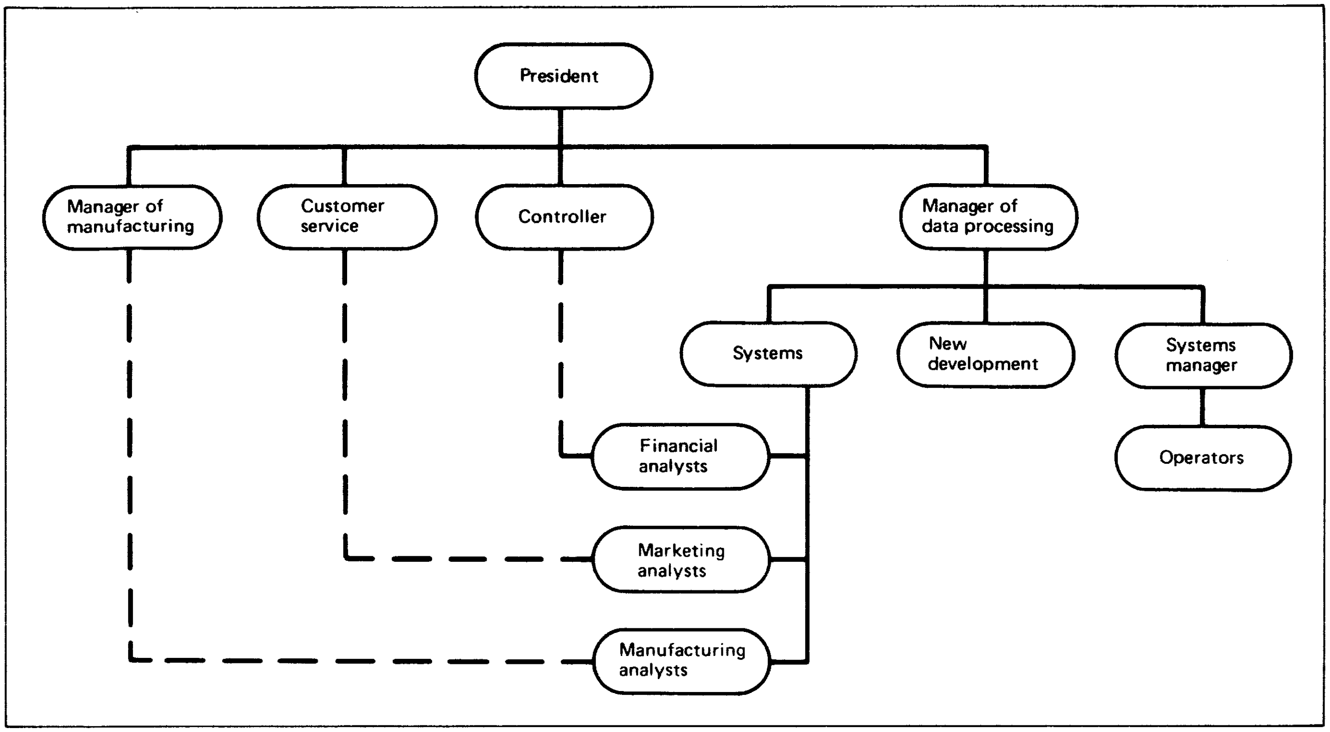


Figure 4. Organization chart: The dotted-line concept

hardware. In a central batch environment, the principle function of operations was scheduling and running different applications. In the distributed environments, files remain on-line, and new data is entered as it becomes available. Thus, the scheduling function is unnecessary with the exception of any batch-produced reports.

Even in an on-line system, some batch reporting will surely remain. For example, the printing of paychecks, W2 forms, and payroll registers is processed in the batch mode. Similarly, the general ledger printout occurs at one point in time. Logically, weekly or daily production schedules are printed at one time. The operations function can assume the responsibility for any reports that remain in the batch mode. Our experience dictates that this is not a major, but it is an essential, function. The biggest problem seems to be the special forms, such as invoices, paychecks, accounts payable checks, and bills of lading or shipping papers. In the latter instance, a small printer may be dedicated to printing these documents inexpensively, if the volume justifies it.

No, the major function of operations is not scheduling in a distributed environment. We should contend that the major function is to "police" the computer and the users. The new responsibility is to upgrade routine operations, and it justifies a new title: the systems manager. The systems manager stays with the computer.

In policing the computer, an important consideration in distributed processing is security. Payroll data is normally considered confidential. Not everyone should have access to other employees' salary. Other confidential information may be resident in the general ledger or in customer sales records. In addition to confidentiality, certain information should be available only to designated departments; for example, inventory balances should be available to planners and purchasing agents, but not to the payroll department.

One of the key responsibilities of the policeman, therefore, is maintaining security through *passwords*. In our new on-line environment, a user signs on in the morning by typing HELLO. The user then requests the system by name, such as, PAYROLL. The computer will ask for the password, which unlocks the payroll data base. If the user does not know the password, that ends the processing. If the user enters the correct password, processing may begin. Incidentally, this new computer can be programmed *not* to display the password on the video display terminal, so that no one can look over a user's shoulder and steal his or her password.

The systems manager assigns passwords, controls access to the system, and periodically changes passwords. The systems manager can also lock out users from the various systems. He can assign only one specific terminal to a specific function, such as payroll, and then closely guard the use of that terminal.

Organizational Design for Distributed Processing

The systems manager is also responsible for backing up the system. Each transaction in an on-line environment is "logged" on tape or disk. In the event of any malfunctions, it is the systems manager's responsibility to "bring the system up" or, simply stated, start it up again and get the system back in use. Normally, the logging tape is listed daily as an audit trail for all transactions. This, too, remains a centralized function associated with the computer itself. Maintaining a tape library and historical audit reports is the responsibility of the systems manager.

The final and undoubtedly the most important function of the systems manager is the maintenance of the organization's data base. We have discussed data bases before, but now it takes on organizational implications. Figure 3 illustrates a data base for a manufacturing concern. Actually, the data base consists of a set of functional data bases. The heart of the data base as a manufacturing concern is the manufacturing data base, which consists of the inventory master, product structure, work center, and routing files or data sets. Other functional data bases include payroll, accounts payable, sales, and order entry. Each of these functional data bases transfers information into the general ledger data base. All together, these functional data bases make up the organization's total data base.

The importance of the data base cannot be overstated. The systems manager is responsible for the integrity of the entire data base, its security, and the efficiency and speed of obtaining data from it. On the other hand, he is not responsible for its content; that responsibility rests with the user. The integrity of the data base means that information from the data base remains protected from destruction tampering, or unauthorized access. Remember that in an on-line environment, these files are always on the computer and are being accessed by many users simultaneously. The systems software, including the data base manager software we discussed earlier, is employed to keep the data bases synchronized. The systems manager must be knowledgeable in the technical aspects of the software and is therefore responsible for the integrity of the data base.

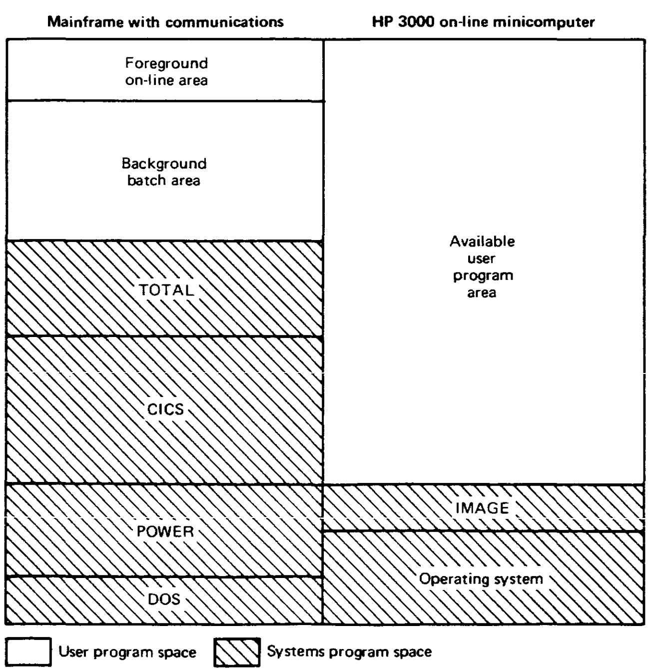


Figure 5. Memory utilization in a real environment: mainframe vs. on-line minicomputer

Because the data base is for the use of many end users, its maintenance rests logically with the hardware and not with the multiple users. Furthermore, it is a rather technical position that requires special training and a knowledge of the hardware rather than the user application. It should remain with the hardware. Finally, the systems manager "reorganizes" the data base periodically to better utilize the physical space on the disk files so that access time is minimized. This is a routine function for the systems manager. Maintenance and the physical environmental considerations remain with the hardware, even in a distributed system.

Systems and Programming

The systems and programming functions in a distributed environment should be returned to the user. It has simply not been effective as a centralized function. Analysts never really come to grips with user problems. However, some central control and coordination is needed. Otherwise, the software development costs will be prohibitive and inefficiencies in using the computer will surely follow. Again, this coordination and control must come from top management—it is simply too important.

There are at least two reasonable alternatives for achieving decentralization while assuring control and coordination of the systems effort. One alternative is used by one of the largest banking institutions in the world, Citicorp, which is the corporate entity of New York's Citibank (formerly First National City Bank). Citicorp is a strong advocate of minicomputers and distributed processing. This corporation uses a variety of minicomputers and distributes them to each major function and organizational entity, which justifies the requirement for computing capability. According to one vice-president, John Olson, the first step was to decentralize the systems function and assign systems analysts directly to the user departments.

This step greatly strengthens the inadequate systems support at the end-user level. The analyst is an employee of the user department and is therefore responsible to it for his performance and the effectiveness of systems he designs. As a major part of his training, the analyst learns the user's systems, his problems, and his opportunities for improvements using the computer.

After the indoctrination of the analyst and the decentralization of the function, Citicorp distributed minicomputers to the various functions and began phasing out the large central data processing department. A coordinating committee of top vice-presidents was set up to coordinate all the systems efforts and assure continuity within the overall objectives of the bank. But this committee does not operate at the detail systems level of the so-called systems planning committees, which produced such dismal results in the 1960s. Rather, it is a high-level steering committee set up to avoid duplication of effort and share new ideas to promote the most efficient processing of information for the corporation. This committee exemplifies the process of returning data processing to management, leaving the technical details to the analyst and putting both in the proper perspective.

An example of Citicorp's use of minicomputers is reported by Citibank, New York City's largest bank. The application is simple: an on-line system that allows clerical personnel to respond to customer requests for their account status. The system, which will replace a manual system of return phone calls and letters, will allow immediate access to customer accounts for more than 2,400 correspondent banks. The control is centralized, but each of the correspondent banks will access its own files through video display terminals. The system includes a minicomputer built by Microform Data Systems. The system will allow a customer service representative to retrieve a customer's account number and his account balance while the customer is waiting on the telephone.

It is a powerful and inexpensive use of a distributed process that epitomizes responsiveness. According to the latest information on the corporation, Citicorp will continue to decentralize, using minicomputers located at the distributed locations of users for still more applications.

A second alternative to distributing the system's function may be more acceptable. In this approach, which was implemented by a medium-size manufacturing firm, the systems analysts report to a manager of data processing in the corporate headquarters, but are assigned permanently to the end-user departments. This gives the analyst a straight-line reporting relationship to his manager and a dotted-line relationship to the end user. Figure 4 illustrates this approach. Here, the centralized data processing relates only to the hardware.

Organizationally, only the coordination of systems and the systems management function remains an entity. The systems analysts are assigned to the specific function of controlling their systems absolutely through video terminals. In a larger environment, each function may justify its own hardware, in which case the same relationship between central and distributed functions prevails.

In Figure 4 the analysts are located in the data processing department. Their key function, however, is with the user, as we shall see:

The financial analysts are responsible to the controller for all financial systems, including payroll, general ledger, accounts payable, and monthly profit-and-loss reporting. These analysts are familiar with the computer programs and the user's systems.

The marketing analysts are responsible to customer service and the marketing operation for the customer order/entry system. They are also responsible for the sales and marketing reporting and they, in fact, designed and programmed both.

The manufacturing analysts are responsible for the inventory and production control systems at this manufacturing concern. These analysts work with the manager of production and inventory control and have become familiar with the manufacturing function, as well as the sophisticated on-line system that supports the manufacturing requirements.

In a dynamic environment with ever-increasing systems requirements that fall outside the principal departments (marketing, finance, and manufacturing), a separate analysis may be established and termed "new developments." Under the direction of the manager who reports directly to the president, this group will handle special systems requests. In the case of this manufacturing concern, this group provided the key to implementing new systems in subsidiaries and making whatever minor customizations were required to fit in a diverse division.

THE COST-EFFECTIVENESS OF CENTRALIZED DATA PROCESSING

This organizational change will not come easily. It will be resisted. Distributed processing itself will not be welcomed by some because it represents a break with the traditional central data processing department. It raises the planning function to the level of top management and leaves the technical considerations to the more specialized data processing manager and his trimmer staff. How will such a change affect costs and responsiveness to user demands?

The Cost of Hardware

Distributed data and processing is an idea whose time has come because for the first time it is cost-effective with the new minicomputer, which also happens to be mini-priced. The logical argument that centralization reduces the overall cost of hardware is now questionable and even suspect. The large overhead factor in a centralized data processing setup makes this an expensive alternative. No generalities can be made here, for each organization has unique requirements and unique organizational structures. Each must evaluate its own potential for minicomputers and distributed processing.

But certain facts remain. A large centralized computer requires much more sophisticated software, such as operating systems, data base managers, and communications interfaces. It requires massive memory. Figure 5 is an example of memory utilization in a real environment. The two memory maps (as they are called) show that the two computer systems are configured with similar capacities of 192,000 to 196,000 bytes. But the large mainframe on the left has so much software that it has only 75,000 bytes for use by the applications program, as shown in the unshaded areas.

The newer minicomputer technology is more efficient and simpler in concept. In a prototypical system we evaluated, the usable core was 132,000 of the 196,000 total bytes. The minicomputer memory was dynamically allocated to each user on the basis of the amount of memory required to process each application. The mainframe, on the other hand, was configured to run one job at a time in the batch partition (54,000 bytes), regardless of how much memory was actually required.

The result is obvious inefficiency and the need for more and more memory. In addition, the mainframe and the typical central data processing organization normally translate all systems programs into COBOL, which is not a memory-efficient language. Again, more memory is required with more overhead, more cost. Although COBOL is available, the minicomputer is usually programmed in the more efficient FORTRAN, or the interactive BASIC languages. In the same experience reflected in the memory maps (Figure 5), the 192,000-byte mainframe could not support even one more on-line application. Because of its design, the minicomputer supported five major applications on-line. The cost of the minicomputer was roughly 25 percent of the mainframe.

There are other overhead considerations which, in many cases, negate the argument that centralization reduces the total cost of hardware. The central large-scale mainframe requires special raised floors for special electrical wiring as well as special temperatures with air conditioners and dehumidifiers. But most importantly, it requires a specially trained, expensive staff of technicians.

As a result, some companies are now finding that it is cost-effective to replace mainframes with several minicomputers, or at a minimum to shift major functions from the mainframe to a minicomputer. In each case, the minicomputer is distributed to the end user. One example is Lowe's Department Stores, a large chain of retail stores in the Southeastern United States. Lowe's replaced a central data processing department with multiple Eclipse minicomputers manufactured by Data General Corporation. We previously cited several manufacturing firms that have dedicated manufacturing and order/entry systems operating right at the user's location, as opposed to the central mainframe.

The Cost of Software

The second argument for centralized data processing is equally suspect today. In the first place, it applies only to multidivisional environments. For the single government or private organization, this is not an issue. But in the multidivisional company, if management blindly agrees to the design of dissimilar systems for multiple locations, it would indeed be costly. This danger exists with distributed processing, as responsibility is also distributed. In larger organizations, the position of corporate information systems takes the responsibility for preventing duplication of effort, beginning with a *corporate computer plan.*

The corporate computer plan defines the long- and short-range objectives for information processing in the corporation (or organization) as a whole. The plan defines the hardware requirements for the multiple divisions and sets up a schedule for implementation with responsibilities clearly defined. This plan should consider only what can benefit the organization and, in the profitmaking sector, the bottom-line effect of the computer. It should be an objective evaluation of the organization's needs, rather than its present situation. Most importantly, the plan should set up a schedule for software or systems development. These steps plus a close monitoring of the plan will preclude duplication of effort.

A good example of a corporate plan with control over software development costs is the Kodak Corporation in Rochester, New York. Here the corporation was originally responsible for designing and programming the major systems, which would be identical in its worldwide divisions. The systems were implemented under corporate control in each division. Software is identical and is written for one model computer. Each division is autonomous from a data processing point of view and has its own systems staff. Kodak maintains a staff of analysts in Rochester to monitor modifications submitted by the divisions, establish a communications link between divisions, and document all major systems. It works and reduces the total software development costs for Kodak.

In a single computer operation, where we recommend distributing data entry and systems expertise to the user departments, the cost of software does not differ in the distributed scheme. In fact, the on-line aspect of the minicomputer reduces the cost of systems development. As noted earlier, some maintain that interactive programming improves programming efficiency by up to 35 percent. We agree. Interactive programming operates much the same as interactive conversation between the user and the computer. The programmer makes changes to his program using the terminal. Once the change is made, it is compiled and tested directly from the terminal.

Compare this simplicity with the central mainframe environment: Program changes are made on coding sheets. The coding sheets are keypunched and run in a batch mode with other program changes. A listing, similar to an edit listing, advises the programmer of any errors he has made. He then corrects the errors and resubmits the change for another listing. When the programmer has a "clean" compilation, he can test his system changes. Thirty-five percent is probably a conservative estimate. So the idea that software costs are less in a large-scale central data processing organization is subject to analysis by the enlightened manager.

Responsiveness

In many instances, distributed processing is organizationally sound. Whether it uses shared resources or the network of remote computers, the Automated Distributed Data And Processing Technique (ADDAPT) is a concept in phase with the times. The arguments against it, such as hardware and software costs, must now be challenged. In smaller organizations, the idea of placing a terminal at every desk makes very good sense. Even in the largest of organizations, the potential benefit of breaking down the large central data processing organization is real. Certain functions can be processed more efficiently by the user. This has been proved in many Fortune 500 companies that took the initiative to decentralize.

The principal reason is responsiveness. Central processing, with its multiple demands, cannot respond as rapidly to user requests as the user can with his data processing system under his direct control. Remembering that three of the key elements in distributed processing are local data entry, local data editing, and data base access, management—including data processing management—must try to put these elements into effect, with responsiveness and profit improvement the potential benefits.

Here's another example: A small, concerned company in Pennsylvania took analysts from the user departments and trained them in the concepts of programming. Management wanted to decentralize its new data processing department. This was attained to the benefit of the organization as a whole.

The same can be attained with the newer, simpler computers, which put new tools in the hands of the users. Such tools as flexible data base managers and inquiry languages allow users to get to the information on the data base without needing assistance from a technical programmer.

The challenge is to create an environment in data processing that breaks down the language problems and the communications gap that exist between data processing professionals and the end users. The job content of the systems analysts needs to be redefined so that new emphasis is placed on data management and systems integration, with an end toward getting away from their role as one-time report generators, trouble-shooters, maintenance programmers, and interpreters of the end users' daily problems. The result will be more effective interpretation of valuable resources and increased responsiveness to user demands and needs.□

System Architecture for Distributed Data Management

Problem:

The term Distributed Processing has been in common use among data processing and data communications people since 1975. The term Data base has been in common use even longer and many EDP installations operate under the Data Base concept. The way these terms are casually tossed about one would think that "Distributed Data Management" or the management of a data base which is geographically dispersed but which is available in its entirely to multiple processors at these geographically dispersed sites, had been researched, solved and implemented in numerous software systems available today. Such is not the case, but many mainframe and minicomputer manufacturers are working on network architectures aimed at providing this capability.

This report focuses on some of the problems of Distributed Data Management and suggests approaches to solutions. This should be of primary interest to suppliers of EDP and communications equipment, but the user also can benefit from an awareness of these problems.

Solution:

INTRODUCTION

Successful implementation of most distributed processing systems hinges on solutions to the problems of data management, some of which arise directly from the nature of distributed architecture, while others carry over from centralized systems, acquiring new importance in their broadened environment. Numerous solutions have been proposed for the most important of these problems.

In a distributed computer system, multiple computers are logically and physically interconnected over "thin-wire" (low bandwidth) channels and cooperate under decentralized system-wide control to execute application programs. Examples of thin-wire systems are Arpanet, the packet-switched network of the U.S. Defense Communications Agency, and Mininet, a transaction-oriented research network being developed at the University of Waterloo. These may be contrasted with high-bandwidth or "thick-wire" multiprocessor architectures, such as the Honeywell 6080 or the Pluribus IMP. A practical consequence of thin-wire design is that processing control is in multiple centers. No one processor can coordinate the others; all must cooperate in harmony as a community of equals.

The key issue is that interprocess communication is at least an order of magnitude slower when the communicating tasks are in separate computers than it is when they are executing in the same machine.

From "Distributed Processing," 2nd Edition by Burt H. Liebowitz and John H. Carson. © 1978. Reprinted with permission, from *Computer*, January 1978. pp.

System Architecture for Distributed Data Management

Therefore, no single process can learn the global state of the entire system nor issue control commands quickly enough for efficient operation, so that multiple centers of control are implied.

This definition does not imply that the computers are geographically distributed. On the contrary, the machines might all be located in one room; they could be considered a loosely coupled multiprocessor rather than a network. The choice of machine location is a result of economic and political considerations; that is, the distributed system architecture is amenable to *both* single-site and widespread machine location.

The debate between centralization and decentralization has thus taken something of a new twist. The old arguments for centralized systems were based on economics of scale in computer architecture, simplification of computer center operation, and the need for an integrated data base. The arguments for economies of scale are highly questionable. Arguments for simplified computer system management are countered by the desire for corporate management to distribute computing power, together with the fact that distributed systems can be operated in one room, and are therefore as easily manageable as centralized systems. The elegance of the distributed system architecture is that these arguments are independent of the other issues of system design. If you plan for thin-wire communications from the outset, the machines can be located wherever considerations of economy, availability, and management combine to dictate.

However, the advocates of centralized systems still hold one trump card: the need for an integrated data base. This need is stronger now than ever. Nevertheless, an integrated data base doesn't necessarily require a single machine or a thick-wire multiprocessor architecture. On the other hand, imposing an integrated data base on a distributed processor architecture presents several problems.

ARCHITECTURE APPROACHES

Distributed data management systems possess a kind of architecture of whole computer systems. Various approaches have been taken to the study of distributed data management architecture;[1] we prefer a simple classification, Figure 1, that divides systems into integrated and federated classes. In the integrated architecture, each component data base management software package "knows about" the existence of all the others. This approach is possible only if the entire system is designed and built from the ground up.

If, on the other hand, existing data base systems must be used, the result is one of two forms of a "federated" architecture. In the simplest case, each of the software data base managers supports the same model and command language. Such a system is said to be homogeneous. However, an integrator must be superimposed, containing the data and procedures required to make the collection of data base managers behave as a single entity. For example, a request or a data item that is not found locally may be routed to one or more remote data base managers by data search integration routines. Problems of integrator design are discussed later.

The third class of architectures is a heterogeneous federation, in which totally disparate data base systems are linked. This type of architecture is clearly the most difficult to support. Both integrators and translators are necessary in heterogeneous federations. When the data base managers support different data models—for example, owner-coupled sets at one site and relations at another—both commands and data must be translated in order to pass between them. If application programs are to use the local data model, then the figure is correct as shown. If they are to use a common data model, then the translator resides between the integrator and the data base manager.

An integrated architecture does not obviate the need for, nor desirability of, specialization at the system hosts. Specialization of data base content and processing is certain to occur. As more complex data models individually evolve from the common base, the need for translation is reintroduced. However, the common base remains as the vehicle for this translation; it is, therefore, much simpler than the interface problem in a heterogeneous federation.

Naturally, life is not quite so simple. These three system architectures represent points on a spectrum rather than the only possible alternatives. Within this spectrum, the problems of distributed data management are of two types: fundamental problems, common to all of the architectures; and incidental problems, a result of trying to integrate systems that were built to stand alone. The latter occur in all but the most purely integrated architectures.

FUNDAMENTAL PROBLEMS

Any data management architecture must address five principal issues: how to provide an integrated data base, where to store data in the system, how to locate data, how to control concurrent access, and how to provide security and integrity. The distributed environment introduces new complications to each of these problems. They are intimately entwined and vary in complexity with network size, data base size, availability requirements, and response time needs. Some design constraints, such as high avail-

System Architecture for Distributed Data Management

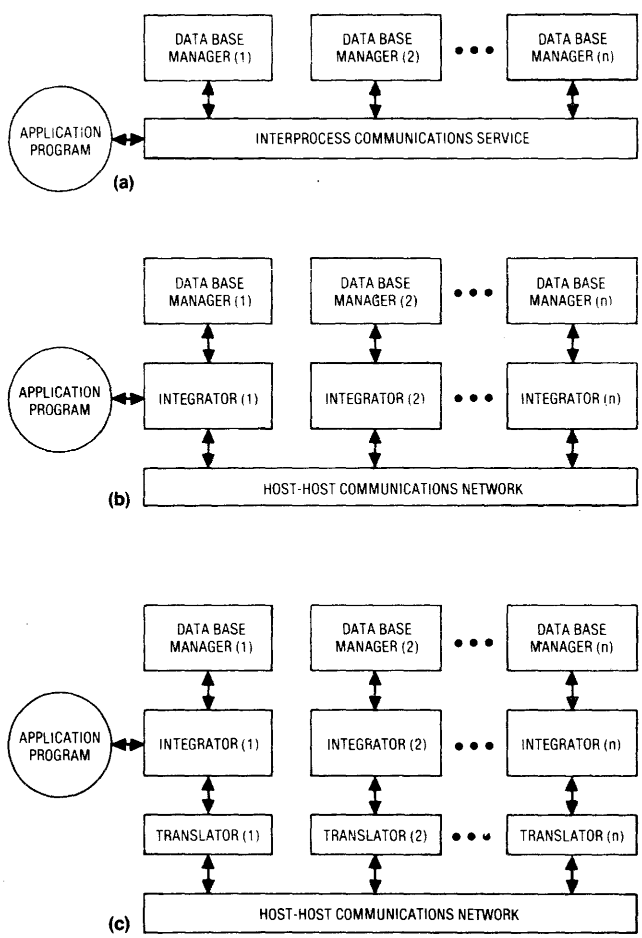


Figure 1. Architectures for distributed data management systems classified as (a) integrated, or loosely coupled multiprocessors, (b) homogeneous federation, or (c) heterogeneous federation

ability, may be easier to meet in a distributed system; while others, such as response time, may become harder to satisfy. In this short report we can give only a brief description of the nature of these problems and point to some of the work that has been done toward solving them.

If the distributed system is to be more than a communication channel plus a set of independent systems, then the data base must span more than one processor site. The purpose of the integrator components of Figures 1(b) and 1(c) is to achieve this span. All of the remaining problems discussed in this section must be solved in order to build such a distributed data base, but the first requirements is that the notion of distributed data be expressed in some way.

Data distribution is desirable at both the information structure level and the storage structure level of data modeling. Distribution at the information structure level means that data location is an explicit part of the data model seen by application programs. Suppose, for example, a customer file is partitioned by city of residence. Then an application program can ask for data from the customer file for a particular city—say, Omaha. Distribution at the storage structure level implies the opposite; the location of data is hidden from application programs. Rather, data is allocated to processor sites by storage management algorithms. This requires that mechanisms be built into (or on top of) the data base managers to locate data objects and to allow for their migration.

Information structure distribution should not force storage structure distribution. That is, the Omaha customer file might be moved, temporarily or permanently, to a processor located in Hartford, without affecting the way application programs use the file.

WHERE TO STORE DATA

In a distributed system there is a clear benefit in storing data at the processor where it is most frequently used, since thin-wire communication implies that access to remote data is much slower than access to local data. If the frequency of access is known from all processors for each file or other unit of retrieval, then an optimal allocation of data objects to processors can be defined as one that minimizes the average access time. But this model is rather naive; several additional factors should be considered as well. For example, you should distinguish between update and retrieval access, since the former requires at least one extra transfer over the network to store the new data. Storage and transmission costs can be considered; and you may be free to adjust communication channel capacity as well as storage location. Storing more than one copy of a data object tends to reduce read-only access times, but also tends to increase the time required for update, since all the copies must be updated consistently. Access patterns may change over time, so that today's optimal allocation may not be optimal tomorrow.

The data allocation problem has been shown to be "polynomial complete"[2]—i.e., execution times of the best known algorithms grow exponentially with the size of the problem. Today's practical applications involve more data than can be optimally allocated by any known algorithm; therefore heuristic techniques are necessary. These are techniques that are known to work well, but have not been proved mathematically to find the optimum solution.

A number of researchers listed in the related readings have applied mathematical programming optimization models. Certain recent contributions represent the most realistic sets of assumptions. One of these introduces the notion of imperfect knowledge of access statistics, noting that these statistics may change with time.[3] These researchers divide time into periods during which they take the access patterns to be static, and find an optimal static assignment of data to processors. If the access patterns for the next period are known, another assignment can be computed. But if these assignments are not identical, then the method incurs the cost of moving the data into the new configuration. Thus their aggregate model minimizes both access and reconfiguration costs.

Another study developed a very comprehensive model that distinguishes query and update traffic.[4] Updates require more message transmission between the processing host and the storing host than queries do. The problem is how to allocate copies of data objects to processors and also to allocate communication bandwidths, to minimize the combined storage and transmission costs. The technique imposes the constraint that average access time is to be bounded by a parameter supplied by the designer. Then the average message delay is expressed in terms of the network traffic (using Kleinrock's formula), which in turn depends upon the data objects and capacity assignments. A further constraint is that availability of object j must exceed a parameter A(j). Again heuristic methods must be applied if problems of even moderate size are to be tackled. The investigators programmed an absolute optimizer and a heuristic optimizer. For five processors and 20 files the absolute optimizer was still computing after 20 minutes. The heuristic program found a better solution in 5 minutes.

Clearly, more research is needed in this area. A problem with five processors and 20 files is hardly a very large one in the commercial world, yet this study showed that it sorely taxes current solution techniques that seek an absolute optimum. But researchers must be warned by the polynomial complete character of the data allocation problem; they should seek improved heuristics, because it seems unlikely that absolute optimum-seeking algorithms will be successful.

Locating Data

When data is needed, how is it found in the distributed system? Presumably a request for a data object can originate at any processor; the data base manager at that processor must have some algorithm for finding where it is stored. Several alternative strategies are available for locating a data object, assuming that it has an identifier unique in the network. The choice of strategy will depend on the number of objects from which one is to be found, the fraction of requested objects that are stored at the requesting site, and the dynamics of updating.

The simplest procedure is to encode location information in the identifier. But then moving an object becomes extremely onerous, because every reference to the object must be found and updated, and one can never be sure all references have been corrected. So this solution is almost certainly not acceptable. Another simple approach is to first look for the object locally—at the site where the request is made—and, if it is not found, to broadcast a request for it to all other sites. This procedure creates an enormous quantity of communication traffic; in a network of N processors, it invokes N—2 data base managers that need not have been involved. Only if broadcasts are very rare will this be tolerable.

A third alternative is to provide each data base manager with a full index that identifies the location of every data object in the network. However, each manager must keep its index somewhere locally; thus the problem is one of storage space, especially for times when the units of retrieval are small and numerous—short files or records within files. Furthermore, every time an object is added, deleted, or removed, every index must be updated.

There are intermediate solutions. For example, full indexes could be kept at only a few service locations. Or if broadcast search is to be used, it may be possible to define "neighbors" for each processor, X, as those processors most likely to hold data requested at X.

An analysis of this problem,[5] comparing costs and access delays for a number of alternative index structures, shows that for update rates of less than 10% of the total access rate, distributed full indexes are better than a centralized full index. On the other hand, the choice reverses for higher update rates. The local index with broadcast search procedure is almost never preferred.

CONCURRENCY CONTROL

To maximize the concurrent use of system resources by multiple users, shared access is allowed to the system resources. A data object is "locked" by a user only when he must be assured that it is not in some transient state. As soon as locking of resources is permitted the possibility of deadlock arises, when two or more users each are trying to reach an object locked by the other. Either prevention or detection and recovery is necessary.

System Architecture for Distributed Data Management

Deadlock control in a single system is well understood,[6] but in a distributed data management system the locking problem is made more difficult by the existence of multiple centers of control. Since in a thin-wire system no one processor is allowed to control all the others, concurrency must be controlled through cooperating algorithms executed at each host. The problem is to ensure that the distributed data base remains consistent despite attempts at concurrent update from different processors. Objects at more than one host, such as duplicate data or structural information, must be concurrently lockable.

There have been many proposed algorithms for maintaining the consistency of multiple copies of data. One algorithm assumes that accesses to an object are either queries of the contents or updates that append new data or overwrite old data.[7] A "time stamp" is recorded for both queries and updates, and the most recent value is used as a stored item's true value. The algorithm achieves "eventual consistency," although at any given time there is no guarantee of consistency. More recently, this algorithm has been expanded to work with general updates, which include modification of old data as well as augmentation and replacement.[8] Two other proposals impose a linear order on resources that are to be locked, and permit lockings only in the sequence defined by this ordering.[9, 10] This must be done if deadlock prevention is desired and a lock cannot be preempted.

Another strategy for deadlock prevention with multiple copy objects is to perform multi-step locking. This strategy adds an in-preparation state between the two extremes of locked and unlocked, or available.[11] Locking all copies of the object requires the requesting processor to first issue a Prepare request, and the data base managers to acknowledge the request. Only when the Prepare acknowledgments have been received can the requester issue a lock command to each. The sequence of events that can occur is complex.

Another investigation compares centralized and distributed detection schemes.[12] In a centralized scheme, one processor has the responsibility of monitoring potential lockups, departing from true system-wide control, as defined at the beginning of this report. In a distributed procedure, each data base manager periodically broadcasts the local current status to each of the other managers. From this information, all the managers then construct an approximation of the global picture. Both of these schemes involve substantial overhead and are probably not acceptable if the locking granularity—the size of the unit of retrieval—is small.

Further research on both detection and prevention schemes is needed. In particular, far too little is known about the probability of interference and deadlock in concurrent data base accesses. For transaction processing systems there is strong reason to believe that interference is rare, and that elaborate avoidance algorithms would not be economical. Research on this issue is under way.

SECURITY AND INTEGRITY

Data base integrity can be compromised by inadequate concurrency control, erroneous software, security breaches, or by system failure ("crashes"). Concurrency has already been discussed; the correctness of software is beyond the scope of this report. The security problem can be rendered more or less difficult in a distributed system than in a centralized one, depending upon details of the system architecture and structure of the application. Problems that arise include securing data during transmission and assuring that any site to which data is sent enforces the same security policy as the site that normally holds it. The literature is almost totally devoid of solutions to problems of the latter type.

In a distributed system that is subject to component failure (a likely proposition), correct operation is hard to guarantee. Suppose, for example, that a transaction updates three records, each stored at a separate host. None of the three updates is in effect until all have been completed and acknowledged. Suppose the source of the transaction, having received all three acknowledgments, sends out the message, "OK to put into effect," to the three data base managers, but one of those three goes down before it has put its update into effect. This is closely analogous to the problems of data transmission protocols, where correctness in a certain sense cannot be absolutely guaranteed.[13]

A recent investigation attempts to resolve the problem by writing to "stable storage" an "intentions list" of actions necessary to complete the updating parts of a transaction.[14] The stable storage ensures that either write operations are executed completely or make no change to the stored data; this is true even if a crash occurs in the middle of a write (the investigators suggest an architecture). An intentions list must have an "idempotency property," which guarantees that repeated execution of any subsequence of the intentions list has precisely the same effect as executing it once, provided that those sequences cover the whole list. Thus, if operations are restarted after a crash, it is acceptable to repeat some previously executed writes. Simple write operations have this property. With these tools, single- and multi-machine algorithms can be defined that guarantee the eventual execution of any properly defined transaction. (Users must define the start and end of a transaction. If the end is not signaled before a time-out occurs, the transaction is aborted.)

The fundamental problems of distributed data management have not yet been given adequate attention, but a great deal of research is in progress. Security and integrity control are vitally important if distributed systems are to operate correctly. However, most of the work cited in each of the problem areas we have discussed consists of designs on paper only; implementation and efficiency studies are needed.

INCIDENTAL PROBLEMS

Unfortunately, none of the generalized data base management packages that are currently available meets the requirements for distributed data management. Therefore, either vendors or users will have to devise "cut and paste" techniques to support distributed systems. For example, where does one express data distribution in a model that adheres to the Codasyl Data Base Task Group (DBTG) recommendation? Few users are prepared to modify their model software to handle such definitions. This means that system integrators must be built on top of the current software until such time as the vendors provide modified software.

Writing such integrators is not trivial. If, for example, a customer file is to be spread over several processors, the procedures that correlate customer identifications to processors must be written by the user. Furthermore, this is only a partial solution. If a DBTG set linking all customers with overdue accounts were desired, then the applications programs would have to recognize that this set would be implemented as several sets, one per host. This violates the desirable property of keeping the network structure invisible to applications.

Additional problems arise because the current operating systems on which data management facilities are built do not support efficient interprocess communication, which must be available for data location algorithms, concurrency control procedures, and crash recovery.

The construction of federated systems is not impossible, but it is likely to be very difficult. This is especially so in the heterogeneous case where data must somehow be translated in format when moved between hosts. This problem is unsolved in general, although a committee of the Codasyl Stored Data Definition and Translation Task Group has recently published a detailed proposal that addresses this issue.[15]

EXAMPLES

Few distributed data management systems are operational. Those operated by the airlines and by banks are impressive, but they are "home-built" and do not represent generally available packages. Some manufacturers, such as Digital Equipment Corp. and Prime Computer, Inc., modestly support distributed data management, but it exists chiefly in the form of support for interprocess communication and non-transparent data distribution for homogeneous hosts. We know of no commercial systems that support all of the requirements outlined in the previous sections on fundamental problems. However, many manufacturers are actively studying the problem.

At the University of Waterloo (Ontario, Canada) we have been working on the logical design of distributed systems. Our objective is to do transaction processing on distributed data bases through an integrated architecture. Since transaction processing has modest computational requirements, the design is based on minicomputers. We view the data base as a single logical entity, but we assume that accesses to it exhibit geographic locality of reference. This means that the data base can be partitioned into components so that the majority of hits on a given component come from terminals in one geographic region; the network has a processor for each partition. But locality of reference is a statistical property, and infrequent access to remote components are required. Furthermore, management is assumed to have occasional need to process the data as a whole. Therefore, communication channels link the computers, providing the basis for a distributed system.

This system, called Mininet, has been described in detail elsewhere.[16] Its design is summarized in the following paragraphs, as an illustration of an integrated architecture.

The support of transaction processing leads naturally to the design of a message-switched operating system. A transaction is handled by a number of small tasks. Each task performs a portion of the computation and invokes the cooperation of other tasks by sending them messages. An example is shown in Figure 2, where the circles represent the tasks that participate in processing a sales transaction, and the lines represent messages sent between tasks. Messages are passed by the communication nucleus of the system. This consists of a dedicated task, together with the communication subnet and its interface. The task is called the switch; a copy of it is executed at each host. An application task sends a message by invoking the switch with a Send command that identifies the receiver.

System Architecture for Distributed Data Management

The Send command has precisely the same format, whether the receiving task is in the same host or in a remote host. The data management function is also implemented with this task structure. This is the basis for transparency of the network structure; the distribution of tasks among hosts of the system can be made invisible to the application code. All knowledge of network structure is concentrated in descriptive tables that are interpreted by the communication nucleus. (Given the transparency of task location, it would therefore be a simple matter to implement "back end" data management machines for Mininet.)

Transparency of data distribution at the storage structure level is provided by cooperating groups of data management tasks. Each member of a group resides in a separate host. A request for data sent to any member of a group may involve all members. They cooperate to find the requested data file, using any of the data search strategies outlined previously in the section on locating data.

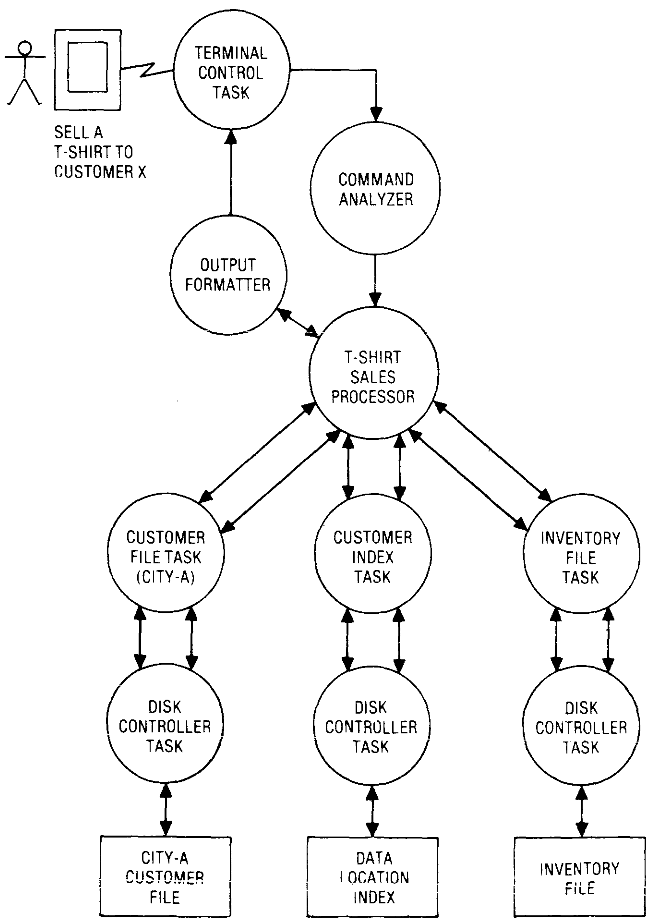


Figure 2. Transactions are processed by tasks (indicated by circles). Lines between circles indicate communication paths. Rectangles at bottom are secondary storage units; double rectangle at top is the point-of-sale terminal

This design has three key features that form the basis of constructing an integrated system on a distributed architecture. Dividing the computations into multiple small tasks distributes the workload over several processors. Transparency of task location in the interprocess communication mechanism allows reconfiguration of hardware and mobility of tasks over the system hosts. And finally, the use of data access task groups provides data location transparency.

The Mininet system is being implemented on PDP-11/45 processors. The communication nucleus is operational, and the data management system is being constructed.

CONCLUSION

Distributed processing offers exciting possibilities. It promises improved performance, improved system availability, easier adjustment to a growing workload, and some individualized control over machine resources. But wide application awaits the implementation of distributed data management services and solutions to several difficult problems. Most notably, crash recovery has received insufficient attention. Even when these problems are solved, the problem of migrating from current systems to the new architecture must still be addressed. Nevertheless, the dramatic drop in the cost of computers makes distributed systems inevitable. People will purchase small machines to solve small local problems, and then recognize the advantages of connecting small machines to their current big systems. Data processing management would be wise to plan for growth in this direction by recognizing the requirements of distributed data management, and by demanding integrated systems from vendors.

REFERENCES

[1]M.E. Deppe and J.P. Fry, "Distributed Databases; A Summary of Research," *Computer Networks,* Vol. 1, No. 2, September 1976, pp. 130-138.
[2]K. Eswaran, "Placement of Records in a File and File Allocation in a Computer Network," *Proc. IFIP Congress,* 1974, pp. 304-307.
[3]K.D. Levin and H.L. Morgan, "A Dynamic Model for Distributed Data Bases," *Proc. ORSA/TIMS Conference,* Spring 1975.
[4]S.A. Mahmoud and J.S. Riordan, "Optimal Allocation of Resources in Distributed Information Networks," *ACM Transactions on Database Systems,* Vol. 1, No. 1, March 1976, pp. 66-78.
[5]W.W. Chu, "Performance of File Directory Systems for Data Bases in Star and Distributed Networks," *AFIPS Conf. Proc., Vol. 45, 1976 NCC,* pp. 577-587.
[6]P. Brinch-Hansen, *Operating System Principles,* Prentice-Hall, 1973.

System Architecture for Distributed Data Management

[7]P.R. Johnson and R.H. Thomas, "The Maintenance of Duplicate Databases," ARPA Network Working Group, Request for Comments #677, Network Information Center, Document #31507, January 1975.
[8]R.H. Thomas, "A Solution to the Update Problem for Multiple Copy Databases Which Uses Distributed Control," Bolt Beranek and Newman Inc., Report #3340, July 1976.
[9]J. Gray, "Locking in a Decentralized Computer System," IBM Corporation, Technical Report TR-RJ1346, February 1974.
[10]J. Day, "A Principle of Resilient Sharing of Distributed Resources," Transcript of the Distributed Processing Workshop, *Computer Architecture News* (ACM), Vol. 5, No. 6, February 1977, p. 12.
[11]A.P. Mullery, "The Distributed Control of Multiple Copies of Data," IBM Corporation, Technical Report TR-RC5781, December 1975.
[12]S.A. Mahmoud, "Resource Allocation and File Access Control in Distributed Information Networks," PhD dissertation, Carleton University, 1975.
[13]C. Sunshine, "Issues in Communication Protocol Design—Formal Correctness," Reprints, ACM Interprocess Communication Workshop, March 1975, p. 118.
[14]H. Sturgiss and B. Lampson, "Crash Recovery in a Distributed Data Storage System," Transcript, Distributed Processing Workshop, *Computer Architecture News* (ACM), Vol. 5, No. 6, February 1977, p. 10.
[15]Committee on Data Systems Languages, Stored Data Definition and Translation Task Group, "Stored-Data Description and Data Translation: A Model and Language," *Information Systems*, Vol. 2, No. 3, 1977.
[16]Jacques Labetoulle, Eric Manning, and Richard Peebles, "A Homogeneous Network for Data Sharing," (two-part article), *Computer Networks*, Vol. 1, Issue 4, 1977.

RELATED READINGS

Canaday, R.H., et al., "A Back-End Computer for Data Base Management," *CACM*, Vol. 17, No. 10, pp. 575-582.

Casey, R.G., "Allocation of Copies of a File in an Information Network," *AFIPS Conf. Proc., Vol. 40, 1972 SJCC*, pp. 617-626.

Chandy, K.M., and J.E. Hewes, "File Allocation in Distributed Systems," *Proc. Conf. on Modeling and Measurement*, Harvard, Massachusetts, March 1976, p. 10.

Chen, P.P.S., "Optimal File Allocation in Multi-Level Storage Systems," *AFIPS Conf. Proc., Vol. 42, 1973 NCC*, pp. 277-282.

Chu, W.W., "Optimal File Allocation in a Computer Network," *IEEE Trans. on Computers*, Vol. C-18, No. 10, October 1969, pp. 885-889.

Chu, W.W., and G. Ohlmacher, "Avoiding Deadlock in Distributed Data Bases," *Proc. ACM National Conf.*, November 1974, pp. 156-160.

Chupin, J.C., "Control Concepts of a Logical Network Machine for Data Banks," *Proc., IFIP Congress*, 1974, pp. 291-295.

Holler, E., "Files in Computer Networks," *Proc., First European Conference on Computer Systems*, May 1973, pp. 381-396.

Karp, R.M., "Reliability Among Combinational Problems," *Complexity of Computations*, Plenum Press, 1972.

Kleinrock, L., *Stochastic Message Flow and Delay*, Dover, 1972.

Levin, K.D., "Organizing Distributed Databases in Computer Networks," PhD dissertation, University of Pennsylvania, 1974.

Metcalf, R., "Strategies for Operating Systems in Computer Networks," *Proc. ACM*, 1972, pp. 278-281.

Morrissey, John, "Distributed Processing Systems—Some Economic and Technical Issues," Transcript, Distributed Processing Workshop, *Computer Architecture News* (ACM), Vol. 5, No. 6, February 1977, p. 13.

Peebles, Richard, and Eric Manning," A Computer Architecture for Large Distributed Data Bases," *Proc., First Conference on Very Large Data Bases*, Framingham, Massachusetts, September 1975.

Roberts, L.G., and B.D. Wessler, "Computer Network Development to Achieve Resource Sharing," *AFIPS Conf. Proc.*, Vol. 35, 1970 (Spring Joint Computer Conference), pp. 543-549.

Whitney, V.K.M., "A Study of Optimal File Assignment and Communication Network Configuration," PhD dissertation, University of Michigan, 1970.

Additional reference material appears in the *Proceedings of the Second Berkeley Workshop on Distributed Data Management and Computer Networks*, May 25-27, 1977, available from National Technical Information Service.□

A Language for Distributed Processing

Problem:

This report is directed to advanced systems programmers who are familiar with, and use, Concurrent Pascal as a programming language in the design of large scale systems. The author proposes two basic modifications to Concurrent Pascal which will extend the applicability of the language, and also alleviate some of the problems which are encountered in its current use. The resulting modified language is referred to by the author as Distributed Pascal.

Solution:

The main question being addressed here is, what is a good way to program a multiple processor system (whether tightly or loosely coupled) to accomplish an integral distributed processing application? Writing concurrent programs for a uniprocessor is tough enough, but writing programs which interact and operate simultaneously in parallel can be a most difficult and frustrating experience. Opportunities abound for operational failures due to race conditions, for time-dependent bugs and for deadlock situations.

Help is on the scene, though, in the form of new concurrent languages as typified by Concurrent Pascal.[4] The new software technology embodied by these languages can be applied to multiple processor problems as a methodology regardless of the implementation mechanisms.[23,25] Nevertheless, the utility of having an effective language is beyond question, even if only as a design tool.

A key feature of Concurrent Pascal is the monitor construct that protects critical data regions shared among cooperating sequential processes. With a mutual exclusion mechanism, only a single process is permitted to access the critical region at any given time. This notion was first suggested by Dijkstra,[11] formalized by Hoare,[13] and implemented by Brinch Hansen in Concurrent Pascal. Monitors, or an equivalent construct or capability, have since been incorporated in many other languages.

Although different linguistic variations are possible, Concurrent Pascal was selected as a base for implementing distributed processing programs because of its track record and extensive documentation. The language has proved to be a powerful and effective tool in practice for building structured concurrent programs.[5] Brinch Hansen recorded improvement in programmer productivity while building a complete operating system with his language,[6] and the utility of the language has been tested for many diverse applications.[29]

There has been some criticism of the language, however. For one thing, it depends on a run-time kernel facility that is invariant and built with a different language.[20] For another, critical system design decisions have been assumed by the language.[24] Researchers are also actively pursuing improved language constructs, most notably the manager concept,[18,27] which ultimately may lead to simpler and even more reliable concurrent programming concepts.

"A Language for Distributed Processing" by Ronald J. Price, Perkin-Elmer Data Systems Group. From the 1979 PROCEEDINGS of the National Computer Conference. Reprinted by permission of AFIPS and the author.

A Language for Distributed Processing

The purpose of this report is to propose two fundamental modifications to Concurrent Pascal that not only will alleviate many of the above concerns, but more importantly, will extend the language's applicability to distributed system environments.

In many respects, the proposed changes are adaptations of principles incorporated in Wirth's real-time language Modula.[31] As presented in the next two sections, they would enable the kernel and system control operators (i.e., the lowest levels of an operating system) to be written in the language itself and would enable partitions of a global, distributed multiprocessing program to be mapped to physical processors, but yet represented as an integral program.

The last section of the report summarizes the proposed concepts and applies them as a methodology for constructing systems—from kernels, across processor boundaries, and up through application programs. As such, the extended language is a systems description language in that it can be employed to describe the algorithmic behavior of a multiple processor system (not to be confused with a hardware description language that prescribes physical circuits). It offers the systems designer a tool for:

- Synthesis
- Documentation
- Modeling
- Simulation
- Verification

and implementation if used directly as an implementation language.

Although the emphasis of this report is on distributed processing, the proposed extensions increase the power of the language for solving complex operating system problems irrespective of the multiprocessing issues. For example, the following problem areas are difficult under Concurrent Pascal as defined, but are quite amenable with the modified language:

- Data communications
- Process creation
- On-line system generation
- Dynamic software restructuring

The main intent of this report is to justify and to explain the benefits of the proposal, not to specify the language nor to suggest a method of implementation. Semantic details and the mechanics of integrating the new constructs within the language need further study and exposure to actual practice.

The level of presentation assumes the reader is familiar with Concurrent Pascal, but the definition of a few items might be useful. A program that can be described by the language is called a concurrent program. It consists of system (or program) components defined as process types and monitor types (and class types not mentioned here); redefinition of the monitor type and the definition of a task component as a partition of a concurrent program are described. A Concurrent Pascal program includes a programmable initial process that directs the initialization of the components in the program. The interpretation of concurrency is the execution of multiple processes overlapped in time, either by multiplexing periods of execution on a single machine or by simultaneous execution on multiple machines. When important, the latter connotation of true concurrency (i.e., parallelism) will be explicitly denoted in context; multiprocessing implies parallelism, for example.

CONCURRENT PASCAL WITH A PROGRAMMABLE KERNEL

As represented in Figure 1, Concurrent Pascal is based on a virtual machine kernel that implements process switching, mutual exclusion on access to monitors, and the various control operators (DELAY, CONTINUE, etc.). The definition of the virtual machine interface can be a problem for system builders interested in different kernel features and/or in multiple machine operations. The problem is that the virtual machine has been abstracted

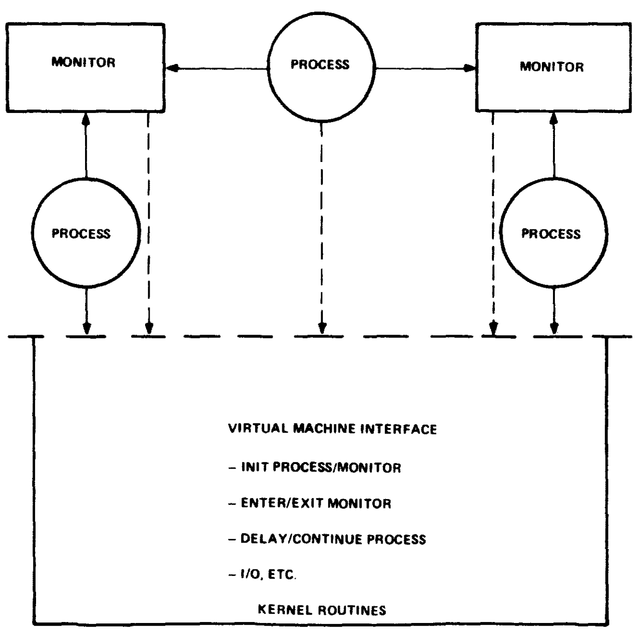


Figure 1. Concurrent Pascal system

Note: The arrows in Figure 1 and in the following figures that depict a concurrent program represent access rights as defined in Concurrent Pascal, and not the flow of data. Further, circles represent processes and boxes represent monitors.

A Language for Distributed Processing

away from the systems programmer to the point of existing literally in another world as defined by its unique language (typically assembly). Moreover, the line between the real and virtual machine might not be optimum for a given application. There are simply too many variables, parameters, factors and extenuating circumstances to consider in general.

In some situations, the programmer of the concurrent program would like to have an influence on the design of one or more of the virtual machine modules, sometimes even to interact with the internal machine dynamically. A prime example of this is programming interrupt service routines. Interrupt handling (typically for I/O processing) is related more to an application than to central general purpose kernel routines; this is clearly so in dedicated systems.

The handling of interrupts has historically caused untold grief and frustration for system programmers. The interrupt is an indeterminate and irreproduceable happening. Contemporary systems researchers recommend against using it as a synchronization mechanism and avoid preemption in general. The notion of an interrupt does not even exist in Concurrent Pascal. Instead, synchronizing primitives are provided (DELAY and CONTINUE) that allow system programs to be designed with so-called cooperating sequential processes.

Unfortunately, the processes embodied by most peripheral devices on even modern computers cannot be considered cooperative. Modula was designed to handle them.[32] But even if we stopped using the interrupt as a synchronizing mechanism, we still need it as a signal with which to measure time and to build real-time functions.

So although we might want to hide the interrupt in some abstract way, we still have to deal with it. Today this is generally accomplished through the kernel. However, not only is the interrupt hidden by the kernel, it is also typically unaccessible to the high-level software in a direct manner. Brinch Hansen and Hoare point out that scheduling cannot rely solely on built-in abstractions and that high-level software should be in control of response times at the lowest level.[3] Indeed, the interrupt is the simplest form of low-level scheduling for machines that can switch an instruction stream automatically upon recognizing an external signal. (Some machines provide multiple priority states where an interrupt level may be interrupted by yet another level, but for purposes of discussion, a single level is assumed here.)

The ability to dispatch programmable service routines in rapid response to external signals and to manage them in a disciplined manner could be afforded to Concurrent Pascal by extending the language with a new construct that allows procedures to be called with interrupts disabled. To allow controlled sharing of the uninterruptable procedures and their data structures, they could be treated much like the ordinary "virtual-time" (i.e., interruptable) monitors. This new construct could then take the form of another system type in the language—a "real-time" monitor (for want of a better name). Generally speaking, the idea being presented here is to incorporate the real-time principles of Modula within the framework of Concurrent Pascal. Actually, we need not add a new system type to the language, but only have to redefine the monitor to include statements that execute in real-time.

The use of "real-time" monitors for interrupt handling is illustrated in Figure 2. Different delay (wait-on signal) and continue (send-signal) operators would be needed that are consistent with the real-time environment. With appropriate entry and exit mechanisms, processes would communicate directly with interrupt service routines without going through pre-defined intermediary kernel routines; even interrupt service handlers could directly intercommunicate.

The "real-time" monitor construct would have far more application than just for programming interrupt handlers. For example, when multiprogramming a single machine, mutual exclusion on access to a monitor is assured simply by having interrupts disabled. In fact, there can be no busy queueing of processes on a "real-time" monitor. Consequently, they could be used in certain situations as a more efficient substitute for the ordinary "virtual-time" monitors in Concurrent Pascal.

Moreover, since the procedures in "real-time" monitors represent indivisible operations to their using program components, they can be employed to implement Concurrent Pascal's "virtual-time" monitors with the language itself. That is, in Concurrent Pascal a process does not directly call a monitor procedure. The call is actually intercepted by a kernel routine to perform mutual exclusion and busy queueing if necessary. This kernel intervention is installed by the compiler in a transparent manner to the programmer. Under the proposed language, this kernel routine would be programmed explicitly and not automatically installed by the compiler (except possibly by default as an implementation-dependent feature).

The various monitor operators and, for that matter, any kernel-like function the systems builder needs would also be programmed in a direct manner. Even conditional critical regions with different

A Language for Distributed Processing

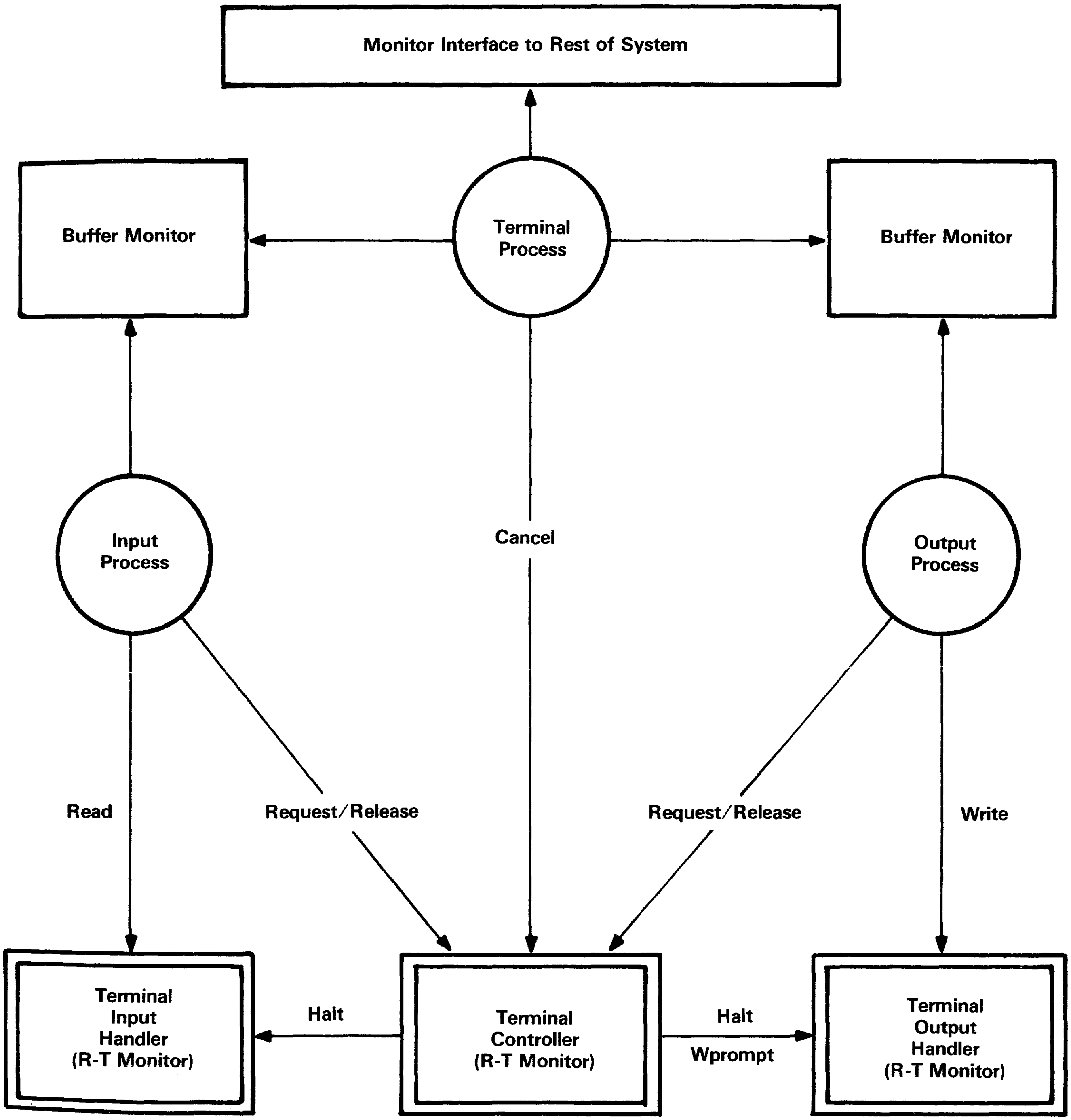


Figure 2. Concurrent program with real-time monitors for terminal I/O handling

scheduling algorithms (guarded regions[8]) can be implemented with this "real-time" construct. In other words, the "real-time" monitor is a means for implementing explicit kernel routines, although the compiler could still support standard implicit kernel calls in a transparent manner.

Figure 3 is an extension of Figure 2 with kernel modules illustrated. An important aspect of this viewpoint is that the full power of the language can be brought to bear on the construction of the lower-level software when it is included as an integral part of the entire system. Such capability is important for embedded systems, process control environments, and data communications applications.

Kernel-like functions could be "hidden" through levels of abstraction, but this would be up to the systems

A Language for Distributed Processing

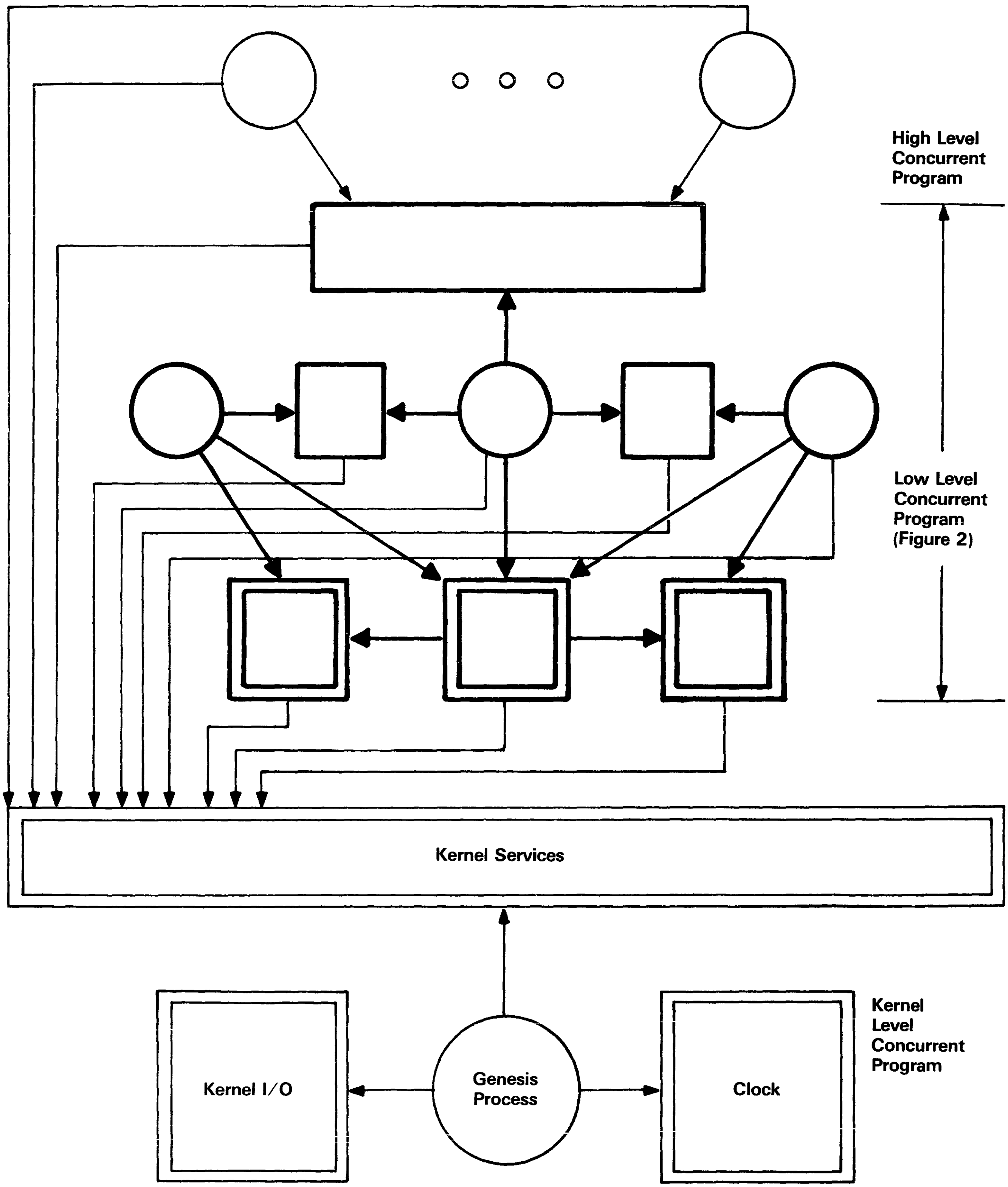


Figure 3. Multiple levels of a concurrent system program

builder and not a condition of the language. In fact, no run-time program, nor a pre-defined kernel definition, is required to support the proposed language.

The kernel can be treated as a concurrent program in its own right,[19] and Figure 3 also illustrates this point. The Genesis process interacts with external processes in peripheral equipment through real-time monitors. It also performs system initialization and takes on the role of the initial process of a concurrent program as per Concurrent Pascal, including in this case the explicit creation of the high-level abstracted user processes. The Kernel Services real-time monitor in this example provides the standard Enter, Exit, Delay, Continue, etc., procedures and a Dispatch procedure for multiplexing processes. The kernel might control private devices as illustrated, but interrupt handling for the higher-level software would also be supported (typically with considerable hardware assist) for dispatching processes in real-time monitors in response to interrupt signals.

The Genesis process selects high-level processes to execute with the Dispatch procedure and executes them much like a subroutine with interrupts enabled. So the high-level abstracted processes are in reality still the Genesis process in disguise. When the Genesis process recognizes an interrupt signal (presumably with hardware assist), it enters Kernel Services (with interrupts disabled) and takes appropriate action. In the event this action results in activating a waiting process, the Genesis process can decide whether to preempt (reschedule) the current running process or to schedule the waiting process. Typically, the action in response to an interrupt signal would be to dispatch the recipient process immediately in its real-time monitor which in turn would initiate scheduling actions as required.

Many of these low-level functions could be implemented in hardware or firmware. Nevertheless, they can be accurately represented and programmed with the "real-time" monitor construct.

Incorporating the real-time feature does not make the proposed language machine-dependent. From a language point of view, the new proposed construct simply represents the sequential state of the machine. However, escape mechanisms would have to be provided in the compiler for programming machine-dependent features in the low-level software modules; or provide machine-dependent statements as an adjunct to the high-level machine-independent language.

A MULTITASKING CONCURRENT PASCAL

The representation of a kernel as a concurrent program becomes more important when we consider a multiple processor system. Figure 4 is an example expansion on Figure 3 to illustrate kernels for a three-processor system; the surrounding higher-level software is not illustrated. The Inter-Kernel Communication (IKC) monitors are real-time monitors designed for exchanging information between kernels.

As should be evident from the previous discussion, the Genesis process together with the support modules in each kernel's partition is actually a sequential program running on a sequential machine; parallelism is just an illusion to the higher-levels of software. Even in Saxena's verification of the monitor concept,[26] he had to represent the idle state of multiple physical processors with an idle process for each processor, the equivalent of the Genesis process. Consequently, in order to represent true concurrency (i.e., parallelism) we need a mechanism for representing the multiple processors, or at least the actions of their kernels.

Even if we were to assume the prior existence of a collection of cooperating kernels on multiple machines that form a virtual multiple instruction, multiple data path machine on which we somehow apply the high-level concurrent program, we still could not take full advantage of the parallel machine with Concurrent Pascal as defined. Loading and initialization of the program, for example, must take place sequentially, either on a single processor or in sequential phases on multiple processors because the initial process of the concurrent program is really the Genesis process of a single kernel.

So we need a way of dividing the global concurrent program into logical partitions that can be delegated to separate processors for initiation and execution. Indeed, we have no viable alternative but to divide the program into physical partitions if it is going to be run on a loosely-coupled configuration that does not share memory.

The partitioning mechanism proposed here is to extend Concurrent Pascal with a *task* block structure somewhat analogous to *module* in Modula. The task construct, however, defines a concurrent system component that contains a collection of processes and monitors. It specifically includes an initial process for initializing the task. And tasks cannot be nested. A multiple task program represents parallelism in that different tasks, via their initial processes can be dispatched and executed simultaneously by separate processors. In other words,

A Language for Distributed Processing

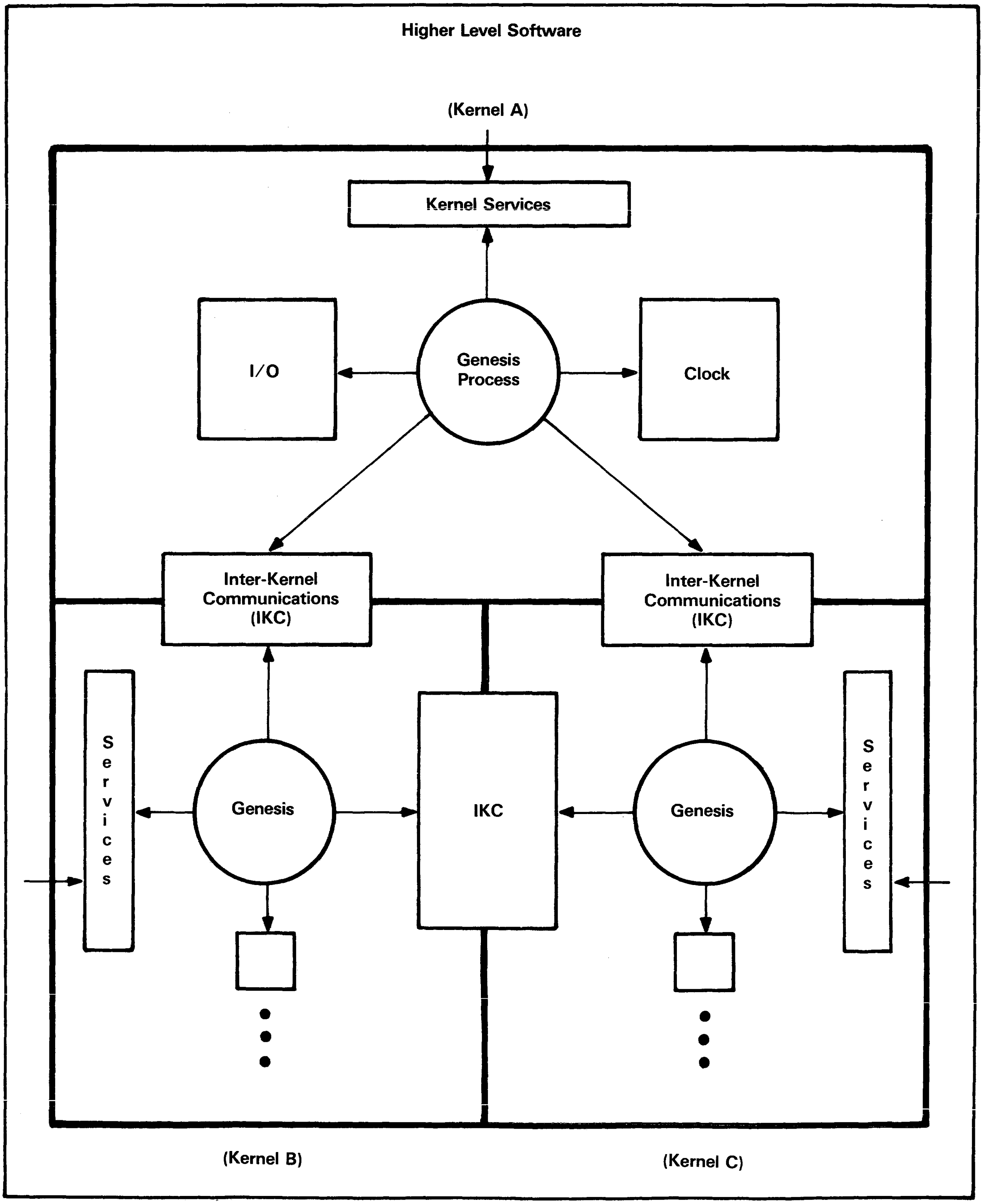


Figure 4. Multiprocessing kernels

A Language for Distributed Processing

a multiprocessing program can include multiple initial processes which represent abstracted extensions of multiple kernels.

Each kernel in Figure 4 would be represented by a separate task, and each would be dedicated to a specific processor. The higher-level software could be implemented as extensions of each kernel or as separate tasks. As one or more tasks on a tightly-coupled system, the high-level modules need not be dedicated to specific machines and could be dispatched by any of the three kernels.

Each task is, in essence, an independent concurrent program and can be compiled into a separate load module. Tasks are linked at run-time to form a global system.

The correctness of the system can be tested with an integral compilation where the tasks interact through monitors at the interface of the task boundaries. The compilation of any given task, however, need only include its predecessor tasks in the system and not any task outside its view of the system.

Regardless of the issue of being able to express parallelism in the language, the task construct is a tool for partitioning a multiprocessing system program. Access rights as implemented in Concurrent Pascal will assure a structured design.

We can divide a concurrent program into sections by taking advantage of the isolation property of monitors. That is, processes intercommunicate and synchronize their operations through monitors, and consequently, they need not know anything about each other—even their existence. For example, in Figure 5 the User B process need not know of the presence of the User A process when calling the Buff 2 monitor, nor for that matter, even if multiple job processes interface the Buff 2 monitor. Therefore, we can safely cut the program between monitors and processes as illustrated. The trick is to keep the access right arrows pointing in the same direction across the task boundary. (Whether task initialization is performed by a separate initial process or one of the application processes in each task is not relevant to this example.)

Note that the system structure and hierarchical order of the program components, as required by Concurrent Pascal, is preserved if we define and initiate Task A before Task B, even if one physical processor dispatches Task A and another processor dispatches Task B. This would not be the case, however, if the Job 3 process were included in the Task A partition because then each task would have access rights to each other in a cycle.

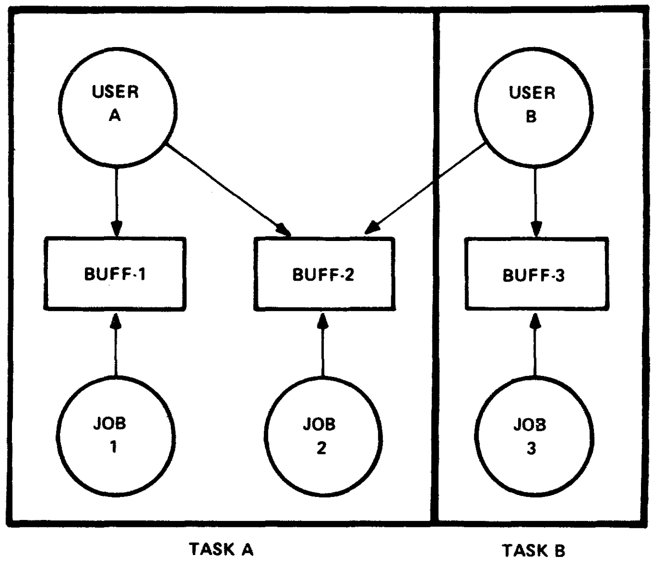


Figure 5. Partitioning a concurrent program

Sometimes the initial layout of a concurrent program does not lend itself to partitioning. For example, if we tried to apply the tasks in Figure 6 to two different machines, the multiprocessing program could easily crash when started (even if one task is initiated before the other) because the design does not guarantee that the monitors will be initialized before being called. But then the program might not crash; the program is a time-dependent race condition.

Start-up is only part of the problem. We also need orderly ways of stopping a multiprocessing program, and more importantly, mechanisms for detecting error situations across processor boundaries and recovering from them. This is what partitioning is about.

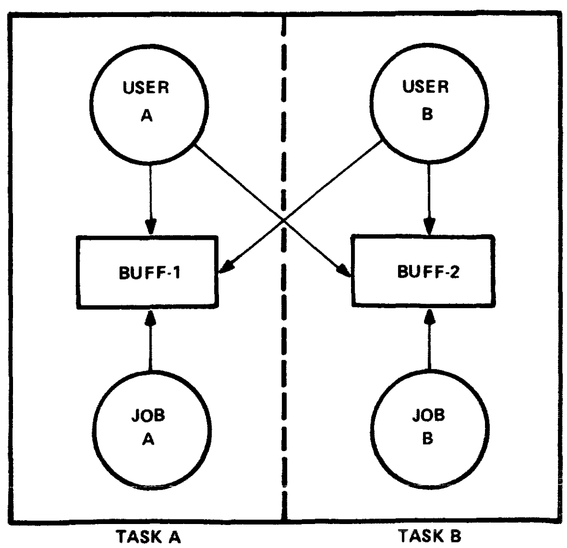


Figure 6. Invalid task partitioning

A Language for Distributed Processing

Figure 7 shows how we can take advantage of the insertion property of monitors to resolve this task layout problem. Here, a message exchange monitor and server process are inserted in the User A process access path. This gets the arrows pointing in the same direction across the task boundary. The server process acts in behalf of the User A process in the Task B partition.

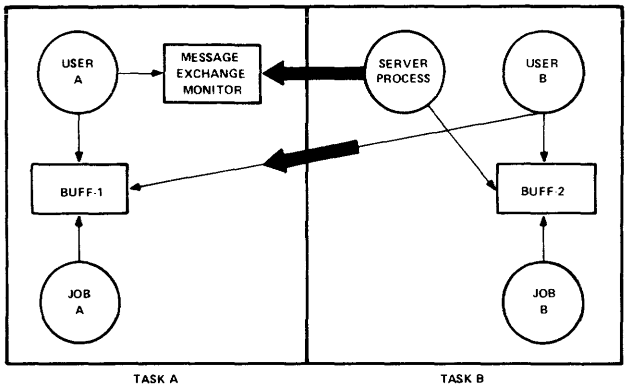


Figure 7. Partitioning with insertion

A correct way to partition a multiprocessing program is to group logically-related processes and monitors into separate tasks in such a manner that their access rights point in the same direction across the task boundaries and by arranging the tasks in a hierarchy such that tasks which access other tasks are ranked below their predecessors, as in Figure 8. This ranking assures an orderly initialization (and termination) and eliminates race conditions and deadlock situations that otherwise might occur with a cyclic control structure.

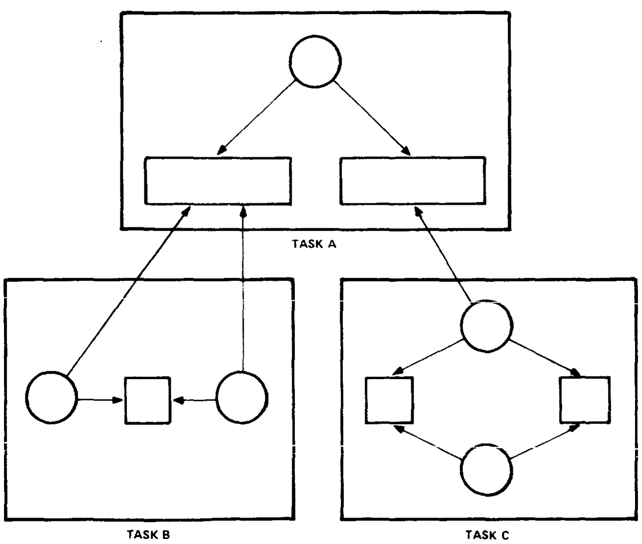


Figure 8. Hierarchical structuring of multiprocessing programs

In some arrangements, tasks, such as Task C in Figure 8, can be literally removed and brought back on-line without disturbing the rest of the system. The status of Task A does have to be known to Tasks B and C, however. In essence, the kernel tasks (not illustrated) and Task A form a virtual machine for Tasks B and C. This capability allows a system program to be generated and restructured dynamically.

DISTRIBUTED PASCAL

A key feature of this proposal for implementing distributed programs is the ability to describe interface monitors between processors. The characteristics of a given interface can be programmed with the "real-time" monitor construct. Parallel kernels can then be described with the task block structure where the legality of their interface monitors is tested with an integral compilation. Higher-level tasks are built on top of the kernel tasks.

Interface monitors can be implemented in shared memory employing "thick-wire" communication techniques or in shared "thin-wire" I/O facilities. Mutual exclusion between machines is achieved by mutual cooperation in adhering to a protocol.

In the thick-wire case, permission to access the data structures is achieved by locking the monitor with a read-modify-write operation (e.g., Test and Set instruction) and then the data structures are manipulated in place. The logic for manipulating the data (i.e., the monitor's program code) can also be located along with the data if the hardware configuration allows code to be executed out of shared memory, or otherwise the logic can be replicated in the private memory of each processor.[9,25]

In the thin-wire case, data are physically copied from one location to another. Although Concurrent Pascal's monitors cannot be directly supported across a thin-wire boundary, an abstracted user's environment illustrated by Figure 9a could be supported by an underlying message communications system as depicted by Figure 9b. This software message system is conceptually the same thing implemented in hardware to support shared memory; however, the flexibility of a thick-wire emulation in software has to be highly constrained because of the limited bandwidth and long response times of the communication facilities.

A good case can be made for adopting a standard thin-wire communication technique for multiple processor systems which is adaptable to networks as well as to tightly-coupled architectures.[15, 22, 30] The overhead normally associated with a message-based system can be ameliorated by implementing message exchange facilities in hardware.[16,28]

A Language for Distributed Processing

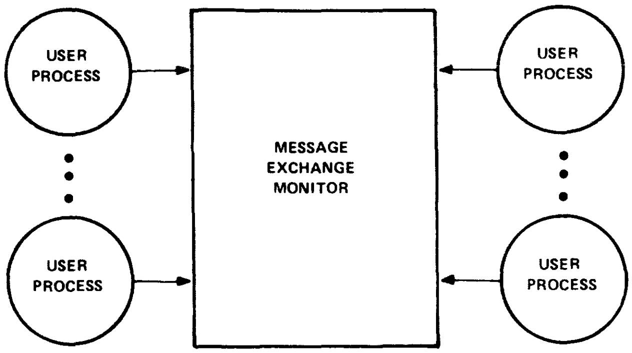


Figure 9a. Message-based concurrent program

Special languages have been proposed for message systems,[1,21] but Concurrent Pascal is a very suitable language for expressing networked systems,[7] including the communications protocol.[2,9] Moreover, the language offers the flexibility of general monitor designs where appropriate in addition to any built-in message exchange monitors of the communications system.

In any case, Concurrent Pascal as proposed to be modified is open-ended in the sense that both communication approaches can be accommodated. For example, if an operating system built with the language establishes a message-based inter-process communications protocol for conventional system use, the underlying implementation can still be based on thick-wire techniques where appropriate and more efficient.

Figure 10 depicts a multiple-task, multiple-processor system employing both thick-wire and thin-wire communications. The kernels dedicated to the processors in each tightly-coupled dual processor complex interface through multiple real-time monitors in shared memory, whereas the two complexes interface through a single real-time monitor over a communications channel. Kernels are represented by tasks and form the lower levels of the system. Higher-level tasks in the global system intercommunicate through monitors in a hierarchical fashion, as well. It is important to note that the different levels do not necessarily imply physical levels; that is, virtual kernels that emulate process switching, interrupts, etc., on top of real kernels is not a requirement to support high-level concurrent programs.

The point being made here is that this total system can be described with a single program (although part of it might be implemented in hardware). The program consists of a set of cohesive routines (program components) that implement the behavior of the global system. Indeed, the whole system operates as a harmonious confederation of cooperating sequential processes, some of which may run in parallel.

Even if the language is used only as a modeling tool, it can help us to design reliable systems by applying computer programming technology to their construction. This is so because Concurrent Pascal is based on proven software engineering techniques.

When building the "THE" multiprogramming system, Dijkstra suggested employing hierarchical levels of abstraction as a methodology for dealing with the complexity of operating systems;[10] that is, modules are built on top of others with well defined interfaces and interactions. This technique, actually a formal method of structured programming, is an invaluable aid for proving program correctness and is an inherent capability of Concurrent Pascal.

The axiomatic definition of Pascal[12] and the treatment of critical regions (e.g., monitors of Concurrent Pascal) and other research efforts have led to many proofs of program correctness relevant to concurrent

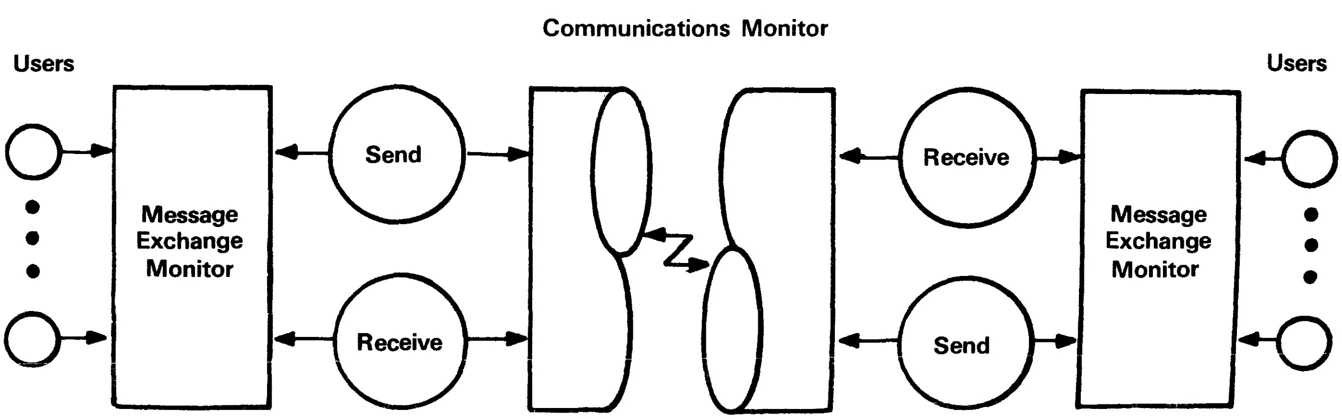


Figure 9b. System implementation of message communications

A Language for Distributed Processing

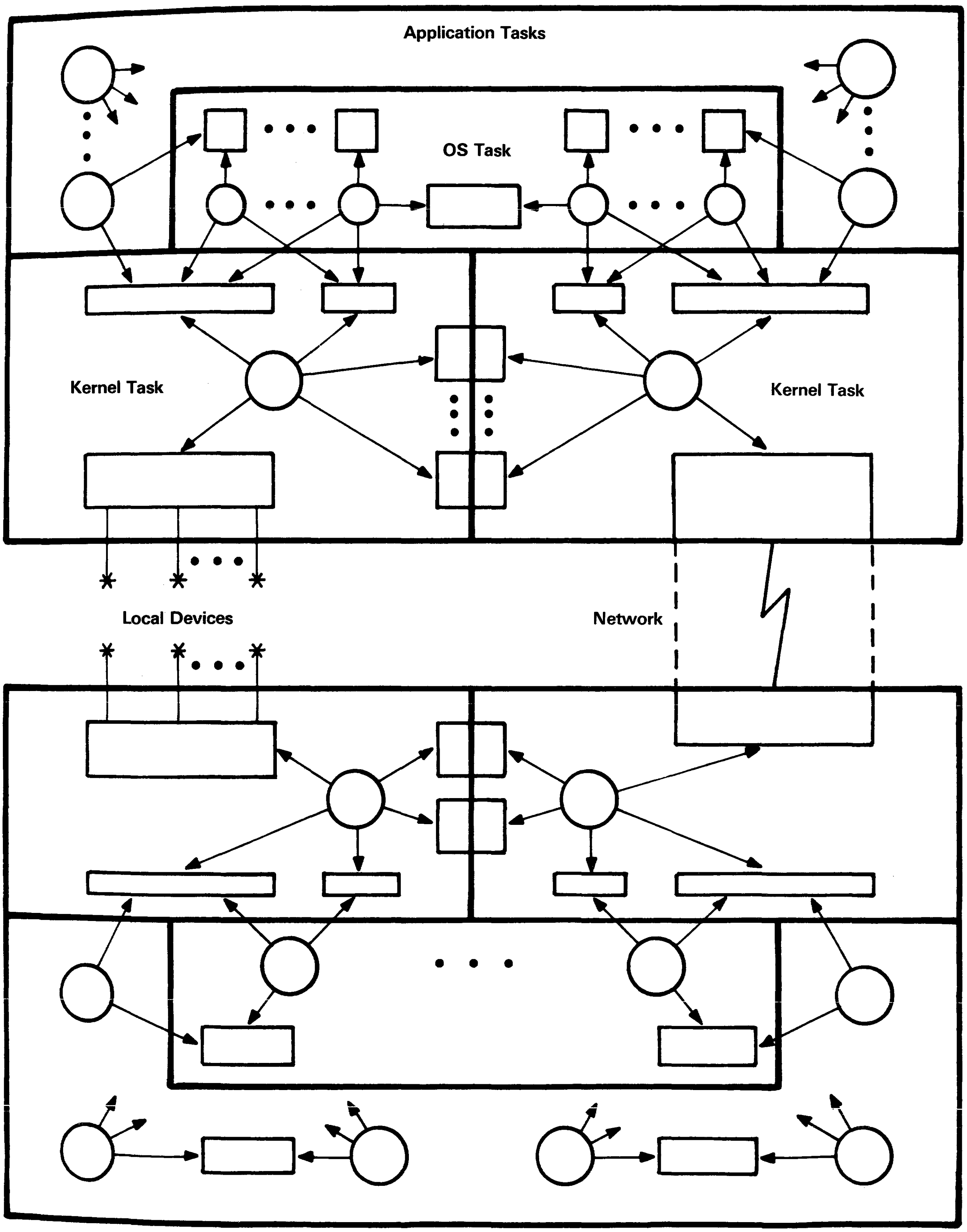


Figure 10. Distributed processing system

A Language for Distributed Processing

programming (e.g., References 8, 14, 17, 26). These principles are now being applied in attempts to discover simpler, more flexible and more reliable techniques for constructing monitors and by enlisting the aid of the compiler itself.

The fact that formal constructs can lead to provably-correct programs may sound academic in reality. However, they actually do in practice lead to rapid program synthesis and to program correctness by inspection. Testing becomes much more systematic and takes on more of a verification role than a debugging operation. Modification and maintenance are also assisted.

By maintaining the consanguinity of Concurrent Pascal as proposed, we can apply these formal constructs to the construction of distributed systems. The language serves as a synthesis aid by enabling the system designer to decompose a system in terms of task components which in turn are decomposed into logically-related processes and monitor components; this can be illustrated in diagrammatical form. Moreover, the language allows a system to be designed in incremental stages, and it can be used to simulate and to evaluate different implementation strategies. In particular, the system designer can describe proposed solutions as models that accurately represent the physical environment and that can be demonstrated to run correctly. The language can also serve as a vehicle for documentation and testing. Finally, it becomes a piece-part of the end product where it is employed as an implementation language.

CONCLUSION

Changes to the language Concurrent Pascal are proposed that enable it to be used to:

1. Describe the algorithmic behavior of the physical system.

2. Express the physical parallelism of a distributed multiprocessing program.

As such, the new language acquires the connotation of Distributed Pascal.

REFERENCES

[1]Ambler, A.L., et. al., "GYPSY: A Language for Specification and Implementation of Verifiable Programs," *ACM SIGPLAN Notices,* Vol. 12, No. 3, March 1977, pp. 1-10.

[2]Bochmann, G.V., "Logical Verification and Implementation of Protocols," *Proceedings Fourth Data Communications Symposium,* October 1975, pp. 7-15 to 7-20.

[3]Brinch Hansen, P., et. al., "Process Dispatching Techniques," *Operating Systems Techniques,* C.A.R. Hoare and R.H. Perrott (ed.), Academic Press, New York, 1972, pp. 201-207.

[4]Brinch Hansen, P., "The Programming Language Concurrent Pascal," *IEEE Transactions on Software Engineering,* Vol. SE-1, No. 2, June 1975, pp. 199-207.

[5]Brinch Hansen, P., *The Architecture of Concurrent Programs,* Prentice-Hall, Englewood Cliffs, New Jersey, 1977.

[6]Brinch Hansen, P., "Experience with Modular Concurrent Programming," *IEEE Transactions on Software Engineering,* Vol. SE-3, No. 2, March 1977, pp. 156-159.

[7]Brinch Hansen, P., "Network: A Multiprocessor Program," *IEEE Transactions on Software Engineering,* Vol. SE-4, No. 3, May 1978, pp. 194-199.

[8]Brinch, Hansen, P., "Specification and Implementation of Mutual Exclusion," *IEEE Transactions on Software Engineering,* Vol. SE-4, No. 5, September 1978, pp. 365-370.

[9]Cavers, J.K., "Implementation of X.25 on a Multiple Microprocessor System," *Proceedings 1978 International Conference on Communications,* 1978, pp. 24.6.1-24.6.6.

[10]Dijkstra, E.W., "The Structure of the 'THE' Multiprogramming System," *Communications ACM,* Vol. 11, No. 5, May 1968, pp. 341-346.

[11]Dijkstra, E.W., "Hierarchical Ordering of Sequential Processes," *Acta Informatica,* Vol. 1, 1971, pp. 115-138.

[12]Hoare, C.A.R. and N. Wirth, "An Axiomatic Definition of the Programming Language PASCAL," *Acta Informatica,* Vol. 2, 1973, pp. 335-355.

[13]Hoare, C.A.R., "Monitors: An Operating System Structuring Concept," *Communications ACM,* Vol. 17, No. 10, October 1974, pp. 549-557.

[14]Howard, J.H., "Proving Monitors," *Communications ACM,* Vol. 19, No. 5, May 1976, pp. 273-279.

[15]Jensen, E.D., "Distributed Processing in a Real-Time Environment," *Distributed Systems, Infotech State-of-the-Art Report,* Infotech International Ltd., Berkshire, England, 1976, pp. 303-318.

[16]Jensen, E.D., "The Honeywell Experimental Distributed Processor—An Overview," *IEETC,* Vol. 11, No. 1, January 1978, pp. 23-38.

[17]Karp, R.A., and D.C. Luckham, "Verification of Fairness in an Implementation of Monitors," *Proceedings of 2nd International Conference on Software Engineering,* October 1976, pp. 40-46.

[18]Kieburtz, R.B., and A. Silberschatz, "Capability Managers," *IEEE Transactions on Software Engineering,* Vol. SE-4, No. 6, November 1978, pp. 467-477.

[19]Lister, A.M., and P.J. Sayer, "Hierarchical Monitors," *IEEE Proceedings 1976 International Conference on Parallel Processing,* pp. 42-49.

[20]Lohr, K., "Beyond Concurrent Pascal," *Proceedings of 6th ACM Symposium on Operating System Principles, ACM Operating Systems Review,* Vol. 11, No. 5, November 1977, pp. 173-180.

[21]May, M.D. et. al., "EPL—An Experimental Language for Distributed Computing," *Proceedings NBS-IEEE Trends and Applications: Distributed Processing,* May 1978, pp. 69-71.

[22]Metcalfe, R.M., "Strategies for Interprocess Communication in a Distributed Computing System," *Symposium on Computer-Communications Networks and Teletraffic,* Polytechnic Institute of Brooklyn, April 1972, pp. 519-526.

[23]Paquet, J.L., et. al., "Concurrent High-Level-Language Machines and Kernels." *Proceedings of the IEEE International Symposium on Mini- and Microcomputers,* November 1977, pp. 293-298.

[24]Parnas, D.L., "The Non-Problem of Nested Monitor Calls," *ACM Operating System Review,* Vol. 12, No. 1, January 1978, pp. 12-14.

[25]Price, R.J., "Multiprocessing Made Easy," *Proceedings National Computer Conference,* AFIPS, 1978, pp. 589-596.

[26]Saxena, A.R., and T.H. Bredt, "Verification of a Monitor Specification," *Proceedings of 2nd International Conference on Software Engineering,* October 1976, pp. 53-59.

A Language for Distributed Processing

[27]Silberschatz, A., et. al., "Extending Concurrent Pascal to Allow Dynamic Resource Management," *IEEE Transactions on Software Engineering,* Vol. SE-3, No. 3, May 1977, pp. 210-217.
[28]Swan, R.J., et. al., "Cm*—A Modular, Multi-microprocessor," *Proceedings National Computer Conference,* AFIPS, 1977, pp. 637-644.
[29]Vallentine, V., and R. McBride, *Concurrent Pascal—A Tutorial,* Department of Computer Science, Kansas State University, November 1976.
[30]Wecker, S., "A Design for a Multiple Processor Operating Environment," *Proceedings 7th Annual IEEE Computer Society International Conference,* COMPCOM '73, February 1973, pp. 143-146.
[31]Wirth, N., "Modula: A Language for Modular Multiprogramming," *Software Practices and Experiences,* Vol. 7, No. 1, January 1977, pp. 3-84.
[32]Wirth, N., "Toward a Discipline of Real-Time Programming," *Communications ACM,* Vol. 20, No. 8, August 1977, pp. 577-583.□

Maintaining Data Order and Consistency in a Multi-Access Environment

Problem:

Within the distributed processing environment a file, application program, or data item is accessible by, and available to more than one location. Supposing two or more locations want to access the same file or record simultaneously. If both wanted only to read the contents of a record for example, there would be no potential conflict. However, if both locations wanted to write to the record, some control over the sequence of events must be exercised. The problem grows with the addition of more locations simultaneously attempting to do the same thing. This report describes the approach taken to solve this and related problems for the SIGMA message service within the Advanced Research Projects Agency (ARPA). The message service had been operational for three years, at the time of the original presentation of this material occurred.

Solution:

The problem of controlling simultaneous access to shared data runs throughout the history of computer science. In order to preserve the consistency and integrity of such data, computer scientists and programmers have developed locks,[3] semaphores,[3] the notion of critical sections,[4,3] monitors,[5] and innumerable other techniques, both concrete and abstract. The great interest in such mechanisms throughout the computer community, both in the literature and in practice, is indicative of the importance of the problem.

Classical Examples

Control over simultaneous access is necessary in countless practical applications, one of the most common of which is the need to guarantee strictly sequential access to critical resources by competing—conceptually or literally—parallel executing processes. To provide the necessary interlocking, implementors have used techniques such as the simple lock and its more sophisticated cousin, the mutual exclusion semaphore. In each of these, before being allowed to access the critical resource, a competing processor must pass a decisive test, engineered to ensure that

- Only one process can pass the test, no matter how many processes are competing or how frequently they attempt access.
- Once a process has passed the test and is granted access to the resource, no other process can pass until the first has "released" the resource.

A different type of control over simultaneous access is necessary to satisfy another problem, referred to in the literature as the "Readers and Writers Prob-

"Maintaining Order and Consistency in Multi-Access Data" by Ronald Tugender, USC/Information Sciences Institute. From the 1979 PROCEEDINGS of the National Computer Conference. Reprinted by permission of AFIPS and the author.

Maintaining Data Order and Consistency in a Multi-Access Environment

lem."[2,1] In this case several readers accessing shared data wish to ensure that the data they are reading remains constant during the reading period. It can be summarized by the following rules:

- Any number of Readers may access the data simultaneously, but if any Reader has access, no Writer may also have access until all Readers are finished.
- Once any Writer has gained access, no Reader or other Writer may have access until the Writer has finished.

Multi-Access Objects in SIGMA

Neither of the above techniques is appropriate if any of the competing processes need extended access to the resource. Any process doing so would block other processes from completing their tasks, potentially lengthening the time required to complete them significantly and, in an interactive environment, slowing the response time. In a situation in which a process needs to read access to a data object for a long time, possibly concluding with a need to modify the object, neither of the above techniques can offer the guarantee of integrity of the data as well as provide timely interactive response when many processes are competing.

In the SIGMA message service[7,8] the two most important and frequently used data objects, messages and folders, are both typically shared among several users. At any time, any or all of the users allowed to share one of these objects may "open" it (request read and potentially write access), read through it for an arbitrarily long time and modify it in any of several ways. A scheme had to be developed which allowed simultaneous access by several users for arbitrary lengths of time, preserved a consistent image of the object for each user even while it was being modified by others, and both permitted and correctly assimilated modifications made in parallel by several users. At this point a description of the users' perception of the two SIGMA shared objects is in order, to illustrate the issues involved in parallel modification.

SIGMA Messages

A SIGMA message is conceptually similar to a formal business letter with a sender, addressees, a body and various other information. In draft form (i.e., before it is sent), it may be read and revised by several reviewers before it is actually approved for sending. Several people may be involved during the draft/revision process, and for time efficiency it is often desirable for them to work in parallel. SIGMA provides this parallelism by giving each reviewer an up-to-date copy of the current draft message, allowing him to edit, revise and comment as he sees fit. Each user's rendition of the message, called a messagette, contains all the original, unmodified parts of the original draft, his own modified sections replacing their original counterparts, plus any new text and comments.

After the message has been sent (referred to as a transmitted message), its character changes. It is no longer a draft entity, subject to revision. Just like a letter which has been dropped into a mailbox, it has become an official document, the contents of which are now on record. Users may read it and make comments, but are not allowed to change its contents in any way.

SIGMA Folders

A SIGMA folder can be likened to a file into which messages are placed. In SIGMA, however, the folder contains not the messages themselves but rather abstracts, called entries, containing a pointer to the actual message (for retrieval purposes) and a subset of the information contained in the referenced message. When messages are sent to a user, the entries referencing them are automatically placed in a particular folder named "Pending" (analogous to a mail in-basket). A user can create other folders and copy entries between them, where the number of folders needed or the significance of the entries placed into them is left entirely to the user's discretion. To help him locate specific entries within folders, a user is provided a rich set of searching tools, including the ability to associate entries with user-chosen keywords. A user can also place comments on entries and delete entries that are no longer needed in the folder.

A user may permit any or all of his folders to be accessed by other users, in which the other users are allowed the same searching facilities available to the owning user, as well as to read the abstracts, make and read comments, and retrieve referenced messages. To allow users to peruse folders conveniently, SIGMA also maintains a place-marker in a folder for each user, marking the last entry he has referenced; when he next accesses that folder, he will be returned to the same entry.

Further Complications

The requirement that many users be allowed to modify an object in parallel posed several logical as well as implementation problems. It became apparent that the solution involved more than simply providing the correct form of interlock apparatus; it also had to preserve as much as possible the intent of the modi-

Maintaining Data Order and Consistency in a Multi-Access Environment

fiers. This caused two complications well beyond the scope of the classical techniques described earlier:

1. "Whoops, where'd it go?"—From the application system standpoint, even assuming that writers are prevented from simultaneous modification, logical inconsistencies can still occur if they can write arbitrarily when it is their turn to write. Consider an object represented as a linked list, where the types of modification allowed are add, replace and delete elements of the list. What does it mean to replace an element just deleted by another writer, or to add an element adjacent to one just deleted?

2. "That's not how I remember it!"—Just as important as the ability to produce a logically consistent object is the need to have the updated object conform to each user's expectations of how it should appear after modification. Consider the typical situation which occurs when several authors or reviewers are allowed to edit a draft document in parallel, and they specify disparate sets of changes to a common secretary. When the updated draft appears, one or more authors may be surprised that the new draft does not reflect the changes they specify.

Clearly the above situations could be avoided if some restraints were placed on the types of modifications allowed. In the latter case, for example, the problem lies not in the authors nor in the secretary, but rather in the revision process which allowed parallel modification of a common data object. To avoid the confusion the secretary could have requested that each author confine himself to a different section of the document during a given review cycle, which would ensure that the various sets of modifications would not seriously conflict. While not a perfect solution, some restraint on the types of modification permitted was necessary to allow SIGMA to preserve the basic intent of the modifying users.

AUGMENTING THE CLASSICAL SOLUTIONS

In light of the previous observations, the approach taken to provide parallel modification comprised two main components. The first was to carefully constrain the types of modifications permitted to the several users so they would cause neither irreconcilable conflicts nor unexpected results. The second involved finding a representation by which the modifications could be expressed.

Limiting Modification

Determining in which ways to limit modification was a delicate issue. If the limitations were not strict enough or not along the correct dimension, the several sets of changes would conflict too much; if too restrictive, the users would be prohibited from expressing the changes they wished (and should be allowed) to make. The data objects involved, messages and folders, and the operations supported on them were thus designed with the goal of making the necessary limitations seem natural rather than confining.

Limitations Applied to Draft Messages

In draft messages, SIGMA imposes its limitations on parallel modification by slightly constraining the draft/revision model. Rather than portray a message as a shared entity, with all reviewers attempting to edit the one version into the form they want it, SIGMA effectively gives each user his "own" rendition of the message, to modify as he wishes. The process of successively refining a draft involves the selective reading and inclusion of desired segments of various users' messagettes, reading and possibly taking action on comments and suggestions, eventually resulting in a new draft. While the assimilation of the various changes and suggestions would normally be expected to be performed by the original author, any reviewer would have access to the same information and tools to create his own rendition.

The approach of giving each user his own rendition of a draft message avoids the complications described earlier.

1. Messagettes are logically and physically separate. Since each user may write only into his own messagette, modifications specified by several users contain no conflicts to produce logical inconsistencies.*

2. Each user works with his own rendition of the message. While he has access to other users' messagettes, from which he can incorporate desired sections, his message always accurately reflects only changes he has made.

Users are thus allowed a free hand in composing or revising draft messages, while still retaining the ability to read, reference and comment upon other users' versions.

Limitations Applied to Transmitted Messages

As previously described in the analogy to conventional mail, the content of transmitted messages is not subject to arbitrary modification. Users are restricted to a small set of operations.

- Comment—A user may place comments on any desired part of a transmitted message.

*Subject to additional considerations discussed later.

Maintaining Data Order and Consistency in a Multi-Access Environment

- Forward—A user may specify the message be forwarded to one or more other users. A notation of the users receiving such forwarded copies is appended to certain message fields.

The limited scope of these operations avoids the undesirable complications.

1. All the types of modifications that users can specify are non-conflicting.

- Inserting a new comment is strictly additive, requiring no change to the message's basic contents. Multiple comments specified in the same place in the message are simply added one after the other.
- Editing an existing comment causes no conflict, as each user may modify only his own comments.
- Forwarding is also an additive operation, simply appending the names of the forwarded addressees to the end of the appropriate field.

2. With the possible exception of the order in which comments or forwarding entries appear in the message (since users may comment or forward in parallel), the updated transmitted message conforms to each user's expectations.

Limitations Applied to Folders

The structure and use of folders prohibit providing a copy for each user, as in draft messages. The replication of storage to keep the multiple copies would be too expensive (folders tend to become quite large), as would the processing necessary to add each new entry to all of the copies. Consequently, all users access and modify the same folder object. Clearly, not all users may be allowed to modify a folder arbitrarily. As shown earlier, such a situation would lead to chaos unless some limitations on modification are imposed.

Fortunately, a compromise was found which provided the appropriate limitations without unduly constraining the capabilities of the users. Since it was logical to permit the owner of a folder more latitude in modifying it than other users, the limitations imposed differed, as follows:

- Owner—The owner is allowed all possible modification capabilities, including the abilities to

  —Delete the folder entirely
  —Delete and modify entries
  —Add keywords to entries
  —Append entries to the end of the folder
  —Add or modify comments (his own)
  —Keep a place-marker in the folder

- Non-owner—Non-owning users are permitted only the latter three capabilities. Note that these are all basically append-like operations, causing no structural changes to the folder.

This two-level capability scheme avoids the undesired complications described earlier.

1. The owning user alone can specify "dangerous" (structure-modifying) changes, so conflicts cannot occur. As in messages, appending data to the end or modifying data pertaining only to a specific user (comments, place-markers) do not produce conflicts.

2. The only departures from users' expectations occur when entries are deleted by the owner: comments specified by other users for deleted entries simply disappear; a place-marker specifying a deleted entry is adjusted to the nearest entry remaining. In all such cases the behavior is reasonable and does not constitute a significant departure from users' expectations.

Representing and Applying Parallel Modifications

Once the appropriate limitations had been established to eliminate textual inconsistencies between the modifying users, there remained the issue of providing a mechanism which allowed changes to be assimilated into a common data object regardless of the number of users reading or updating the same object in parallel. It was also considered desirable to preserve a consistent image for readers during their access of an object (the goal in the Readers and Writers Problem); changes performed by other users should not cause objects to "change underneath them."

To provide the consistent image of an object during a user's session, a temporary copy of the object is made upon access. The reading and modification are then performed upon the copy, ensuring that no unexpected changes occur during the session. However, when a user's modifications have been completed, the net effect of the specified changes must then be performed on the central copy (called the base copy), which constitutes the "real" object to the rest of the users. Note that the modified copy cannot in general simply be substituted for the base copy; if any other users were making changes in parallel, such a strategy would cause all sets of changes but the last to be lost. Rather, the mechanism developed for SIGMA centered upon a construct called a Δ-file.

The Δ-file Concept

When referring to a change in a variable (say χ), mathematicians often express it as $\Delta\chi$, meaning "the change in χ." To obtain the new, updated value

Maintaining Data Order and Consistency in a Multi-Access Environment

of χ (call it χ_{new}), one must take the old value (χ_{old}) and apply the change ($\Delta\chi$), which in mathematics is done by addition, i.e.,

$$\chi_{new}=\chi_{old}+\Delta\chi$$

The notion of Δ-files in SIGMA is conceptually similar to the mathematical Δ. Each user, when accessing an object, is given a local copy. As he modifies it, his changes are remembered. When his changes are complete, SIGMA places into the Δ-file a record of the effective changes applied by the user to the object. A Δ-file thus represents the distillation of changes made by a user to his local copy of an object in an editing session, which, if "added" to the base copy, would produce the changes specified by the user.

The analogy to the mathematical Δ is complicated by the possibility of having several Δ-files produced in parallel by different users, each referring to the same base copy. Since the users work independently, the order in which the Δ-files are applied cannot be guaranteed. If one user makes changes which conflict with those of another user, the consistency of the base file and maintenance of users' intentions cannot be guaranteed. But, as previously described, the types of modification allowed cause no significant conflicts.

What's in a Δ-file?

When a user modifies an object, the Δ-file produced contains not the new contents of the object but rather a specification of the modifications that need to be performed on the base (original) copy to make it conform to the user's changes. The following types of change specifications used in SIGMA Δ-files, called Δ-operations, are sufficient to describe all possible changes to SIGMA objects.

- Add—Add a new item of data to the object adjacent to some other data item.
- Delete—Delete a data item from the object.
- Replace—Replace the contents of a data item with a new value.

For efficiency in applying the changes to the base copy, the identification of the data items to be affected by the Δ-operations is done by an absolute, rather than symbolic, addressing scheme. This avoids costly searching to locate affected items, but requires that the addresses in the base copy at the time the Δ-operations are performed match those which existed at the time they were generated. The constancy of these absolute address references is guaranteed by the non-conflicting nature of Δ-files and the internal structures of messages and folders, which do not require address manipulation in response to modifications.

Mechanics of Assimilating Δ-files

Once the various Δ-files have been generated, the task remains of applying them to the base copy. Rather than assign this task to the SIGMA user processes, it was decided to create separate processes (one for messages, one for folders) to execute this assimilation function, implemented in the SIGMA system as shared background processes known as daemons. When a SIGMA process creates a Δ-file as the result of user changes, it enqueues a request to the appropriate daemon, supplying the name of the object to be modified and the name of the Δ-file in the form of a physical location identified (PLID), an operating-system-dependent path name describing the location of the Δ-file information. The daemons execute these requests in order, finding the referenced PLIDs and applying the contained Δ-operations to the named objects. A diagram depicting the process of Δ-file incorporation is shown in Figure 1 (although the SIGMA message service processes many shared objects, the figure concentrates on just one such object, in the process of being modified in parallel by several users). Note that the order in which Δ-files are processed is arbitrary, but the consistency of the resulting updated object is preserved by the non-conflicting nature of the changes they contain.

The division of labor in the updating task between the SIGMA processes and the daemon provides several benefits. It allows the Δ-file incorporation task to be performed by a separate background processor, insulating the SIGMA processes from a significant processing burden (hence user-perceived response time delay) which they would otherwise encounter whenever an update operation was performed. And since the SIGMA processes have no need to modify the base copy of an object (all writing is done in the Δ-files instead), the daemon can be given exclusive write access to the base copy. Combined with the "copy-on-access" discipline imposed to maintain a consistent view of objects for users, no costly interlock is necessary to prevent access conflicts between the daemon and the SIGMA processes.

Flaw in the Non-Conflict Assumption

Despite the careful considerations described earlier regarding limitations on modification, a significant flaw developed, although it was not specifically connected to parallel access. Rather, it developed as a side-effect of the division of labor between the generation of changes (by users through their SIGMA processes) and their application to the base copies

Maintaining Data Order and Consistency in a Multi-Access Environment

(performed by the daemon). Since the daemon is an asynchronously operating request-driven process, there is typically a measurable time interval between the generation of a Δ-file and its assimilation into the base copy. During this interval, whose duration depends upon system load and daemon backlog, a user's pending changes are not reflected in the base copy. If the user accessed the object during this period, he would find that the object did not contain his changes, violating the requirement to preserve user expectations. Moreover, if the user were allowed to access the old copy anyway, and made changes which conflicted with those of the pending Δ-file, the assumption of non-interference of Δ-files would also be violated.

The inescapable conclusion was that a user could not be allowed to access the old (unmodified) copy of the object; he could only access the object with his previous changes incorporated. The following approaches were considered:

1. The user's SIGMA process could keep track of the pending Δ-file and note whether its assimilation had occurred. If not, the SIGMA process could perform the assimilation itself to recreate the modified object. This approach required significant bookkeeping and additional computing in the SIGMA process to duplicate that soon to occur in the daemon. Also, it would not be able to account for changes made by other users.

2. The SIGMA process could prevent the user from proceeding further in his terminal session until the assimilation of the Δ-file were complete. This would introduce an unnecessarily servere degradation in user-perceived response time at the conclusion of editing of each object.

3. The SIGMA process could prevent a modifying user from accessing the object again until the previous changes had been assimilated. While this approach could temporarily prevent a user from accessing a particular object, it would not inhibit him from executing SIGMA operations not pertaining to that object, unlike (2) does.

Since in practice it is rare that a user attempts to re-access an object before the daemon has had the opportunity to perform the Δ-file assimilation, Approach 3 was chosen. While not an elegant solution, the infrequent temporary denial of access to a specific

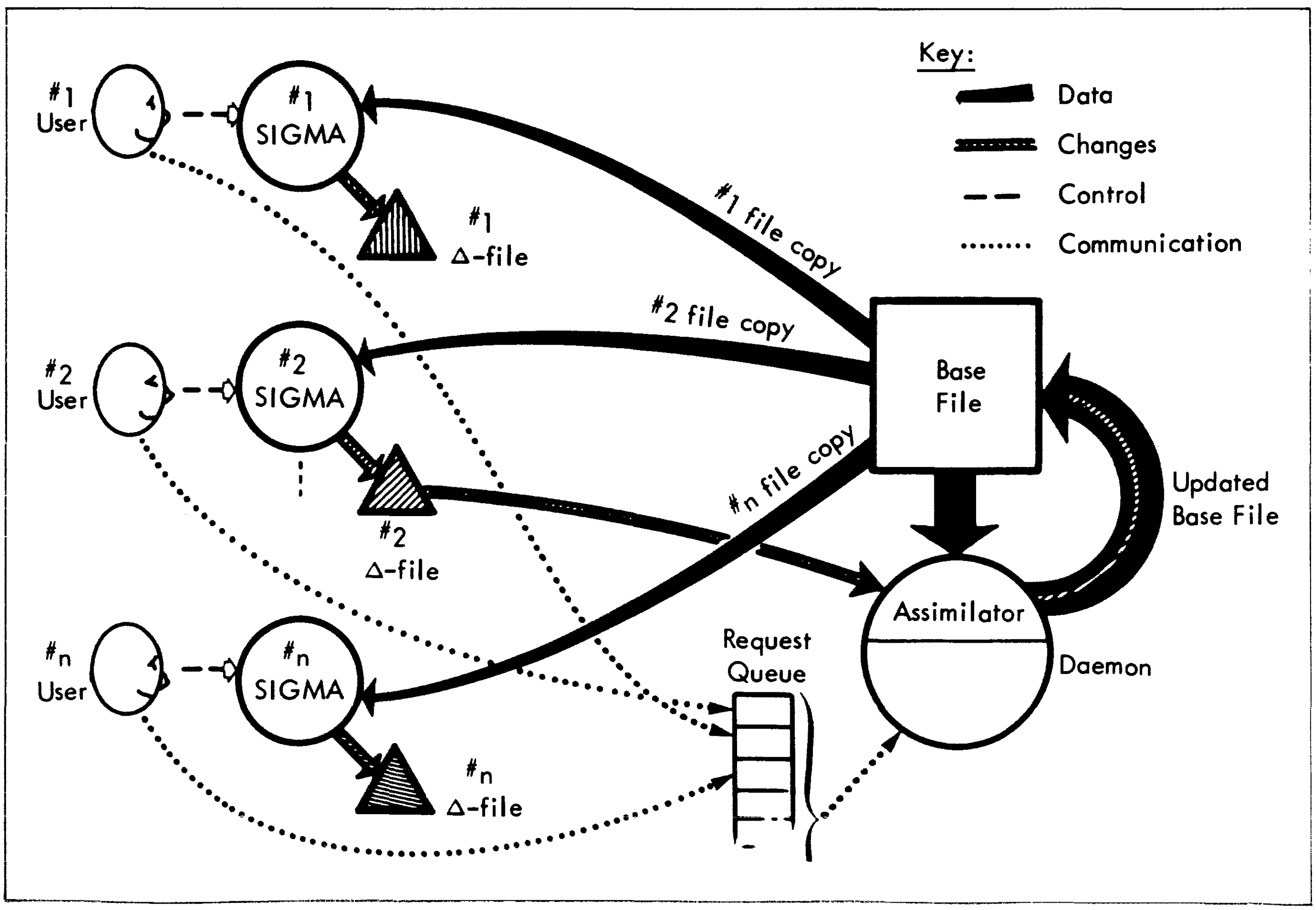


Figure 1. The Δ-file mechanism

Maintaining Data Order and Consistency in a Multi-Access Environment

object poses a negligible inconvenience and is thus a minimal impediment to users.

CONCLUSIONS

The multi-access scheme described in this report has been in operation within the SIGMA message service for over three years. This experience has shown the approach to have successfully satisfied the requirements needed to provide multi-access to SIGMA's shared data objects. Following are several of the most successful aspects of the SIGMA multi-access methodology:

- Parallel access to shared objects by an arbitrary number of users (processes) occurs with no conflict.
- Each user is satisfied that his changes to a shared object are faithfully recorded.
- Users accessing shared objects are not confused by other users' changes during their own editing sessions.
- The limitations imposed on modifications are natural rather than cumbersome, and do not overly constrain users.
- The concept of a Δ-file and the resulting nonconflicting, incremental update of shared objects is a powerful technique to apply to the multi-access problem.
- The division of labor between the foreground (SIGMA) and background (daemon) processes provides two significant benefits—the ability to avoid expensive reader/writer interlocks on shared objects, and the shifting of the processing burden away from the user process to achieve better user-perceived response.

While this methodology was developed specifically for the SIGMA message service application, the concepts involved are much more general. Similar techniques could be successfully applied to many other shared data applications, such as command-and-control, data base management systems, information retrieval systems, or any other application in which many users can have access to common data and in which preserving the intent (and hence the trust) of those users is an important concern.

REFERENCES

[1]Brinch Hansen, P., "A Comparison of Two Synchronizing Principles," *Acta Informatica,* Vol. 1, 1972, pp. 190-199.
[2]Courtois, P.J., F. Heymans and D.L. Parnas, "Concurrent Control with 'Readers' and 'Writers'," *Comm. ACM,* Vol. 14, No. 10, October 1971, pp. 667-668.
[3]Dijkstra, E.W., "Cooperating Sequential Processes," in *Programming Languages,* F. Genuys (ed.), Academic Press, 1968, pp. 43-112.
[4]Dijkstra, E.W., "Solution of a Problem in Concurrent Programming Control," *Comm. ACM,* Vol. 8, No. 9, September 1965, p. 569.
[5]Hoare, C.A.R., "Monitors: An Operating System Structuring Concept," *Comm. ACM,* Vol. 17, No. 10, October 1974, pp. 549-557.
[6]Saltzer, J.H., "Traffic Control in a Multiplexed Computer System," MAC-TR-30, Project MAC, Massachusetts Institute of Technology, 1966.
[7]Stotz, R., R. Tugender, D. Wilczynski and D. Oestreicher, "SIGMA—An Interactive Message Service for the Military Message Experiment," *Proceedings of the National Computer Conference,* AFIPS, May, 1979.
[8]Tugender, R., and D.R. Oestreicher, "Basic Functional Capabilities for a Military Message Processing Service," ISI/RR-74-23, USC/Information Sciences Institute, 1975.

This research was performed for the Advanced Research Projects Agency under Contract No. DAHC 15 72 C 0308, ARPA Order No. 2223. The views and conclusions expressed in this paper are not necessarily those of any person or organization except the author(s).☐

Planning for Electronic Mail

Problem:

To the old post office motto, "Neither rain, nor sleet, nor snow . . .," we will soon need to add ". . . nor static . . ." as the electronic mail application of communications systems rumbles on toward reality. The expression "electronic mail" is an excellent example of a category of modern expressions that we have perversely dubbed technological "glibbery" (new word, don't look it up). To qualify, an expression must roll trippingly off the tongue and glibly imply an underpinning of all sorts of technical marvels—as in truth there usually are, but the concept of electronic mail, stripped of its glibbery, amounts to an almost absurdly simple modification of one of the most ancient forms of point-to-point communication—the letter. The simple modification is that the letter is not transported physically from the sender to the receiver. Its electronic image is the only thing that crosses the space between sender and receiver. The modification is totally transparent to the sender and receiver, except that if the receiver checks the letter's postmark and then glances at his watch, he may casually note that the letter was sent about two minutes ago! That's the essence of electronic mail—an astounding compression of the letter-carrying function from days (sometimes weeks) to minutes.

Of course, there's an awful lot of electronic gimcrackery involved—packet switching, store-and-forward, facsimile, CRTs with hard copy options, and so on almost interminably—all of which supports or somehow augments the illusion of a transparent, instantaneous connection between sender and receiver. As a user, you will reap the benefits of transparency and the "instant letter." As a planner/designer, you must get to know the techniques that will make the letter-writing users oblivious to the intervening electronic intruder.

Solution:

Many telecommunication paths are now being built as digital links rather than as analog links. The wire-pair lines that fill the cities and suburbs and that stretch along country highways can carry more telephone calls if they are digital than if they are analog. Digital technology is spreading to the bigger telecommunications highways such as coaxial cable systems and the new high-capacity waveguide systems.

From *Future Developments in Telecommunications,* 2nd Edition, by James Martin. Chpt. 15, pp 259-269. © 1977 by Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Planning for Electronic Mail

Satellites, the most promising of new communication technologies, can also transmit more telephone calls if used in a digital fashion, and the even more promising technology of optical fibers offers a literally unlimited capacity to handle all the traffic we could ever produce. A main trend throughout telecommunications will be the swing to digital techniques.

When telephone calls are carried in digital form, the relative cost of data transmission and telephone transmission swings in favor of data. Data is transmitted on terrestrial telephone lines at between 1200 and 9600 bits per second. Telephone calls are transmitted using 64,000 bits per second (an AT&T and CCITT standard). In space, 60 million bits per second of existing satellite modems represents a vast amount of data traffic.

Furthermore, telephone traffic has to be transmitted in real time, i.e., when a person speaks, his speech must be transmitted almost immediately. To achieve a good grade of service, i.e., low probability of a caller encountering a busy signal, there have to be idle channels ready for immediate use. No real-time systems achieve 100% utilization of their facilities. Probability calculations determine how much idle capacity is needed to provide a given grade of service.

The transmission capacity must be designed for the peak telephone traffic. The traffic during the peak hour of the day is several times higher than the average traffic. The traffic during the peak hour of the peak day is substantially higher than that of the average day. The unshaded part of Figure 1 represents idle capacity. On a typical corporate network, the channels are idle 80% of the total time. Such is the nature of telephone traffic.

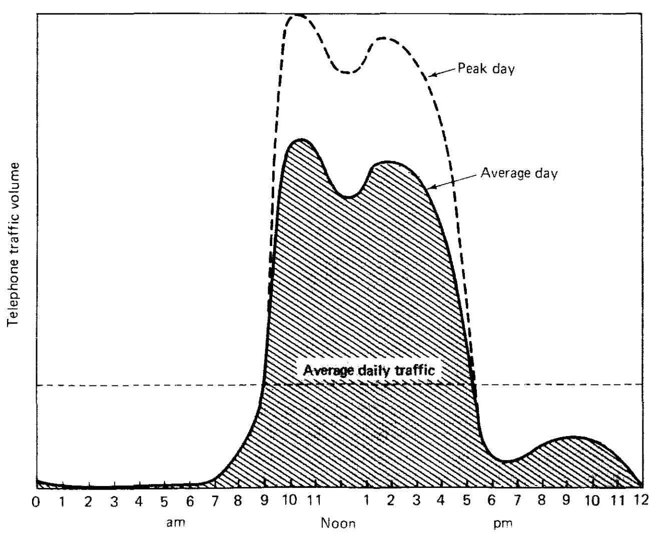


Figure 1. Telephone traffic. The unshaded part of the chart represents idle channel capacity

Mail can be sent over telecommunication networks. We have already reached the time when it is cheaper in certain circumstances to send mail electronically than to send it by conventional methods. The cost of electronic mail will drop substantially, whereas the costs of typing, addressing, delivering, receiving, opening, distributing, and filing paper mail are rising. Mail delivery is not "real-time." We are happy if it is delivered an hour, or a day, later. We write letters, leave messages, send telegrams, order catalogues, transmit batches of computer data, and request books from libraries. This information transmission has two important characteristics. First, it can wait until channels are not occupied with telephone or other real-time traffic. Second, it can be interrupted in the middle of transmission provided that the interruption is done in such a way that no information is lost.

Approximately 70% of all first class mail in the USA is originated by computer. Most of this, invoices, orders, receipts, payments, etc., is destined to be fed into another computer, often in another firm. It should be transmitted directly in alphanumeric form. Instead it is usually printed, bursted, fed in envelopes, sent to a mail room, stamped, sorted, delivered to Post Office, sorted again, delivered to the destination Post Office, sorted again, delivered to a corporation, handled in the mail room, opened, and laboriously keyed into a medium that the receiving computer can read. All this, when a packet sent on a value-added network costs a small fraction of one cent.

The most efficient way to utilize communication channels is to organize them so that real-time and nonreal-time traffic can be intermixed, and so that real-time traffic has absolute priority over nonreal-time traffic. Nonreal-time traffic should never delay the real-time traffic for more than a small fraction of a second—50 milliseconds, say. Most telephone networks today do not carry interruptable nonreal-time traffic, and consequently 75% to 85% of their total daily capacity is unutilized. When digital equipment is used for multiplexing voice signals, a variation of that equipment can accommodate the intermixing of real-time and nonreal-time traffic. If the mechanism exists for doing it, a remarkably large capacity for mail and other nonreal-time traffic is available. It is worth looking closely at society's nonreal-time uses of information to see how the unused transmission capacity could be employed.

DIGITIZED MESSAGES

It is possible to convert any type of message into a digital form for transmission. Different messages require different numbers of bits. Figure 2 gives some examples of approximate message lengths when converted to bits and compressed ready for transmission or storage. The figures are conservative in that a

Planning for Electronic Mail

smaller number of bits could be used in some of the items if complex encoding techniques are employed. Telephone speech, for example, has been economically encoded into far fewer than 32,000 bits per second with a level of distortion acceptable for one-way messages. A vocoder is a device that synthesizes speech from encoded parameters describing the speech-production mechanism and the sounds it is making. Vocoder telephone speech, can be designed to make the spoken words intelligible but not necessarily make the speaker's voice recognizable. Using vocoder techniques, spoken messages can be transmitted in one tenth of the number of bits of speech using PCM or delta modulation.

Much more compaction can be achieved with still less natural sounding speech if the messages are composed of prerecorded words. A prerecorded vocabulary might have up to a thousand words, each addressed with 10 bits, thus permitting a 30-word message to be sent with 300 bits plus addressing and error-detection bits. After transmission, the 300 bits would trigger a spoken message from a voice response unit (many of which are in use today). A code book of words commonly used in business was employed for a similar purpose in the early days of telegraphy.

Varying degrees of compaction can be achieved with other types of message delivery. There is a trade-off between message length and encoding complexity.

A digital transmission system with the high bit-rate of today's PCM telephone links can be given the capability to transmit any of the message types in Figure 2; the nonreal-time messages fit into gaps between the real-time traffic. The signals are interweaved in the bit stream with whatever is the most appropriate form of multiplexing.

PRIORITIES

It is necessary to have some form of priority structure in the system. At its simplest, there could be two priorities: real-time and nonreal-time. There are, however, different degrees of urgency in the nonreal-time traffic, so several priority levels may be used to help ensure a fast delivery of messages. The system organization may be designed to permit the following categories of end-to-end delivery time:

1. Almost immediate (as with telephone speech).

2. A few seconds (as with interactive use of computers).

3. Several minutes.

4. Several hours.

5. Delivery the following morning.

Message type	Bits
1. A high-quality color photograph	2 million
2. A newspaper-quality photograph	100,000
3. A color television frame	1 million
4. A picturephone frame	100,000
5. A brief telephone voice message (voicegram)	1 million
6. A vocoder telephone voice message	100,000
7. A voice message of codebook words	400
8. A document page in facsimile form	200,000
9. A document page in computer code	10,000
10. A typical inter-office memo	3000
11. A typical flip chart	1000
12. A typical computer input transaction	500
13. A typical electronic fund transfer	500
14. A typical telegram	400
15. A typical airline reservation	200
16. A coded request for library document	200
17. A fire or burglar alarm signal	40

Figure 2. Numbers of bits needed for different message types

When more than one message is waiting for transmission at any point, the higher priority messages will be sent first. The fact that much of the traffic is not in the highest priority category will make it possible to achieve a substantially higher line utilization than on a network which guarantees real-time transmission for all messages. In addition, the facilities will be well utilized at night, when on other systems they would be largely idle.

TRAFFIC GROWTH

If transmission systems come into existence for transmitting nonreal-time information at a low cost, there can be a major growth of such traffic, and the growth will probably incorporate traffic which is not sent electronically today.

Table 1 is a forecast from a NASA study showing demand trends of data record transmission. In addition to data records, there are other types of traffic of very high volume, such as mail.

		1950	*1960*	*1970*	*1980*	*1990*
Stolen vehicle information transfer	cases/yr × 10^3	160	320	820	1950	4600
Facsimile transmission of "mug shots," fingerprints, and court records	cases/yr × 10^6	2	4	7	13	25
Stolen property information transfer	cases/yr × 10^3	430	880	1700	3500	7000
Motor vehicle registration	items/yr × 10^6	49	74	110	164	245
Driver's license renewal	items/yr × 10^6	38	48	60	75	90
Remote library browsing	accesses/yr × 10^6	0	0	low	5	20
Remote title and abstract searches	searches/yr × 10^6	0	0	low	8	20
Interlibrary loans	books/yr × 10^6	...	...	low	40	100
Remote medical diagnosis	cases/yr × 10^6	0	0	20	60	200
Remote medical browsing	accesses/yr × 10^6	0	0	20	60	200
Electrocardiogram analysis	cases/yr × 10^6	0	low	20	60	200
Patent searches	searches/yr × 10^6	6	6	6.5	7	7
Checks and credit transactions	trans/yr × 10^9	11	25	56	135	340
Stock exchange quotations	trans/yr × 10^9	0	0	1	2	4
Stock transfers	trans/yr × 10^6	290	580	1200	2500	4900
Airline reservations	pass/yr × 10^6	19	62	193	500	1400
Auto rental reservations	reserv/yr × 10^6	0	low	10	20	40
Hotel/motel reservations	reserv/yr × 10^6	...	...	25	50	100
Entertainment reservations	reserv/yr × 10^6	...	...	100	140	200
National Crime Information Center	trans/yr × 10^6	0	0	6	20	70
National legal information center	trans/yr × 10^6	0	0	low	5	30

Table 1. Demand trends for transmission of records (from a study of satellite uses commissioned by NASA[1])

Planning for Electronic Mail

Type of Mail		Percentage	
Individual households to:			
Business		5.8	
Individual households		14.0	
Government		0.4	
	TOTAL		20.2
Government to:			
Business*		1.8	
Individual households**		3.8	
Government*		0.6	
	TOTAL		6.2
Business to business:			
To suppliers*		3.9	
Intracompany*		1.4	
To stockholders*		0.7	
To customers:	order acknowledgement*	0.2	
	bills*	6.7	
	product distribution	1.3	
	promotional materials	5.4	
Other*		6.2	
TOTAL		25.8	
Business to households:			
Letters:			
Bills*		10.1	
Transactions**		1.2	
Advertising		12.6	
Other**		4.5	
TOTAL LETTERS		28.5	
Postcards:			
Bills**		0.7	
Advertising**		2.1	
Other**		0.4	
TOTAL POSTCARDS		3.2	
Newspapers and magazines		13.6	
Parcels		1.3	
TOTAL BUSINESS TO HOUSEHOLDS		46.7	
Business to government:*		1.2	
	TOTAL BUSINESS		73.6

*Potentially deliverable by telecommunications to the end user (22.7%).
**Potentially deliverable by telecommunications, sorted, to a post office (22.8%).

Figure 3. The composition of the U.S. mail

ELECTRONIC MAIL

The total cost of mail delivery is gigantic, especially in North America. Americans are not only the most communicative people by telephone; they also receive the most mail. More mail is sent in New York City than in the whole of Russia.

Some types of mail could be sent and delivered by electronic means, and where the volumes are high, this could be done at a fraction of the cost of manual delivery. To send a handwritten letter electronically, it is fed into a facsimile machine, transmitted, and received by another facsimile machine, which produces a copy. Most of today's facsimile machines transmit an analog signal over telephone lines; but they can be designed to transmit a digital signal, and digital facsimile machines are now in use.

Figure 3 breaks down the U.S. mail by type. The asterisks indicate which mail could be sent by electronic means, and hence potentially by digital channels. A single asterisk refers to mail that could be delivered electronically to the end user. It is assumed in the table that individual households can neither send nor receive electronic mail. They have neither the equipment nor the desire to change their mail-sending habits. At some time in the future, electronic mail will reach into consumers' homes, but we will make the conservative assumption for this report that for the time being only businesses and government will use it. When government and businesses send mail to households, this mail could be delivered to the local post offices already sorted for delivery. All local post offices could have a receive-only satellite antenna on the roof (like the Musak antenna) and a high-speed facsimile printer. Advertising letters and promotional materials have not been included as potential electronic mail because they may contain glossy or high quality reproductions. Some advertising letters could be sent by facsimile machines. Newspapers and magazines have not been included although there has been much discussion of customized news sheets being electronically delivered to homes.

On this basis, 22.7% of all mail is potentially delivered to end users by telecommunications, and a further 22.8% is potentially deliverable to post offices. In 1980 this will be a total of about 50 billion pieces of mail per year. The 50 billion pieces of digitizable mail would require on average approximately 200,000 bits each to encode; some would need more than this; many would require less because alphanumeric/encoding rather than facsimile would be used. The annual total would be roughly 50 billion x 200,000 = 10^{15} bits.

CONTROL MECHANISM

It is possible (some would say almost certain) that the control mechanisms necessary to interleave telephone and nonreal-time traffic will not be built as part of the traditional common carrier systems. We are, however, embarking on the construction of new types of networks—value-added networks, satellite systems, government and defense networks, and private industry networks. The designers of the networks, seeking to maximize the utilization of their facilities, are likely to intermix real-time and nonreal-time traffic in a way traditional telephone systems do not.

Particularly interesting are the satellite networks now being designed and discussed. A multiple-access satel-

Planning for Electronic Mail

lite, designed to transmit telephone and other traffic between many earth stations, can operate most efficiently in a digital fashion. With today's equipment, one voice call typically requires 32,000 bits per second. If a satellite system were designed to carry a peak traffic of 100,000 voice calls, the total capacity of the satellite would be 100,000 x 32,000 x 60 x 24 x 365 = 1.68×10^{15} bits per year. If 80% of this capacity were unutilized by telephone traffic because of the uneven traffic distribution shown in Figure 1, as would probably be the case, then 1.35×10^{15} bits per year would be available for nonreal-time traffic.

In other words such a satellite system could carry all of the above 1980 digitizable mail, the required channel capacity being an unused byproduct of a telephone satellite. However, if only 1% of such mail were transmitted, it would pay for a large satellite system. The practicality of satellite mail is aided by the fact that three quarters of all U.S. mail originates in only 75 cities, and only 20.2% of all mail originates from individuals—the rest is from businesses and the government.

ELECTRONIC FUNDS TRANSFER

Nonreal-time digital transmission capacity will be required in the future for the transfer of funds between banks and other institutions. The Federal Reserve Board has made it clear that electronic funds transfer is essential for America, if only to halt the growing burden of paperwork such as check processing.

About 30 billion checks per year are written in the United States representing $20 trillion per year. An electronic fund transfer network could speed up the clearing time for checks by at least one day on average; probably more. This represents a float of

$\frac{\$20 \text{ trillion}}{365}$ = $54.8 billion, savable by electronic check transfer.

At 8% interest this gives a saving of $4.38 billion.

If one check requires 500 bits of transmission, the total capacity needed is 30 billion x 500 = 1.5×10^{13} bits per year—a little more than 1% of the spare capacity in a 100,000-voice-channel satellite.

The number of credit transactions is almost double the number of checks, and the payment delay with these is much longer. In the electronic-funds-transfer society, when a customer makes a payment in a store or restaurant with a machine-readable card, a transaction would travel to a bank and a response would be received from the bank computer. If all credit card transactions in the United States in 1980 were handled this way, 200 billion such messages would be needed, or about 10^{14} bits per year. In fact such transactions will still be a small proportion of the total by 1980 and most will be to local banks on local telephone loops, but the quantity of long distance credit transactions will be growing rapidly.

FASTER-THAN-MAIL TRAFFIC

Today, message delivery that is faster than mail costs substantially more than mail. As the cost of long distance telephone calls has dropped, the number of telegrams sent has steadily declined. In many cases today it is cheaper to telephone than to send a telegram.

With digital transmission links, however, the relative costs of transmitting telegrams and telephone calls change. A typical telephone call lasts four minutes and with telephone company PCM requires about 15 million bits transmitted in both directions. A typical telegram requires 2000 bits. Furthermore, the telegram can wait to be fitted into gaps on the real-time traffic that would otherwise be unused.

Even with such a dramatic cost difference, most corporate communications users would probably still use the telephone because of its convenience, friendliness, and the cheapness of the instrument. If computerized corporate telephone exchanges (PABXs) place controls on user's communication expenditures, then digital transmission may bring new life to telegraphy.

TELEPHONE MESSAGES

The total number of telephone calls in North America is far higher than the total number of written messages. AT&T alone plans a capital expenditure over the next ten years more than 20 times higher than the likely capital expenditure in the U.S. Postal Service. Telephone callers are often greeted with busy signals or no answer, and many of these callers would leave a brief message if they could. Thirty-two long distance calls out of 100 are not completed today[2], because of busy signals, no answers, or equipment failures. Of the business calls that are completed, on only 35 per cent does the caller reach the called party. It is estimated that this wastes 200,000 man-years of callers' time, which at $10,000 per year is equivalent to $2 billion.[2]

A one-way telephone message could be digitized and stored so that the person it is sent to can retrieve it at his convenience. It could be transmitted and stored in any of the three forms mentioned earlier:

1. True-to-life:	32,000 bits per second
2. Vocoder:	2,400 bits per second
3. Words from a prerecorded vocabulary:	10 bits per word

Planning for Electronic Mail

Such a service could be designed so that subscribers could leave either a coded telephone message from a list of such messages, dialed on a conventional telephone, or they could leave a brief spoken message. The system would ring the called party periodically until it could speak the message. The called party could use his telephone dial to ask for repetition of the message, give confirmation of its receipt, or dial a response. The system may be designed so that a user can dial his stored message queue from any telephone, key in a security code, and have the messages spoken to him.

If 10% of all business telephone callers left such a message when they failed to complete their call, this would amount to approximately 100 million messages per year. If 10% of all business callers who failed to contact the individual they telephoned also left such a message, that would be amount to about another billion messages per year.

Telephone messages could be sent for which simple responses are required. The receiver would dial the response on his telephone after a local computer speaks to him, and the computer would receive the response and deliver it. The voice answerback unit would inform the called party what form of response was expected.

Organizations could send bulk messages in this way, in which one telephone message is sent to many individuals. Unsolicited messages could be composed by computers for verbal delivery, possibly expecting a response. One can imagine such a system being programmed to carry out opinion polls or to gather statistics from individuals.

Whatever the applications or the form of the messages, it seems clear that future satellite systems or networks for corporate or government use should be designed to intermix all types of traffic, real-time and nonreal-time, on digital channels. There are many uses for telecommunications other than traditional telephony and telegraphy.

REFERENCE

[1]Roger W. Hough, *Information Transfer in 1990,* AIAA 3rd Communications Satellite Systems Conference, Los Angeles 1970, presenting work supported by NASA Contract NAS2-5369.
[2]Harry Newton, *Communications Lines,* Business Communications Review, Sept., 1976. □

Facsimile—How to Analyze Its Role in Your Organization

Problem:

EDP managers, confronted with the need to provide overall solutions to the problem of information transfer between geographically dispersed locations in a large organization, or to points external to the organization, have many alternatives to consider. Among these are electronic mail systems implemented via message terminals, direct on-line control of systems/processors from a remote location, and facsimile. Each of these electronic information transfer systems can be connected with a variety of dial-up or dedicated communications lines and assorted terminal equipment.

One of the least understood and possibly most useful methods of communication for certain types of information is facsimile. This report will explain the fundamentals of this technology and offer a methodology for analyzing its role in your organization.

Solution:

Facsimile is the transmission and reproduction of an image over a distance by electronic means. Facsimile machines accept input originals (which may be documents, photographs, microimages, or other forms of graphic information) at a scanner or other appropriate transmitting terminal, and translate the image into electronic signals. These signals are sent over communications facilities to a receiving terminal or printer where a facsimile or reproduction of the input original (possibly at a different linear scale) is created.

BACKGROUND

The basic idea of facsimile was conceived in Scotland and demonstrated in primitive form over 125 years ago, but for many years it attracted no commercial interest. Facsimile had to await the growth of long-distance telephone communication networks in order to become a practical possibility and begin to attract users.

The first routine application of facsimile began some 50 years ago, for news wirephoto. Pictures accompanying news stories could be transmitted by wire or radio in parallel with the stories. The news services found a broad and eager acceptance of the novel service by newspapers across the country and abroad. A second early application was found in the transmission of weather maps which are needed by meteorologists in preparing weather forecasts. This application also found and retains wide acceptance.

Both of these applications had several common denominators. The information was graphic and could not be sent by the use of printed or spoken words or symbols. Also, the information was "time-vital." Weather maps that are 3 days old are of no use to forecasters. Three-day-old news photos are of no interest after the story has left the pages of the newspaper.

By 1960, businesses had begun to experiment with facsimile to solve some of their record communications problems, even though the equipment then available was primitive and but little improved over what had been available for many years. During the 1960's, the need became relentlessly more urgent: Computers began to become important and required accurate and timely input information; the postal system began to show symptoms of a growing inability to handle efficiently the enormous volume of mail generated by modern society. Fortunately, the fields of data communications and of integrated circuits yielded technologies that would enable increased speed and flexibility for facsimile communications. By the end of the decade major advances in technology, plus favorable government regulatory changes, had created the bases that would permit facsimile to become a convenient and economical communications tool.

SYSTEM DESIGN

Today there is a considerable array of facsimile equipment. The user must understand his or her requirements, develop selection criteria, and acquire a system that complies with them. The most important considerations are the overall usage cost of the system (i.e., equipment costs and communication costs, both of which have fixed and variable elements), output copy quality, the convenience of use of the equipment, and the assurance that a transmission sent is in fact received and printed at the destination.

The user must decide what operating speed is needed, or whether several speeds are required—that is, whether top-of-the-line equipment is needed at certain offices, but less sophisticated can be used at other offices. Selection must be made among transmitters, receivers, and transceivers, or some mixture. A decision must be made on communication facilities, and terminals must be selected with the proper degree of automation for demonstrated needs. The user must select the communication network interfaces that best fulfill the company's needs, and consider which optional attachments should be ordered with the equipment. Thus the user is the final designer of his or her own system.

UNDERSTANDING THE REQUIREMENT

The burden is upon users to understand their communications needs and to develop a clear understanding of what a graphic system is worth to them, in dollars. The marketing personnel of the equipment vendors are well experienced in these analyses and are of course anxious to help and should be consulted. However, users are wise always to bear in mind that the salesperson is out to sell; such counsel must be evaluated objectively.

Only a very small fraction of potential or actual facsimile users have a good grasp of what fast graphic communication can do for them in improving the efficiency of business and government administration. In many organizations, facsimile is "saved" either for use of the top executives, or for only super-urgent communications, while the great bulk of traffic is committed to other and slower means. But companies should evaluate the cost of having receivables documents and order entries tied up in the mails for two days or more. The 1975 average delivery time for first class air mail, coast to coast, office to office, was over 4 days. In litigations, what is the value of access to distant documents in minutes as compared to days? In management and control, what is the value of having the weekly operating reports, with total accuracy, in the hands of the supervision in minutes instead of days? As to accuracy: a facsimile machine, in contrast to all other means of electronic record transmission (all requiring an operator at a keyboard at some point in the process) cannot transpose a number.

USAGE COST

The usage cost of a facsimile system is made up of the fixed charges for the machines and the communication channel terminations (both in general are monthly rentals), plus machine variable charges and consummable supplies, plus the cost of the use time on the transmission channels. Some organizations have dedicated and private communication networks, or have "unlimited-use" long-distance call contracts with the communication companies. In such instances, a serious error is sometimes made by the analyst who assumes that the variable cost of the transmission channel is free, because the channel is in being and paid for regardless of the facsimile decision. Of course, all channel users should absorb a duration-of-use proration of the transmission costs to ensure that each contending user makes an economical use decision.

The trade-off between speed (or throughput) and cost must be evaluated. Equipment is available in various speed configurations. In general, higher transmission speed ability means high monthly rental charges, coupled with lower transmission line costs. The break-

$$V = \frac{R_b - R_a}{(L_a + S_a + M_a) - (L_b + S_b + M_b)}$$

Where V = break-even transmission volume (in pages) per average unit time
R = rental per unit
L = line charges per page
S = supplies cost per page
M = meter cost per page

As an example, suppose a comparison is desired between a 2-minute metered unit and a 4-minute unit with lower monthly cost. Assume the following:

R_2 = \$200/month
L_2 = 2 minute × 38¢/minute = 76¢/page
S_2 = 2¢/page
M_2 = 10¢/page

R_4 = \$50/month
L_4 = 4 minute × 38¢/minute = \$1.52/page
S_4 = 5¢/page
M_4 = 0

The break-even volume is calculated as

$$V = \frac{200 - 50}{(1.52 + 0.05 + 0.00) - (0.76 + 0.02 + 0.10)}$$

= 217 pages/month or 10 pages/day

even monthly volume between two different units (subscripts *a* and *b*) can be calculated as shown in the accompanying boxed calculation.

This formula finds the intersection of two lines, each representing the usage cost of a machine as a function of the volume. These curves have the familiar appearance of Figure 1.

The numbers employed in our example are purely illustrative. For an actual business decision, it is imperative to consult the pertinent vendors and common carriers to ensure that the input figures employed are current and inclusive.

The following figures are representative for 1975 equipment and transmission costs for an even mix of transcontinental and midcontinental calls. The 4- to 6-minute units are least costly (all-inclusive) at a daily page volume per unit of 7. The 1-minute units (for which the rental is higher but the channel costs are less) are most cost-effective at daily page volumes of 14 or more. The mid-speed (2- or 3-minute) units are best suited to the range between those shown in Figure 2.[1]

One more comparison: For reasonable traffic levels, teletype is about five times as costly, word for word, as high-speed facsimile.

OUTPUT COPY

It is worth the cost to employ facsimile equipment with high-quality output copy (high resolution and high constrast) if any of the following are true:

1. Copies of the received copy are likely to be made for further distribution.

2. The received copy is to be marked up, corrected, signed off, or otherwise reprocessed and then transmitted again.

3. Received copy is to be of archival quality, be filed, and kept for any lengthly period of time.

4. Received copy is to be scanned by character reader devices, either of the optical or mark-sense variety.

5. Received copy is to be used (if only occasionally) as input to a duplicator.

OPERATOR CONVENIENCE

In general the internal operating logic of facsimile machines is not interesting to office personnel, and they should not be called upon to be sympathetic with the machine's problems. Therefore, success in routine operations demands that the machines be highly automated so that the operator has only to instruct the machine on the intended address, the page size, and the desired speed. Experience has shown that if the operators have to do more, they tend to fear the machine, and fear begets error.

The 3-hour time shift between the coasts of the United States means that East and West Coast offices are out of step for part of each day. It is thus desirable for some users to have the feature of completely unattended receive in their facsimile equipment. With that feature, the East Coast offices can make routine transmissions west while the address offices are still not opened up in the morning. And similarly, after 2 p.m. in the West, transmissions east remain practical, and will be on the addressees' desks the following morning while the senders are still in bed.

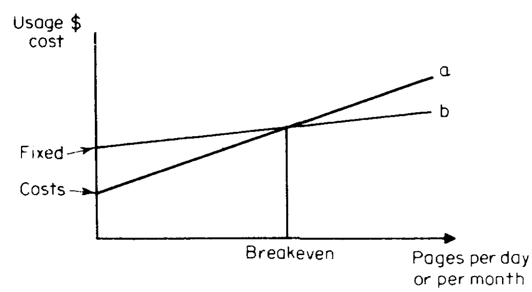


Figure 1. Facsimile usage break-even chart

ASSURED DELIVERY

An important requirement for a true unattended receive function is that the transmission be closed loop, so that the receiver automatically informs the sender whether the copy was properly received or not. If not, the copy can be sent again. Without assured delivery, the transmission may be a failure and neither the sender nor the intended receiver would know. Days (and valuable business!) could be lost. Unattended receive equipment without the feature of assured delivery should not be considered.

[1]The calculations are for fax (facsimile) transmission only (not comparing fax direct cost with mail, for example) and are for an even mix of mid-range (1,000-mile) and long-range (3,000-mile) calls. An average of 14 transmissions per day is not much for an office of any appreciable size, and hence such offices, if they treat with other offices in other cities, should logically use high-speed fax. With any alternative, that office is wasting money and time. Not many firms as yet appreciate this. Rapifax's largest customer, with more than 60 facsimile machines, does. Each of its machines averages about 1,000 transmissions per month, or about 45 per day. The average customer makes about 400 transmissions per month, or 18 per day. Experience shows that usage of equipment always rises from month to month, for each installation, as users become accustomed to the speed and confident of delivery on the other end.

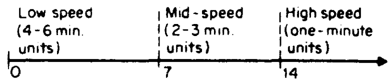


Figure 2. Comparison of facsimile units for maximum cost effectiveness over a range of usage

TRANSMIT-ONLY AND RECEIVE-ONLY UNITS VERSUS TRANSCEIVERS

With separate consoles, electronics, power supplies, and channel interfaces, it is clear that the cost of a transmit-only unit plus a receive-only unit will be greater than the cost of a transceiver. Hence, if a given facsimile transaction point is to have both transmit and receive functions, it is usually true that transceivers are the preferred equipment. There are, however, two circumstances where this simple conclusion does not hold:

1. In the circumstance of very heavy operational use (many hours per day of either transmission or reception at one point) a transceiver could be so busy sending that an outside sender could rarely get "in"—the sender would almost always get a busy signal; similarly, if a transceiver is almost always busy receiving, the local personnel would have to get into a queue to get "out." The solution to such problems is to have multiple communication lines and multiple machines, some for transmit only and some for receive only. Thus, the cost of a transmitter plus the cost of a receiver can be less than the cost of two transceivers.

2. In the circumstance of highly-skewed usage, with almost all transactions being transmit, transmit-only units are frequently preferred. The reason here is that, in general, a transmitter is less costly than a receiver with the same quality of performance. Thus, from points with high outbound traffic, and little inbound traffic, several transmit-only units plus one transceiver should perhaps be considered.

COMMUNICATION FACILITIES AND COMMUNICATION INTERFACES

In addition to the public switched telephone network (PSN), most large firms and almost all branches of government employ one or more networks of privately owned or leased and dedicated communications facilities. Until the early 1970s these networks were usually analog voice circuits set up for economy or privacy or both. More recently such networks have shown a trend to be digital, and to be intended primarily for data links among computers or between computers and terminals. Facsimile equipment can be set up to operate very well on these lines.

For organizations contemplating the routine operational use of facsimile communication, it is usual to set up the equipment for primary operation on the private network, with access to the PSN available as a backup should the private network be out of commission or overloaded. For this purpose, both the public and private lines are brought to an interface switch on the machine, enabling the operator to select the channel to be used for transmission. Normally, this interface is set up to receive an incoming call on either line, if not busy, without operator intervention, and without regard to the switch setting for outbound calls.

Standard interfaces are available which enable a facsimile transmitter to make a broadcast or, simultaneous transmission to all or any subset of the facsimile machines on a network. These arrangements prevent an unauthorized point from eavesdropping, but do not prevent a station on the network not involved in a broadcast from calling any other station not on the broadcast, to transact its own business without any interference to or from the broadcast.

Also, standard interfaces are available for facsimile hot lines. These are dedicated lines among two or more points, and available for use without operator dialing. The operator merely hands the documents to the machine, which automatically "wakes up" and immediately begins transmission to the other point or points.

Hot-line installations find use within large plant sites, among main offices of corporations, and notably in the financial institutions of New York City between midtown and Wall Street.

OPTIONAL ATTACHMENTS

The line interfaces already mentioned (digital line, broadcast, and hot line), and the unattended receive/assured delivery feature may be regarded as optional attachments to a basic facsimile equipment. They do not alter the basic performance of a unit, but rather enhance the convenience of its use. Further optional features available which the system buyer should consider are:

Unattended Send. This feature enables a machine to make a transmission at an arbitrary interval of time after it has been instructed to do so. Suppose, for example, that the field sales offices are to send in, daily, by facsimile, the sales results for main office study. Each office is assigned its time slot during the night for transmission. The document feeder (discussed below) is loaded with the day's sales report, a clock timer on the unit is set, the unattended send button on the facsimile machine is asserted, and the address (phone number) of the main-office machine is dialed in. That number goes into a temporary memory. At the appointed hour, the clock starts the transmission and it proceeds to completion.

Facsimile—How to Analyze Its Role in Your Organization

Store and Forward. The idea of unattended send is carried to its logical end point by the store and forward attachment. This unit enables a number of transmissions to a number of different addresses to be stored on magnetic tape during the course of a business day. Then, at a preselected hour, the unit initiates transmission activity to the various addressees. The unit pauses at prearranged intervals so that the parent facsimile machine can make receptions if someone (e.g., another store and forward) is attempting to call in. The unit is set up to retry calls that do not get through, and to keep a complete record of its work so that the following morning the operator is fully informed about what was and was not achieved.

Multiple-Page. This important feature enables a facsimile transmitter to hold a connection so that many sheets can be transmitted on the same call. This capability is inherent in all voice-coordinated-type machines (i.e., those requiring an operator at both ends). It is not inherent in all unattended receive machines. Machines without this feature should not be considered for any but the most occasional use.

Document Feeder. Once the benefits of graphic communication are appreciated by an organization, the process is organized for efficiency. Transmissions are grouped by addressee, and multiple pages are sent on each call. It is costly to have an operator do manual feeding, a page at a time (e.g., one per minute, or one per 2 minutes, which is just sufficiently frequent that the operator could not do something else in the meantime), and hence an automatic document feeder is usually an attractive investment. Most mid-speed and high-speed facsimile equipment offer this option.

Shorthand Dialer. Experience shows that most facsimile transmissions are made to a very few addressees. Any one machine will make 85 to 90 per cent of its calls to one of nine frequently called numbers. It is time-consuming, boring, and hence conducive to operator dialing errors for an operator to dial the same two or four numbers several times a day, day after day. The facsimile industry is now offering a shorthand dialer attachment which will enable the most frequently called numbers to be stored in memory in the machine. These numbers (including the area code and intermediate pauses for dial tones) can be called forth by the operator by dialing only two digits. The actual network control signals (dial pulses or tone signals) that go to the telephone plant are generated electronically by the shorthand dialer circuitry.

Compatibility. There are many facsimile units, from many vendors, in use throughout the world. Some of these machines will work across family lines with machines of another family, and some will not. There is growing appreciation of the proposition that compatibility between families of machines is desirable for several reasons:

1. It would permit high-speed machines to be used between high-traffic locations, and yet allow low-traffic points to communicate with high-traffic points without requiring the duplication of equipment at the latter.

2. It would facilitate lateral facsimile communication between companies using equipment from different sources.

It is, however, clearly unwise to attempt to fix and finalize the signaling format of facsimile machines, since that act would foreclose progress to more efficient encoding means, higher speeds, or other innovations. The solution lies in the compatibility options offered by several equipment manufacturers. These attachments will allow the host machine to send to and receive from machines in one or more other families of machines, in at least one of the speed modes of the other machines.

The facsimile equipment buyer should consider compatibility requirements and ensure that the compatibilities he or she requires can be obtained.

Computer Compatibility. Facsimile equipment can be used in the role of a computer printer. For this purpose there must be a character-generation function either as part of the computer program, or as part of the internal writing control logic of the facsimile printer. This function receives the character codes from the computer, and from a look-up table stored in memory, outputs the raster scan equivalents to the printer mechanism, line by line.

APPLICATIONS SUMMARY

Facsimile today is finding broad, general-purpose use in a variety of industrial, service, governmental, and professional applications.

In a message communication system, facsimile is used to communicate time-vital information of all kinds from the source point directly to the end user. In this application, facsimile offers the flexibility of maintaining the original format of the message with no possibility for a transposition or omission error by an intermediate operator, while providing the most rapid system possible at a surprisingly low cost. The features of the contemporary equipment such as quiet, simple, and odorless operation that permit its placement in the office instead of in the communication center make a facsimile system practical and desirable.

In data collection, facsimile permits the capture of operating data at the source and their transmission

Facsimile—How to Analyze Its Role in Your Organization

to the central point for study or entry to data processing. Facsimile data entry can be used in the small sales office as easily and effectively as in a major facility.

Acceptance of facsimile as a major means of communication in business and government has been progressing on a large scale since about 1970. But today, both the technology, and the full benefits of its use, are still being discovered and developed. The applications are limited only by the imaginations of the users. Equipment improvements and extensions are coming forth rapidly, and facsimile promises to provide the same widespread, convenient, instantaneous, and readily available communication for graphics as we have all come to take for granted for voice, via the telephone.

APPENDIX—FACSIMILE TECHNOLOGY

A facsimile transmission involves the following seven basic steps:

1. Input paper handling. This function brings the document, from which the image is to be transmitted, into the proper position to be scanned and then to be removed from the machine after scanning. Input can be manual or automatic. All facsimile systems use a raster (i.e., television) scan. The image is scanned along closely spaced parallel lines with the scan head reading the image density along one, two, three, or even more than three, contiguous scan lines simultaneously. The scan proceeds across successive sets of scan lines in the same direction. The same raster pattern must be created in both the scanning and the printing processes. Synchronism between scanner and printer must be maintained so that each starts a new line (or set of lines) at the same time.

There is no limit to the number of ways a raster can be created. Several of the more widely used means are rotating drum, rotating turret, and flat bed.

2. Scanning. This step is the process of translating the light and dark areas along successive scan lines into an electrical signal that represents the image. The resulting signal, called the video baseband signal, is then turned over to the next system component for further processing.

The heart of the scanner is a photosensitive element which outputs either a voltage or a current proportional to the intensity of the incident illumination. The scanner measures the surface optical density at the location of the resolution element (picture element, or pixel) being scanned. Either the document is floodlighted and the scanner optics are sharply focused on the pixel, or the scanner optics may be defocused and only the pixel illuminated by a sharp spot of light. This latter method is termed flying-spot scanning. Both methods can yield excellent results and are widely employed.

3. Encoding. The baseland signal may be encoded in a variety of ways intended to make the subsequent transmission fit into the available bandwidth of the communication channel or (in a digital system) to eliminate redundant information, thus minimizing the number of bits to be transmitted, or to make the transmitted signal less susceptible to errors, noise, and interference, or for crypto purposes. Not all facsimile machines employ encoding.

To discuss encoding it is first necessary to distinguish between synchronous and asynchronous systems. For a synchronous system, one may visualize the raster scan line to be marked off by a fine ruler into individual fixed-position resolution elements. As the scanner examines each resolution element, it must decide if that element is white or black. That decision is easy if the element is all white or all black, but is equivocal if a transition from white to black (or vice versa) takes place toward the center of the element.

Decision rules are built into the scan logic which attempt to make the decision on a consistent basis. However, such rules are imperfect, and in the scanning of curved characters or inclined lines, the scanning yields image signals equivalent to ragged or stepladder edges in the received copy. This result is called quantizing noise. To overcome it, the resolution of synchronous systems is increased to the point that the edge imperfections are too small for the unaided eye to detect. It is a thumbrule in the graphics trade that a synchronous scan requires about 30 per cent higher resolution than an asynchronous scan to yield output copy of equal apparent sharpness to the eye.

An asynchronous system does not have fixed-position resolution elements along the scan line. Instead, the scanner merely outputs a signal proportional to the average image density it sees in the resolution spot. This signal then may or may not be thresholded, or interpreted in binary digital form. If it is not, the printer is commanded to make a mark on the output copy of the same density seen at the corresponding spot by the scanner. If the signal is thresholded or digitized, the printer prints white until the threshold is exceeded, then makes a mark. An asynchronous digitized system is thus also afflicted with quantizing noise, but to a lesser degree than a synchronous system.

4. Modulation, line coupling, and transmission. The encoded or unencoded baseband signal (as the case may be) is passed through a modem (i.e., modulator-demodulator) which puts it into a signaling format appropriate to the transmission channel employed

Facsimile—How to Analyze Its Role in Your Organization

(digital or analog, two-wire or four-wire, or wireless). The signal thus reaches the receiving terminal. An important step in the transmission process is that of establishing the link to the receiving terminal and instructing that terminal to prepare for reception.

Techniques for modulation and transmission of facsimile signals are available in endless variety. Only the more widely used ones are mentioned in this summary.

Modulation. The public telephone network is engineered and maintained for voice communication even though a substantial portion of the traffic is made up of data or facsimile signals. Such signals must then be put in a form that the telephone plant can handle. This function is accomplished by a modem. The most widely used form of modem operates by frequency modulation; that is, the modem emits to the line a continuous tone signal whose frequency changes up and down in step with the input to the modem from the facsimile encoder. The simplest such modems use two tone levels, for example, 1,800 Hz may be "0" and 2,100 Hz may be "1."

Nonencoding facsimile machines designed to transmit not just black and white, but also shades of gray, are served by a modem whose emitted tone signal can assume any frequency, between the mentioned 1,800 Hz and 2,100 Hz. (These example frequencies are arbitrarily chosen. There exists no obligatory standard.) The receiving modem recognizes each frequency and enables the printer to reproduce approximately the same shade seen by the scanner. Such binary FM modems are limited by nature to a maximum of about 2,400 picture elements per second.

High-speed facsimile machines (1 minute per page and less) employ one or more of the following: extensive encoding, and either high-capacity (anolog or digital) channels, or very efficient modems that can make the most use of the bandwidth offered by a channel. The fastest machines (30-seconds-per-page range) exploit both encoding and efficient modems. These more efficient modems do not employ FM, but rather employ a library of (usually) 8 or 16 different waveforms, each of which represents three successive data bits, or four such bits, respectively. Such modems routinely transmit 4,800 bps over direct-dialed long-distance connections. Their ability to talk through noise and interference is remarkable, in that such modems will accept and use almost 95 per cent of long-distance connections.

Line Coupling. The modem can be coupled to the telephone network either acoustically or electrically, the latter requiring the use of an isolation circuit called a data access arrangement (DAA).

Acoustic coupling has been important to the acceptance of facsimile. Today, it means the placing of the standard telephone handset into a cradle with the earpiece and the mouthpiece adjacent to opposing microphone and speaker elements. Isolation from ambient noises is provided by sponge rubber seals. Prior to 1966, all facsimile devices had to be connected electrically to the communications lines through cumbersome isolation devices specified by the common carrier companies. In some instances the cost of the connection, particularly to the voice-grade public switched network, approached the cost of the facsimile device. The introduction of facsimile coupling permitted much simplification. Most of the low-speed facsimile devices now in use are acoustically coupled. That means is both flexible and, on a first-cost basis, inexpensive. It enables portable machines to be offered which can receive or transmit from any telephone in the office, home, or elsewhere. Acoustic coupling is limited to the lower rates of data transfer, for when high rates (above about 3,000 bps) are attempted, acoustic standing waves develop in the coupler and destroy the data stream. An electrical connection is necessary for high data rates.

Transmission: Analog modems operating above 2,400 bps make very thorough use of information capacity of a channel. For data speeds above 3,600 bps, modems employ adaptive equalization, which is implemented by active circuits in the receiver to automatically and constantly retune the receiver circuits to balance out the distortions of the line.

It is also possible to use digital lines (offered now by a number of common carrier companies), which are lines explicitly engineered and maintained for the transmission of data signals. Coupling to a digital line does not require a modem. Instead, the Electronic Industries Association has standardized a set of interface connection arrangements which permit a digital facsimile to be coupled directly, by hard wire, to the channel network. The standard most employed is termed RS232C.

If the traffic load is more than roughly 5 million bits per day (15 to 20 pages of high-speed facsimile), then digital channels, where available, should be considered and in many instances are less costly than analog (voice-grade) channels. Digital channels are substantially free of noise and errors, and can be ordered in almost any speed capacity desired. To exploit them, new superspeed facsimile machines are being developed and will be offered to commercial users.

5. Reception, demodulation, and decoding. The transmitted signal is detected on the communications channel, demodulated into a format suitable for decoding, decoded, and passed to the printer in a form

equivalent to the original video baseband signal. These processes can occur either open loop (no communication from receiver to transmitter about the quality of the transmission and reception) or closed loop (i.e., with such communication).

These processes are simply the inverses of encoding, modulation, and transmission, and their end result is the video baseband signal, fully provided again with all its redundant data bits, which is fed to the printer.

6. Printing. This step recreates a facsimile of the original input image on paper, film, or a display.

Alongside the overall cost of usage (i.e., fixed cost plus variable charges), the companion feature of facsimile systems most important to the end user is the apparent printing quality on the output document. Here psychological factors come into play, for it has been found that the eye attaches roughly equal weight to the contrast between the whites and blacks of an image and to the actual resolution of the process, in forming its subjective assessment of copy quality. Resolution is defined as lines per inch; a resolution of 100 lines per inch means (in facsimile) that if the input copy presents two parallel black lines 1/100 inch wide, spaced by 1/100 inch, then both of the lines and the gap will be discernible on the output copy. If the same image is printed twice with exactly the same resolution, but once with low contrast (gray on white or black on gray) and once with high contrast (black on white), the eye will believe that the high-contrast copy is much more crisp and sharp. Hence, facsimile manufacturers work hard to achieve high printing contrast. Systems have been offered in the trade which, from a speed, reliability, cost, and resolution viewpoint were excellent, but which had low-contrast output copy, and which were commercial disappointments to their sponsors.

Another interesting psychological fact is that the eye attaches greater importance to horizontal than to vertical resolution. (The raster scan lines of a television picture are scarcely noticeable in the normal viewing attitude, but try looking at a TV picture with the head slanted 90 degrees to one side!) Hence, almost all facsimile machines use greater horizontal than vertical resolution, usually by about 20 to 25 percent.

7. Output paper handling. This step is the bringing of the recording medium into position for printing and removing the medium after the image is implanted. Usually processing of some sort is required after printing to develop and fix the image. This step may occur either before or after removal of the medium from the receiver, and may be manual or fully automatic.

Substantially all business facsimile machines in use today discharge the output copy into a tray or pop open a door so that the output copy can be removed easily. The output of most of the high-resolution machines in use for wirephoto or newspaper page plates is manually removed from the receiver for processing. Automation of the chemical photographic processing is coming into use on recently manufactured machines.□

Planning for All-Electronic Multimedia Corporate Communications

Problem:

Data communications, as we have defined the topic thus far, is a totally nonhuman, electronic limbo of digital pulses (or their analog equivalents) in which the terms data and information are given distinctly separate meanings. Data is the stuff that exists on the interconnecting lines among termini and nodes; information is the stuff that results when data is sifted through a human intelligence (even so-called machine intelligence is human in origin). Unfortunately, this neat definitional nicety gets muddied up as we move applications-wise farther from the computer and closer to people. For example, in electronic mail applications, the electronic letter is a complete informational unit in exactly the same way as an ordinary letter. Truthfully, the definition is probably only important to editors because even the system planner/designer does not think in terms of, "This is data and that is information." All he is really concerned with is where, what kind, and how much intelligence to build into the network.

Communications, in its broadest possible sense, is not constrained by the definitional niceties of data versus nondata communications systems. The most often-used people-to-people media are still the telephone and good old face-to-face meetings with flipcharts, slides, and other supporting audiovisual paraphernalia. But we can make these media more effective by borrowing techniques from the data communications technologies. This report explains how the telephone medium can be improved and offers some excellent insights into how voice, visual, and data communications media can be successfully combined into a hybrid communications forum that can simulate the "presence" of face-to-face meetings even though the conferees may be thousands of miles apart. This new hybrid is called teleconferencing.

Solution:

Telecommunications has a vital part to play in improving the productivity of industrial nations. They will become more important as industrial processes are increasingly run by machines. A growing proportion of human work will be concerned with handling information rather than operating tools or lugging items around a factory. The relative costs of physical transportation and telecommunications are rapidly changing. Physical transportation is becoming more expensive as fuel costs rise, while long-distance bandwidth is dropping in cost.

From *Future Developments in Telecommunications,* 2nd Edition, by James Martin. Chpt. 17, pp 289-303, © 1977 by Prentice-Hall, Inc. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Planning for All-Electronic Multimedia Corporate Communications

Most corporations today spend a substantial proportion of their money and human talent on communication of one form or another. Businessmen spend most of their day communicating. Most white-collar workers spend much of their time communicating with superiors, subordinates, customers, suppliers, secretaries, and computers. The information handling process typically costs from 5% to 30% of an organization's total expenses.[1] Given such a large expenditure, it is desirable to ask: how can corporate communications be made as effective and inexpensive as possible.

If we look at telephone communication alone there is much scope for lowering costs. Table 1 shows typical U.S. costs for traditional corporate telecommunications, most of which is telephone traffic.[1] Large corporations spend many millions of dollars on their telephone facilities.

Industry	Range	Average
Airlines	3%—7%	4%
Banking and Finance	0.6%—4.2%	1.5%
Insurance	1%—3%	2%
Manufacturing	0.3%—2%	0.5%
Securities	8%—12%	10%

Table 1. Cost of corporate telecommunications as a percentage of total operating expenses

Surprisingly, about three-fourths of this expenditure relates to internal telecommunications. A widespread corporation can therefore profit by having an internal telephone network designed to minimize costs. Such a network is likely to incorporate leased lines, WATS lines (giving a fixed monthly charge to or from specified areas), lower cost lines from the specialized common carriers and switching arrangements to interconnect these lines. Some corporate networks now include leased satellite channels.

MINIMIZING THE NETWORK COST

Designing a minimum-cost corporate telephone network is a complex operation. In the United States it is much more complex now than in the 1960s because of the competitive tariffs and new service configurations. Among the actions today that will lower the telephone bill are:

1. Selection of an optimum mix of leased lines, WATS lines, satellite channels, privately owned microwave and millimeterwave facilities, and public facilities.

2. Selection of appropriate lines and services from the specialized common carriers.

3. Optimum geographic layout of the leased line network.

4. Selection of appropriate tandem dial and common control switching facilities.

5. Selection of equipment which will route external long-distance calls over the leased-line network.

6. Use of equipment that will automatically send calls by minimum-cost routes.

7. Appropriate PABX selection.

8. Elimination of switchboard operators.

9. Use of the feature on computerized PABXs which prevents specified telephones from making specified long-distance or expensive calls.

10. Use of telephone monitoring facilities.

Surprisingly, in spite of the magnitude of telecommunications costs, most corporations exercise little control over them. There is often no attempt at corporate network optimization: the telephone facilities are acquired by local managers, who are told what to lease by the telephone company. Leasing a 500-line electromechanical PABX often requires a million-dollar 5-year contract, but there is seldom any technical study or high management involvement as there would be with other equipment of similar cost (for example, computers). Telephone expenditures have been taken for granted.

The new computerized PABXs and telephone monitoring equipment can do much to lower telephone bills as illustrated in Figure 1. (Different regulations in different countries prohibit some of the techniques shown).

USER MONITORING

- All users telephone calls are listed so that users can be asked to explain them. Telephone misuse is eliminated.
- Refunds are demanded for personal calls.
- Excessively long calls are eliminated.
- Management is able to promulgate new rules on telephone usage.

EXPENSE ALLOCATION

- Precise allocation of phone costs to departments and budgets.
- Department managers are made responsible for their phone budgets.
- Users dial a code to identify personal calls.
- Users dial codes to indicate subject matter, for billing or budgetary control.

Figure 1. How to cut a corporation's telephone bill

Planning for All-Electronic Multimedia Corporate Communications

CLIENT BILLING

- Detailed bills for clients are prepared in service organizations such as law firms, architects, accountants, etc., permitting 100% billback of telephone costs.
- Call dial codes are used to indicate client and subject matter for client billing.

AUTOMATION

- Elimination of telephone operator positions.

CALL RESTRICTION

- Specified extensions are prevented from dialing certain area codes. Geographical restriction for each employee.
- Specified extensions can dial only within specified hours (e.g., 9:00 A.M. to 5:00 P.M.).
- Users sent warning tones when calls exceed a specified time.

TRUNK MONITORING

- Quick detection of malfunctioning corporate trunks (which often go undetected for months).
- Telecommunication facilities brought to maximum operating efficiency.
- Refunds won from carriers for malfunctions.
- Console attendant indication of heavily utilized trunk groups.

NETWORK REPORTS

- Computer-produced reports recommending adjustments to network facilities (including WATS lines, specialized common carrier lines, etc.) to handle existing traffic at minimum cost. Such reports are based on trunk monitoring.
- User frustration reports summarize uncompleted calls.

LEAST-COST ROUTING

- Each call is routed by the least expensive route at that instant: e.g., first choice: tie-line; second choice: WATS; third choice: DDD.
- Takes advantage of time zones to increase bulk facilities utilization.
- Permits a major increase in bulk facilities utilization. System automatically calls back user when a low-cost route is free.

PRIORITY SYSTEM

- Low priority callers are prevented from using expensive routes.
- High priority callers are given a high grade of service—almost no busy signals.
- Nonreal-time traffic sent when trunks are idle.

Figure 1. How to cut a corporation's telephone bill (continued)

OFF-PREMISES ACCESS TO CORPORATE NETWORK

- Use of corporate network gives executives inexpensive calls from home.
- Reduces employee credit card calls.
- Uses bulk corporate facilities which stand idle after hours.
- Off-premises to off-premises calls are possible via corporate network.

Figure 1. How to cut a corporation's telephone bill (continued)

DATA NETWORKS

In addition to telephone networks, corporations have networks for transmitting data. The volume of computer data transmitted in many corporations is growing at more than 25% per year.

Being controlled by the data processing staff, data networks are often well optimized, unlike the telephone network. However there are often multiple data networks forming parts of different computer systems that were designed separately. The combination of all data transmission facilities in a large corporation does not normally form an optimized whole. Some corporations are now in the process of linking together separate data networks to cut cost and to increase throughput.

To control data networks, appropriate software and control mechanisms are needed. Unfortunately, the different networks often used different line control procedures, and so to combine them the line control software and hardware had to be changed. Often the terminals had to be changed. Users and manufacturers attempted to standardize the network procedures to make future network growth and optimization easier.

Integrating the separate data facilities has several advantages. It can lower the overall cost. It can decrease the network response time if it is done well. It can increase the total reliability of connections. Some terminal users can have more remote computers and more data banks available to them.

However, optimization of the data transmission facilities alone is suboptimization. What is really needed is optimization of all of the corporate telecommunications. Data transmission costs vary from 1% to 20% of the total corporate transmission costs. Because of economies of scale, especially with the newer technologies, it can pay to combine the facilities that are leased for voice and data traffic. It invariably pays to combine real-time and nonreal-time traffic.

In the future, as voice digitization techniques spread, the combining of voice and data traffic will become increasingly important.

Planning for All-Electronic Multimedia Corporate Communications

OTHER COMMUNICATION COSTS

Today, corporate telecommunications is generally thought of as being telephone and computer traffic, possibly with a few small extras such as occasional facsimile messages. The total cost of communicating in a corporation, however, is much greater than the cost of telephone and data transmission—typically between 5 and 10 times as much. Two major contributors to this cost are the sending of mail in a corporation and the cost of physical travel for communications. In many large corporations about three-fourths of these expenses are for internal communications (as with telephone expenses).

An average piece of correspondence in the United States has been estimated to cost $10 or more to be conceived, formatted, copied, transported, received, read, and filed.[2] This is much higher than the cost of making a telephone call. A very large quantity of memoranda and letters is sent within most corporations. Modern telecommunications enables us to ask a new question. Should not corporate correspondence be sent in an electronic form rather than in the form of paperwork. Electronic memoranda have two important advantages. First, given appropriate system design they can be cheaper. Second, they can reach their destination faster. The highest priority correspondence can reach its destination in minutes.

ELECTRONIC MEMORANDA

Memoranda can be handled electronically in the following ways:

1. Telegraphic Message Switching. Memoranda are typed into terminals like telegraph machines and delivered by a store-and-forward system. They are filed in the system, not by filing clerks.

2. Visual Display Message Switching. Memoranda are typed into screen units that permit easy document editing. Retrieval on screen units avoids paper handling. Screen units on executives' desks would be used for many other functions.

3. Magnetic Card Systems. Magnetic card typewriters permit easy editing and storage of documents, and may form the input to a message switching system.

4. Facsimile. Facsimile machines permit drawings, signatures, logos, and handwritten notes to be sent. If transmitted or stored digitally, facsimile documents require about ten times as many bits as alphanumeric documents. In some cases use of facsimile can save typing costs.

5. Voice Message Storage. As electronics costs drop and secretarial costs rise, the cheapest way to transmit and store interoffice messages may be in the form of speech. Speech messages could go straight from the sender to receiver and avoid any intermediate human processing. Widely distributed memoranda can be made available by telephone.

Many of the machines used for message input and output could be machines already existing in offices with communications adaptors added, for example, typewriters, magnetic-card typewriters, and copying machines.

SPEECH MEMORANDA

Executives spend much of their time dictating messages to other people in their corporation and the secretaries have to type them. The most labor-saving, and potentially cost-saving, form in which to deliver such messages is in spoken-voice form. The spoken-voice messages—voicegrams—would be filed in computer storage. Each recipient would be notified when messages were waiting for him, or could check his file periodically from an ordinary telephone. The cost of storing such messages in a large on-line library store would be about 20 cents (U.S.) per year for messages coded with delta modulation, 2 cents per year for messages coded with vocoder techniques, or 0.01 cents per year for messages composed with code-book works.

The use of speech memoranda instead of typed memoranda would be alien to some executives who cherish their present way of operating. However, it would be more convenient to send many memos by telephone. Whenever an executive cannot reach a person he telephones, he can immediately leave a message with no danger of secretaries misinterpreting it. When an executive is traveling, he can telephone his own file, at night if necessary, and listen to the memos that were sent to him. Security controls can be designed to prevent any other person listening to it. Such a system could be more convenient and accessible than a memo-typing operation.

Memos are used in corporations to provide a record of past instructions and communications, and the memo sender signs the document. The voice-print of an individual saying certain words, for example speaking his personnel number, is as good an identification as a signature. The system may be designed to store speech memoranda in long-term files when instructed to do so. When not so instructed, it keeps them for a given period of time after they are delivered (perhaps a month) and then erases them.

Memoranda containing tables, diagrams, or other items not conveniently represented by speech would still be sent on paper. The paper may be delivered by facsimile or alphanumeric coding, and stored in the same system as the speech memoranda.

Planning for All-Electronic Multimedia Corporate Communications

AN ALTERNATIVE TO TRAVEL

Not only is the total cost of typing and filing memos greater than the telephone bill in most corporations, the total cost of business travel is also greater—and the cost is rising. Some large corporations spend more than $100 million on business travel within the United States. This does not include the cost of the time of the persons traveling or the effects of wear and tear on them, which for some executives are considerable.

The technology has now reached a point when telecommunications can form an effective and cost-saving substitute for some travel. It is necessary to ask the question: What are the best ways for people to communicate with people at a distance? This is a complex question and there are many types of answers to it. Many of the answers require transmission at higher rates than that of local telephone loops, often in very brief bursts.

The following facilities can be used to improve communications at a distance and are now being implemented in pilot and experimental systems

Telephone Conference Calls

Conference calls have been infrequently used in business because of the difficulty of setting them up. Some new computerized PABXs give their users the ability to set up multiple-party voice connections without operator intervention. This can be a valuable facility because it is often very useful to consult a third party during a conversation, or include several people in a telephone discussion.

A telephone "meeting" with many parties at separate locations needs a certain discipline imposing upon it to make it effective. It is necessary for each person to know who is speaking and to be able to indicate to the dispersed group that he wants to speak. This can be done if the meeting has a chairman who disciplines the conversation. Another possibility is for each caller to have a small strip with lights. Each participant has one of the lights associated with him, it is on when he is talking, and he can make it flash when he wants to talk. A low bit-rate control channel is derived from the speech channel for operating the lights. The strip may be designed so that a caller can write the names of the parties by the lights assigned to them.

Telephone meetings may also be held between meeting-rooms in different locations equipped with speaker-phones.

Picturephone

Picturephone adds information to a telephone call by permitting the observation of facial expressions. The cost of this extra information is high—many times the cost of a telephone call—and for many corporate calls it is not worth it. Picturephone does not have sufficient resolution for users to read typed documents, contracts, computer printouts, detailed engineering drawings, and so on.

Facsimile

Paper documents can be transmitted fairly quickly by facsimile means. Telephone callers may use facsimile transmission to enhance their conversation so that they can exchange sketches or documents and discuss them.

Freeze Frame Video

Callers may employ a screen, in conjunction with their telephone call, on which still images can be displayed. If the image is to be displayed and discussed while the conversation proceeds, it is desirable that it should be transmitted fairly quickly—say in 5 seconds. Picturephone transmits one frame every 1/30th of a second. A Picturephone-quality frame transmitted in 5 seconds would need $\frac{1}{30 \times 5}$ times the Picturephone bit rate, or 42,000 bits per second. A black-and-white printed sheet would require about 250,000 bits, or 50,000 bits per second, and a much higher resolution screen than Picturephone. If the telephone call is occupying 56,000 bits per second on a PCM channel, then the image could go over the same channel, interrupting the speech in one direction for 5 seconds or less.

To enhance the conversation, both parties should be able to look at the same image at the same time and point to it. A moveable arrow may be provided on the screen for this purpose, along with a low speed sub-channel for conveying its movements.

It may be desirable to see the caller's face, but as a still image rather than the moving image of Picturephone, which takes so much bandwidth. The lens used to transmit documents could also be used like a Picturephone lens to record the face of the caller. A system could be designed in which a telephoner could press a button to capture the image of the face he is talking to. He could then have some idea of a person's expression at a selected critical moment in the conversation. This would not convey as much human information as Picturephone, but could be transmitted over a duplex channel of telephone bandwidth.

Electronic Flip-Charts

Most major locations in corporations will have visual display units connected to computers. Such screens have been used effectively for enhancing person-to-person communications as well as person-to-machine

Planning for All-Electronic Multimedia Corporate Communications

communications. Two or more individuals talk by telephone and discuss information which resides in a computer storage. All participants to the conversation see the same data displayed and can modify it.

At its simplest level, the machine is being used merely to display human ideas with clarity. In many corporations today, flip-charts are used for this purpose. An employee has a set of facts or ideas that he must present quickly and efficiently to management or colleagues. He writes the information in a concise form, so that it can be grasped quickly, on large sheets of paper hung on a flip-chart stand. Information such as given in Figure 1 is typical of what would be written on flip-charts. Employees travel with a roll of flip-charts to make half-hour presentations to management. The information could be conveyed equally well if the data on the flip-charts were entered into a computer system as a set of single-screen displays and edited until they were as concise and clear as possible, preferably in color. The persons talk by telephone, using the electronic flip-charts in the same way as paper flip-charts.

Electronic flip-charts have several advantages other than avoiding the need to travel. First, they remain in computer storage after the conversation. Management rarely admit it, but they probably remember only a portion of the data that is flip-charted at them. It would be useful if they could review the charts again privately at their leisure and perhaps discuss them with persons other than the original presenter.

The production of paper flip-charts is often made a laborious task. Neat magic-marker takes time to write, and the wording is frequently modified. Computer software could make the entering and editing of screen charts a fast operation.

In a corporation, the many flip-chart presentations could be filed and indexed with appropriate security locks. Many flip-chart presentations are made on the same or related subjects, and the indices would permit computer searches to be made for related material.

In some corporations, flip-charts are one of the main forms of communication, with much money being spent on air travel by persons making flip-chart presentations. It can be a highly efficient form of communication and is susceptible to mechanization. Computer-assisted flip-charting has major advantages.

Communication via a Database

Communication links that handle flip-charts could also handle data assimilated and stored by computer database systems. Database technology imposes a measure of precision on the way data is defined and referred to, and database administrators have often been surprised by how different departments or managers call the same data by different names or different data by the same name. When communication takes place between parties using a common database, there is less chance of imprecision.

In a system at Westinghouse,[3] a graphics terminal is used for production scheduling based on sales forecasts of washing machines. This is a complex operation because Westinghouse makes over one hundred models, all available in several colors. Once a month, the production and marketing managers travel to Pittsburgh to work together on the display console. The marketing managers evaluate market forecasts and assist the production managers in working out the production schedule. The use of the terminal permits more options to be explored than were possible before. Before, according to Reference 3, a "seat of the pants" approach was necessary. Now the two groups of managers can communicate with precision. The managers involved, once experienced with the technique, "wouldn't want to do their scheduling any other way."[3]

The same type of meeting could take place via telecommunication links, with the parties involved able to discuss data that all can see. In some cases, the data will be modified or processed during the conversion.

In reflecting on the ideal forms of man-computer dialogue, it seems that they have much in common with ideal forms of person-to-person dialogue via machines. As person-machine communication improves and person-to-person communication becomes more precise, the two will increasingly tend to require the same hardware, channels, and features.

Teleconference Rooms

Because some of the facilities for improved communications are expensive, they may be installed in conference rooms rather than in individuals' offices. A conference room designed for meetings via telecommunications may be equipped with multiple television screens and cameras, with electronic flip-chart facilities, links to database systems, and facsimile equipment. In some cases, teleconference rooms are equipped with voice conferencing facilities and the equipment selected can all be used over voice-grade links.

Video equipment and the requisite communication channels are expensive, but in many situations not as expensive as the alternative physical travel. Several major U.S. corporations are now using video conference facilities, in some instances to lessen the cost and strain of air travel, in others to give better human communications. IBM has a video education network in New York. New England has a video network linking hospitals.

Radio Paging

Radio paging systems make it possible to contact individuals who are not sitting near a telephone. The

Planning for All-Electronic Multimedia Corporate Communications

individuals wear a small inconspicuous radio receiver that can signal them either with an alarm tone or with a spoken message. Individuals roaming about a factory floor can be instructed to pick up the nearest telephone. Service personnel miles from anywhere can be instructed to go to a stated customer. Paging is one of the facilities of some computerized PABXs but is more often done manually. A few corporations make a massive use of radio paging.

Two-way Mobile Radio

Two-way radio may have a major growth ahead. Two-way communication to persons far from conventional telecommunication facilities is being accomplished via satellite. WESTAR links connect to off-shore oil-drilling rigs. The MARISAT satellites connect ships around the world to their head offices. Ships have been million dollar facilities without the transmission capability of a terrestrial office. Now they can be linked into their corporate communications network like any branch office.

Remote facilities on land can also be linked to corporate networks by radio—possibly via satellite. For example, the trucks that service tractors or earth moving equipment in developing countries can have radio links to their offices.

TELECOMMUNICATIONS MANAGEMENT

The design and management of a corporation's telecommunication facilities need to be conducted on a centralized basis, taking all divisions of the corporation and all different users of telecommunications into consideration. Because of the economies of scale and scope for optimization, centralized design and management can save a substantial amount of money, and can provide a corporation with better facilities than it would have otherwise.

Booz-Allen & Hamilton[1] experience shows that organizations carrying out a centralized audit of conventional telephone facilities typically show savings of 8% to 15%. This savings is soon lost if centralized management procedures do not follow up the controls on a permanent basis.

Telecommunications management is rapidly becoming a much more complex job. The complexity is arising first from the new types of tariff structures and services such as those of the specialized common carriers, the interconnect industry, the satellite common carriers, and the value-added common carriers. Second, it comes from an increase in the equipment complexity and diversity, for example computerized exchanges, concentrators, intelligent terminal controllers, data network hardware, satellite demand assignment equipment, radio facilities, and so on. Third, complexity arises from the need to incorporate diverse traffic into the network facilities, such as data transmission, facsimile, video, electronic memoranda, security monitoring, and radio paging. Fourth, it may be difficult to establish the tradeoffs between various reasons for travel and telecommunications, and secretarial services and telecommunications.

Efficient telecommunications management is particularly important today because telecommunications costs in most corporations are rising more rapidly than the inflation rate. Surprisingly few corporations have tight centralized management of their telecommunications resources. Those that do rarely take the broad perspective of communications tradeoffs, including the travel budget, the secretarial costs, and data processing budget.

A number of stages can be observed in the development of telecommunications management:

Stage 1. No centralized managment. Lines and equipment are leased haphazardly by user departments.

Stage 2. Centralized planning and management of the corporate telephone facilities. Optimization of separate data networks by data processing system designers.

Stage 3. An attempt to combine the various data networks into an integrated data transmission facility, possibly including a message switching system.

Stage 4. An attempt to combine the data and voice facilities into an overall optimized plan.

Stage 5. A study of all forms of corporate communications including internal mail and memo typing, executive travel, and use of information resources.

Stage 6. Cost optimization of all forms of corporate communications.

To achieve Stage 6 is a very complex operation. Few corporations have contemplated it yet, let alone succeeded. The technology becoming available will make it increasingly profitable.

REFERENCES

[1]From a draft of an article by Harvey L. Poppel and Anthony G. Ward: *Time to Tame Telecommunications*, Booz-Allen & Hamilton, 1975.

[2]The $10 figure is from a Booz-Allen & Hamilton study and is quoted in the above reference.

[3]William E. Workman, "Which Color Washer Will They Choose?" *Computer Decisions*, December 1969.□

The Communications Framework of Management Information Systems

Problem:

The simple goal of good communications system design is to provide easy paths for people to talk to each other in every conceivable way—through intervening computers, through terminals, by voice, by pictures, by digital signals, and so on. But then what? Once the paths are established, the communication system becomes subordinated to the higher goals of an interlinked information system. In fact, these goals are really the original source of the pressure to design a communications system. The pressure may be very tentative in the beginning—just a generally expressed need to develop some way to get more information from here to there faster—but once those needs are satisfied, the improved communications linkages can provide a framework for all kinds of interesting applications like distributed processing, communicating databases, and management information systems. This report places communications into an information system context and shows how all the pieces can be fitted together to form a very powerful management tool.

Solution:

The success of any information system can only be measured in terms of user satisfaction. Since primary use of information is to provide a basis for decisions, it is appropriate to make an in-depth analysis of the user as decision-maker to assess some of the criteria for user satisfaction.

Advanced techniques of communication, computers, and audiovisual aids have been used extensively to eliminate the drudgery from many business activities and to provide economies in jobs where much clerical help was needed. Systems known as Management Information Systems (MIS) provide means for documenting the trivia of business activities and for generating summary statistical reports to help the middle manager monitor these activities. However, the full potential of modern devices for augmenting human skills has not been realized due to the inadequate integration of the supporting techniques. The many operational systems have at present only one common element—the man who makes use of them. He must learn to use each system on an individual basis. Complete integration, correlation, and synthesis occur in his mind rather than in the system that supplies the information to him.

Surely a level of sophistication has now been reached where a preliminary synthesis of these many elements can be achieved with the aid of computers. The final coordination and evaluation of the integrated data must still remain the responsibility of the man, but much of the redundancy can be eliminated, and detailed processing can be carried out more rapidly and accurately with the aid of machines of various types.

From *Information Systems and Networks* by K. Samuelson, H. Borko and B.X. Amey. Published by North-Holland Publishing Company. © 1977 by Unesco. Reproduced by permission of Unesco.

The Communications Framework of Management Information Systems

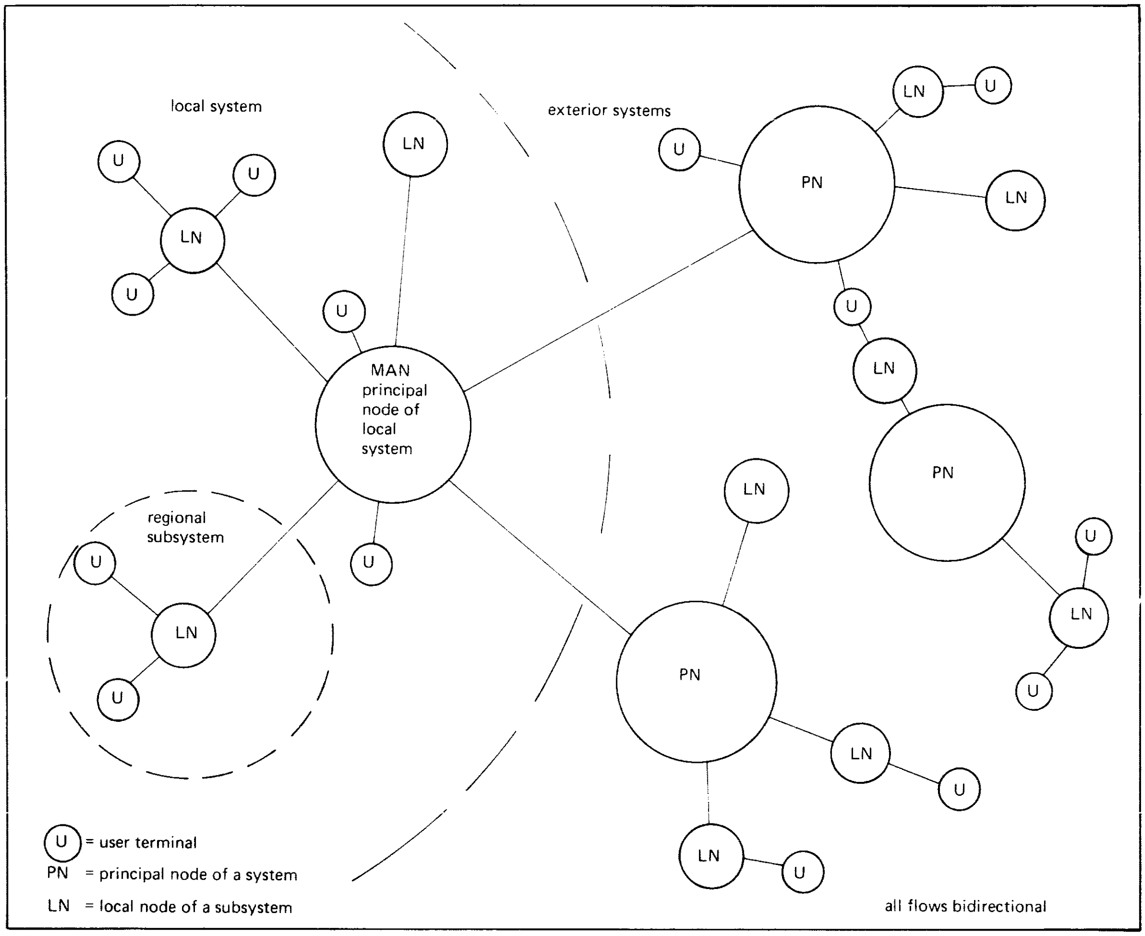


Figure 1. Information network

Man, who is the target (a node in an information network, Figure 1) of converging streams of information, must make decisions based on the data received and must transmit information to the network. For the purposes of this report, the user is considered not only as an isolated individual but also as an element in some large organization. The organization is assumed to have some idealized pyramidal structure (a hierarchy) of the line-and-staff type. Neither the exact type of organization, whether commercial or governmental, nor its size will have much effect on the general analysis.

It is to be expected that full utilization of mind-augmenting resources already available will not be achieved without reconfiguring organizational structures and modifying personal habits to harmonize with the new possibilities. The concept of congregating workers into large groups is not consonant with proper use of modern communications techniques, which effectively annul space and time between people situated at different locations. The depth of hierarchical structures as well as the relative importance of lateral or vertical flow of authority, will probably have to be modified to optimize integrated decision-making (management) systems (DMS) containing advanced machine elements. The final achievement of such a goal can certainly take a decade of effort.

COMMUNICATION HIERARCHIES AND INTERFACES

Communication within an organization or among organizations calls for the making of decisions according to the status of the individuals involved within the respective hierarchies. At the lowest level, the individual carries out tasks in reaction to instructions received and need make few decisions requiring significant amounts of information beyond the rules of the procedure manual defining his task. His timespan for response to a requirement is very short, and he normally cannot initiate communication with others, except in the form of routine data-entry in a rigid format. This is an example of operative information.

At the top of the hierarchy, the individual is not often in the position of having to react to outside events.

The Communications Framework of Management Information Systems

He originates directive information, which may affect all lower-ranking members of the organization, and he makes all essential policy decisions. The time-span of such activities is long. He is concerned mainly with the future and with setting the style of the organization. On the other hand, at crisis times he must make decisions rapidly. In both of these modes of action, as planner or prime decision-maker, he needs to call on a large body of information, which has been predigested according to rules he recognizes, principally by mobilizing his staff officers. Broad policy developed by the top manager forms a directive basis for operative information and rules to be formulated by middle managers to govern day-to-day operations.

Middle management is primarily concerned with day-to-day management of ongoing activities and with implementing the policies initiated by top management. The middle manager ensures an orderly replacement of resources used and monitors the flow of activities of the staff within his functional area. His time-span of action falls between the bottom and top levels just described and he operates primarily from inside information such as provided by a conventional MIS.

Staff officers, not directly within the chain of authority of a line-type structure, are not primarily decision-makers as much as transducers and evaluators of information that can influence current and future policies and operations of the organization. They provide technical advice to top and middle management to help integrate all the activities of the organization with a view to optimizing the overall system and identifying new opportunities. They must be particularly sensitive to changes outside the organization that require continuous adaptation of the local system. They have the longest time span for activities of the classes noted, and the greatest need for both outside and inside information. The essential problem is how to match channels of communication to each user category of an information system.

A conclusion may immediately be drawn as to the place of computer aids in such a system. There is a hierarchy of media suitable for the presentation of information related to the hierarchy of decision-makers. Direct communication with a computer, at present, requires keying of queries or data. It is thus unlikely that top management will make direct use of computer aids until voice communications become practical and easy with a computer. There are some tasks for which humans will be superior to computers for at least the next twenty years , and a new class of staff officers, information aides will be needed during this transitional period to interface between people and specialized information sources. Visual aids, whether to display analog or other data stored in a databank or for use in video-conferences, may soon be used.

At middle-management levels, conventional management information systems (MIS) process and digest internally generated (endogenous) data to document the movement of people or material within the organization (personnel records and logistics), to provide accounting aids, and even to make decisions (reorder time and quantity for stores) according to algorithms designed by the technical staff and authorized by higher management. Better use of the data in these systems may even enable emerging problems to be foreseen before the crisis point is reached and possible solutions to be formulated by the computer. In fact, several of the decisions formerly requiring human intervention may soon be replaced by automatic computer response based on algorithms that are concrete expressions of top management policy. The skills of middle managers will be needed primarily to formulate these algorithms and to cope with exceptions not provided for in the programs.

At lower operational levels, data will continue to be entered by traditional keying methods. There will be an increase in task-scheduling and prompting by programmed computers. Simple interactions with the system will continue to be mediated by codes rather than by natural language, or by simple voice or light-pen type response to multiple choices offered by voice or visual displays (menu prompting).

MANAGEMENT INFORMATION SYSTEMS (MIS)

MIS were developed to monitor everyday operations of organizations by processing clerical data and digital or analog inputs from automated processes imbedded in the work stream, related to routine transactions of strictly defined types, and are now at a quite mature state of development. A high level of integration of subsystems, such as pay, logistics, personnel data, etc. into corporate databases, has been achieved. Statistical reports are generated automatically or on command by standardized report generators. Quick response can be provided when required for individual transactions. The logic of these sysems often requires that large-scale computers be used to manipulate large databases depending on the nature of the application. If the complexity of processing is not too great (i.e., only a limited number of tightly defined queries is allowed) many users can be serviced, as in the case of various reservations systems, e.g., for airlines. To date, attempts to combine complex processing with fast response systems for many users have not been overwhelmingly successful.

INFORMATION STORAGE AND RETRIEVAL SYSTEMS (IS&R)

Information-retrieval systems, such as those used in library or scientific documentation activities, are conceptually at a high level of development. The more sophisticated types provide immediate responses to queries to enable the user to redefine his requirements in the light of the data he receives. Large databases are required with greater flexibility of definition than is usually allowed in MIS or in airline reservation systems. The complexity of file handling and processing is usually such that a dedicated computer system is required, even to handle a relatively small number of users. An IS&R system differs from the MIS in that some of the data may be stored in more than one file in the system to accelerate searching and to ensure adequately fast responses to online queries.

Whereas data in a MIS is usually internally generated, the bulk of data in an IS&R system is externally generated. Even the largest system of the latter type could not possibly hold, online, all the information that a user might wish to search. Thus, complex selection and purging techniques must be employed to ensure that only the most frequently used material is stored in the local, rapid-access files. Since complete control of all information that may be needed is not possible, the internal system must be coordinated with external ones to make fullest use of all available information. In most disciplines, any one nation, such as Canada or Sweden, generates only a few percent of the current production of new knowledge. The contribution to the total cumulated stores of information is even smaller. Information from many external systems may well be required by a user normally served by a specialized system. It would be unfeasible to attempt to create a single giant system (a megalith), which is supposed to combine all data into one exhaustive data-base, serviced by a single vast computer.

Reorganization of data in a MIS within an IS&R system could be worthwhile to determine implicit correlations, e.g., between personnel movements and material wastage; or between incidence of illness and geographical location—which would not be apparent in a conventional MIS. In other words, the original MIS had better be designed to anticipate future desirable uses of the existing information. In practice, this means that exploiting the flexibility of multiple files often necessitates that a generalized file handling or database management system be developed to support the MIS and IS&R systems for swift retrieval.

ADMINISTRATIVE INFORMATION SYSTEMS—THE OFFICE

Up till the present, computer-based techniques have made little impact on office activity. The principal methods of communication for administrative purposes are conversation in person or by telephone with other people and correspondence by mail or memos. The organization of mail is in essence a library process. Incoming mail is date stamped, logged, classified by subject, entered into a file cover, and circulated to the addressee or a number of individuals who might be interested in or responsible for taking action on it. The addressee writes a reply, and a copy of the reply is filed with the original letter or transaction form.

In this case, file-handling entails loss of file integrity whenever some of the dossiers are in the hands of individuals. Since the file cover hopefully contains all relevant documents on the subject, only one person at a time can deal with this subject if duplicate copies are not made. In practice, duplicates of part of the master file are maintained in the office files of individuals. These occupy considerable space, and individual dossiers are never as complete or as up to date as those in the central file ought to be.

Modern techniques have been applied to incoming mail received by telegraph. In the system designed by AT&T for the US State Department, telegraph signals (i.e., messages in machine-readable form) are displayed on cathode ray tubes (CRT) for assessment by message analysts. On the basis of content, the analysts determine a list of individuals to whom the message should be circulated for comment or action. The messages are then reproduced in the required numbers of copies and are transmitted to recipients by internal office mail. The message itself is stored in a data bank that can be queried from terminals.

This system appears to have a gross drawback in that it multiplies the amount of paper stored. It would appear preferable if terminals with CRT displays were provided for all recipients, who could then be signalled whenever a message is routed to them for action. The officer could press a key to display the message and immediately enter his comments or route the message to others. No paper need enter into the transaction.

Such an approach appears to be the way of the future when mailing of paper will most certainly be replaced by electronic mailing.

In the transition period during which typewritten correspondence will remain the norm, machine-readable data can be provided as a byproduct of typing, either at an editing terminal or by sufficiently standardizing type-face quality of paper and typewriter ribbon and letter-format so that Optical Character Recognition (OCR) systems can be used to capture all or part of the typed material.

The Communications Framework of Management Information Systems

Indexing of Transactions for Filing

The greatest weakness of traditional filing methods is in indexing. Items are usually filed uniquely under a single subject heading and are thus difficult to retrieve. This method is similar to that of shelving books according to subject in a conventional library. However, libraries usually have catalog cards to provide cross-referencing and multiple entries. This is not customary in a filing system for transactions, internal memos, external messages, or correspondence, even in a well managed library. An integrative approach should be used even though an overall integrated information system may not be within immediate reach for libraries, archives, or other organizations.

If the item were to be filed under more than one heading, it would need to be duplicated to provide other copies held in other file-covers. This leads to difficulties in keeping the files updated.

The first requirement for an efficient system is to divorce the physical location of an item from subject dependence. Incoming documents or transactions at one node of a network should be filed in strict chronological order of receipt. Ancillary files (which could be card files in the interim stage of changing over to a machine-system) will then locate the accession number according to subject, date of receipt, etc. Minimum indexing requires that codes for sender, recipient, and subject should be entered on each piece of a transaction as it is entered into the sequential file (which may be microform).

Indexing is currently done by clerical help. The quality should be upgraded by the addition of more subject entries and better definition of indexing terms used. Mnemonic or other codes authorized by a standard codebook must be used for entering data that identifies the source and recipient of the letter. Accessions of memoranda, etc. received from sources outside to the system (or cooperating with it) could be coded at the source to reinforce the intent of the sender.

If such subject, source, and recipient files are entered into a computerized data-bank, machine-edits can be used to ensure that valid codes are used, thus simplifying retrieval at a later date.

Retrieval of Filed Transactions

When an individual has been actively working on a task for some time, he will usually know exactly which files are of most use to him. Another person, tackling the problem for the first time, is not in so fortunate a position. Since report or letter writers do not always treat one subject per document item, relevant data may very well be contained in dossiers, the main subject heading of which appears to be very different from the target subject.

Complete analytic cataloguing by a transaction or document analyst can solve this problem, but it is a very expensive undertaking and tends to delay handling of the file and the action on the messages received. In addition, it is not always possible to predict how the information in the file may be used in the future. Thus, indexing that is adequate for current requirements may rapidly become out of date.

If the messages are received in machine-readable form, they can be indexed by computer, using conventional methods. In addition, if it is desired to search an old file recorded in this manner, it is relatively simple to reindex the material for searching according to the concepts then current.

When transactions are received from an external organization, it is evidently difficult to impose on outsiders the requirement that their documents and messages be in machine-readable form. Thus, machine indexing of incoming material will not be practical in the near future without expensive rekeying. However, practically all messages are tightly linked to other messages generated within the organization for which machine indexing is feasible, usually by the subject code (currently called file reference number). In searching a file, this citation coupling can be used to transfer, in effect, the indexing of the inside data (outgoing correspondence) to the outside data (incoming mail). The method suggested is analogous to certain bibliographic coupling techniques. Once an organization such as a library adopts the principle of automation, the full benefits of efficient computer usage should be taken into account, especially for simple clerical tasks.

Integration of Administrative Information Services

By use of a universal query language, to be proposed below, it is possible to tap sources of information from many specialized autonomous systems such as the following:

- Information retrieval systems,
- MIS data banks,
- stored drawings,
- stored correspondence,
- personal files.

CRT displays should be used for graphics. Since both video and digital data need to be displayed, either two

The Communications Framework of Management Information Systems

terminals may be used, or one all-purpose display can be used.

Datapro Comment:

An extraordinary range of products is offered around the basic CRT visual interface. The simplest product is a passive, or slave, monitor, which is even less complex than the ordinary TV set. The next step up is some form of interactive display terminal, but it is a giant step because the degree of interaction can be relatively superfluous, limited to a light pen or some simple form of cursor control, or it can be a highly versatile mixture of keyboard, light pen, electronic pad and cursor controls augmented by full color and unlimited graphics. However, the apparent simplicity or complexity of the hardware is deceptive because almost all of the real interactive power of a display terminal is a function of the behind-the-scenes software (including microprocessor firmware) as soon as one begins to use a CRT terminal as more than just a communicating typewriter. Variations in terminal hardware are thus more cosmetic than intrinsically functional and are generally designed to fit the applications environment (i.e., the display in an office looks physically different than the display in a chemical plant, but the real functional difference between them is determined by the software).

In addition to retrieval services, communication between user terminals should be possible with controlled sharing of user files.

Other possible uses are of the electronic calendar variety:

- recording appointments,
- tickler reminder systems,
- daily planner and organizer, etc.

Reminders of appointments could be routed automatically by computer through, for instance, the telephone system, using synthesized voice messages, to officers at conferences or on business trips if desired, or voice data could be transmitted in response to prearranged codes through touchtone telephones.

Selective dissemination of information, with time-limits for response to action notes, can also be incorporated into the system to aid in integration of responses to messages received.

The ability to modify certain files, under full access control, is important in areas of high technical content, where specific experts in a field may be given authority to change or add to indexed technical data, and to add numerical data or review material to analytic cataloguing data provided by the basic manual or machine system via a "wait" file. Thus information analysis activities can be carried out by those most competent to do so from their own office or laboratory.

In a fully integrated system, the principal means of communication is via the computer system. That is, everyone, down to the lowliest staff member would have a terminal of some kind. The inventory clerk may work with a single-line display and a limited keyboard. The typist will require a full keyboard, with upper-and-lower case displayed on the screen. Terminals capable of multicoloured graphics displays will be required at the higher levels. Teleconferencing facilities will be required in conference rooms to avoid unnecessary travelling to attend business meetings. The inputs and displays must be convenient and acceptable to the level of user for which they are designed. Human engineering is essential to avoid user frustration. The system must be a forgiving one and have acceptable default options when the user does not address it in an optimum manner.

Elements of an Integrated Administrative System

TRAFFIC DENSITY. Certain activities are so specialized that they are most appropriately treated as autonomous subsystems; e.g., information services serving design centres of research establishments working in highly specialized subject areas. Since the traffic associated with these data-bases is concentrated at nodes sited at appropriate centres of specialization, it is not cost effective to maintain the same data at some central node with its prime dedication to general use. Online, reactive (fast-response) querying is usually required of large dedicated disc files. The principal output activity of these files will be stimulated by the information specialists responsible for analyzing and searching the data. Thus, needless cost would be incurred in communication between the central computer and the terminals of the information specialists if all traffic is routed through the centre. Needless processing overheads will be incurred by combining this specialized information system with others of unrelated systems.

A similar analysis applies to data processing relating to administrative detail at local centres. For example, in the military, control and planning for daily food ordering is most appropriately carried out by an autonomous subsystem. Only report information from digested data on gross movement of funds and supplies need be periodically passed to the central node of the information system.

The Communications Framework of Management Information Systems

The existence of natural nodes in a network of communicating computers can be detected by a traffic analysis of messages circulating in the system. Cost effectiveness indicates that heavy local traffic should be switched, controlled, and recorded at a local node.

INTERFACING NODES. Special consideration must be given to ensure that the master node will be able to communicate with and monitor local nodes. This calls for a certain degree of standardization.

- Between data-elements common to the MIS and the subsystem.
- Formats and codes for interchange of information between autonomous subsystems having data elements not part of the MIS for tape or line transmission.
- The query language for querying files serviced at any node within the system (irrespective of local file structures).
- Coordination of system planning with external organizations that will be called upon frequently to supply specialist information, external to the organization. Use of a common methodology is desirable for meaningful communication between system managers.

EVOLUTION OF AN ADMINISTRATIVE INFORMATION SYSTEM

Conceptual studies of formal elements of a system are necessary to ensure that the elements being developed independently will mesh to form parts of the optimal total plan. This will allow for orderly progression from the original hybrid collection of uncoordinated subsystems, which entail much duplication of file storage and overlapping of file-handling activities, to an integrated overall system with fast response and minimum duplication of files or effort. Formal analysis of the essential parameters of the system will allow standardization of parameters that will continue to be used, perhaps stored and processed in different ways, throughout the evolution of the system.

Man-Machine Integration

HUMAN PROBLEMS ARISING DURING SYSTEM DEVELOPMENT. The introduction of computers into operations formerly carried out by other means has usually resulted in frustration and opposition from the people operating the system undergoing change. Proper use of computers appears to require a fundamental reorganization of the managerial structure of an organization. Initially, some individuals lose status or freedom of action as a result. Thus, a revolution in the way administration is carried on inevitably will call forth resistance that can easily destroy the cost effectiveness of the changes. This is most apparent where incumbents of critical positions have held the same post for many years. The problems are not all generated by the original staff members, who probably are doing an excellent job using the traditional approach. System analysts occasionally take simplistic views of operations and tend to find neat solutions for watered down versions of the operation and omit all the virtues of the system being superseded.

Re-education of staff to new ways may be impossible, especially if they are unwilling to cooperate. In organizations where postings are not of long duration, the problem can be less abrasive since phasing in of a new mode of operation can be synchronized with transfer of staff who might otherwise prove to be obstacles to successful implementation of the plan. Change, when properly handled, can even result in improved morale of the workers involved, as in the well-known Hawthorne effect.

The importance of people in the overall system must never be minimized. The system is not an end in itself. It is intended to augment the capability of its human masters. Machine efficiency should be subordinated to the comfort and convenience of system users. As the cost of hardware continuously decreases, it becomes evident that the main costs of a total system will continue to be the salaries of the human components.

Milestones of Implementation

Milestones of implementation can be planned to ensure that independent elements of the final system can be implemented in stages to achieve quantum jumps in efficiency and convenience. A system that demands extra effort and a change in the habits of the people involved, without any apparent increase in convenience or status for the participants, is unlikely to succeed. Implementation should be attempted initially only for part of the organization, while the bugs are being removed.

Model Elements and Pilot Networks

Implementation within the planning group itself would be the logical first step, since its members should have a vested interest in getting the system to work. In operating realistic models of the final system, the essential character of the inputs and outputs (format, response time, display mode, man-machine interface, etc.) should appear the same as in the target system. Ideally, the new services should sell themselves so that the potential users will request that the system be extended to them. It is evident

The Communications Framework of Management Information Systems

that the specific hardware configuration used during this stage need not be that chosen for final implementation, but it should be close enough so as not to affect the credibility of the program.

Program Initiation

In an evolutionary approach, several items can be singled out for immediate action in a multi-pronged approach to the problem.

Information services have already been developed to service many activities. However, clerical transaction handling has not registered much impact from the introduction of computers into the conduct of business. Thus the subsystem dealing with automatic filing, indexing, and display of correspondence deserves special attention. The precise modalities used—whether digitally indexed video-tapes (e.g. Ampex and Sony systems) or microfilms—is not important during the stage in which the mode of retrieval and display to the user is being developed. Such a system should be implemented on a realistic scale, as soon as possible, to iron out conceptual bugs. The main requirement is for flexibility so that continual enhancement is possible based on feedback from a sample population of real users. Provision for automatic feedback of user response should be incorporated in the system from the beginning.

Indexing standards are needed for transactions and other files where they are not currently available or adequate. Coding of the sources of correspondence is an essential requirement to facilitate retrieval of stored information. From the beginning, a method must be incorporated for validating source codes; for updating source codes as names of orgnaizations change; and for linking files relating to both old and new names. Tracing name changes in old files is one of the most time-consuming of clerical tasks.

Traffic analysis of existing correspondence, signals, and other communications within and beyond the organization is necessary to determine the optimal size for data processing operating elements. Perhaps, in the early stages, the more refined processing should apply only to correspondence and other information traffic between managers in higher echelons of the organization. The direction of flow of communication and the receipt-response loops for messages should indicate the most cost-effective application of DP equipment.

For internal efficiency, subsystems will need to operate in modes internally optimized for the specific subsystem. Complete integration of different modes of operation appears to be impractical and uneconomic. However, portions of the data, digested and evaluated locally, may be required to contribute data elements to the central system. If the data in a subsystem is to be queried directly by any user in the overall system (possibly switched through the central computing node), it is necessary to define a system-wide query language, which, with default options, can be used to query any accessible file within the system. Otherwise it would be necessary for each user to know the job control language and local query language of each of the subsystems. This language should be defined early since it is certain to have an impact on the file structures and manipulation capabilities of each of the subsystems. Where desired, the subsystem could provide an interpretative front-end to allow for translation of a standard query into the format required for the local system.

It is possible that the final information system will take over many of the secretary's tasks. However, making use of the complete range of information services that will be available at various stages of development will require skilled technical help to interface between the senior decision maker and the system. Acceptance by users of radical changes in doing business by the methods sketched calls for accurate delineation of this function and training of information aides (special staff officers) at a very early stage of the development. This is all the more important as the system will be in constant flux for many years as each subsystem is brought online and separately debugged. The essential role of this class of staff cannot be too strongly emphasized. It is analogous to the role of the service engineer supplied by major hardware vendors to help introduce a computer and maintain it for a client.□

Network Diagnostic Tools: An Equipment Vendor Survey

Problem:

The problems of selecting just the right diagnostic equipment for a communications network can be as confusing and as complex as selecting the network components, especially for the designer who is not accustomed to the chore of outgoing maintenance for a system whose parts may be physically separated by perhaps hundreds of miles. This report offers a comprehensive survey of currently available network diagnostic equipment, complete with the latest prices, and defines the general diagnostic application for each equipment item. The survey will help to get you started in the right directions, but don't hesitate to contact any manufacturer listed to find out all you need to know about an item before you spend your money.

Solution:

Data communications systems typically share these characteristics: multiple equipment vendors; geographical site separation; and transmission facilities that are supplied by a carrier more attuned to voice applications than to data transmission. In the past, users have found it difficult to even judge the operational efficiency of a system much less to identify potential or actual problems and make the determination as to who should correct them.

In the early days, most processing was centralized, and in the event of a malfunction the mainframe vendor made the service call. Early on however, sprawling low-speed teleprinter networks used for corporate communications made their way from the back room communications center to the computer room. This did not represent a major difficulty, except in terms of service timeliness and user frustration, because the communications manager generally let the two giants (mainframe vendor and communications carrier) resolve any jurisidictional problems in the event of failure.

Then came the timesharing boom of the 1960's and suddenly the number of vendors offering terminals, modems, multiplexers, plug-compatible hardware, and proprietary software packages skyrocketed. Although this period produced some trauma, it was manageable and most erratic problems could be traced to noisy phone lines or incompatibility (timing or interface) between system elements. Although the battlefield was no longer limited to the giants, problem solving was still not too difficult; it simply became the responsibility of the last vendor to contribute to the system (who almost always happened to be the smallest involved) to identify the fault and provide the communications manager with enough ammunition to force the guilty party to correct it.

About the same time, the Carterphone case, the MCI decision, etc., were settled and the ranks of the carriers swelled with packet message switching vendors, value added carriers, and specialized common carriers each offering its own brand of regulated, semi-regulated, and unregulated special services for voice, data, facsimile, video, message, etc. Add to this the international record carriers, the satellite carriers, and the proposed or existing electronic mail carriers and our communications manager was no longer dealing with one friendly account executive but was suddenly confronted by a host of carrier representatives—and "finger pointing" when problems arose gained one more level of density.

While all this was going on, technology hardly stood still. The practical limit (and according to most major suppliers, the "theoretical limit") of transmission speed over a 4-wire, leased, voice grade line was increased from 2400 bps to 3600, 4800, 7200, and finally to 9600 bps, full-duplex. Half-duplex over a dial-up circuit went from 2000 bps to 2400, 3600, and then, comfortably, to 4800 bps. Today, dial back-up using two calls to simulate a 4-wire leased line, is commonly used at 9600 bps for full-duplex operation. All this didn't happen overnight, of course, and in the early 1970's there were many unbelievers. Although 4800 bps for military applications had been in limited use since the late 50's, it was employed primarily over radio links having a wider bandwidth than terrestrial links and it was extremely expensive. Supported by the "theoretical limit" concept espoused by the carriers, most communications managers were of the opinion that—sure, somebody with savvy can fine tune a demonstration unit for operation over optimized facilities, but can they bring it to production at a reasonable cost and will it perform satisfactorily over my facilities, many of which are the next thing to barbed wire; and incidentally, who's going to service it? With these formidable objections, how did the trend ever get off the ground? The impetus came from the computer room.

With mainframe performance levels increasing almost logarithmically, DP personnel were more than mildly upset that the throughput, applications possibilities, utilization, and overall satisfaction with their CPU's was being inhibited by slow-speed, antiquated communications facilities that delivered marginal performance and appeared to cost too much. The DPer's key was speed—in bits or bytes per second—they cared not about the comm. man's baud rate problems. If they had to suffer through all of the NAK's, slow turnarounds, timeouts and disconnects, how could it get worse at a higher transmission rate—it had to improve. Through exposure at trade shows and advertising, DP personnel showed marked interest in the new products being announced and salesmen, being what they are, began to court this possible new source of business—in the computer room.

Speed is what they asked for and that's what the industry gave them. Frequently offering no-obligation trials, the faster machines began to generate acceptance, at least on the DP side of the house. Naturally, the communications people were aware of these tests, if for no other reason than they still ordered up the lines. Beginning with point-to-point lines, polled circuits were then experimented with. Again, the problem was speed; this time not link operating speed but turnaround time, which to the DPer meant degraded throughput. With the advent of LSI (Large Scale Integration) and microprocessors, fastpoll modems were invented, which drastically reduced the training time required to synchronize the modems and greatly improved performance. Again, there was reluctance—"OK, it works all right on this circuit but we have a nationwide network that encompasses various operating companies wtih different vintages of equipment; we certainly can't cut over the entire network; besides, who's going to keep it running?"

Who indeed! Test equipment had always kept pace with the rest of the electronics industry but it was designed for use by engineers and skilled technicians. About the only diagnostics available to an operator or user were simple front-panel meters used by radiomen to tune transmitters and receivers. This period was also explosive in the semiconductor industry with new chips and improved performance reported monthly. The semiconductor industry reported breakthroughs regularly, but it took applications engineers serving the end user community to harness that power. After active filters had been perfected and improved modulation/demodulation techniques had been proven, the telecommunications engineers in the various modem/muliplexer shops realized that there was still capability to spare. With size, power consumption, and cost on the decrease and transmission performance enhanced, they decided to use the additional capability to fight the biggest objection to their success—diagnostics. Soon automatic loopbacks that could be controlled remotely with the distant site unattended were commonplace. The distant point in the system that could be controlled in this manner keep increasing until an operator (or the front end, under software control) could extend the loopback function all the way through tandem multiplexers and distant modem links right to the brass terminals at the far-end CRT, teleprinter, etc.

These diagnostic routines came to be expected and some of the smaller companies who either could not afford the engineering or the chip design/set up costs began to feel the crunch. With tight money and venture capital no longer willing to risk investment in high technology ideas, some of the smaller companies folded or were bought out and the industry experienced a shake out that worried a lot of potential customers. With funds at a premium, many firms had to cut back the frills, one of which was mistakenly identified as service. Again, who's going to fix it?

Although telecommunications test equipment had always enjoyed a certain amount of success, it was hardly big business and was generally catered to by small specialty houses. With their market threatened, the more successful digital transmission firms, primarily modem and multiplexer companies, jumped in with both feet to create a wide range of test, monitor, and control hardware that complemented the diagnostics already built into their primary products.

Exotic network management and control systems emerged and the communications manager was at last comfortable with his ability to diagnose difficulties, even in advance of failure, and frequently to correct them by remote control. This is usually accomplished by initiating a dial back-up circuit around a failed line or to a hot standby spare in the event of a modem failure. Since most of the diagnostic systems available today operate in conjunction with the transmission hardware manufactured by the same firm, sales spur one another and profits can be invested into even more sophisticated systems.

It has long been the contention of the FCC, the US Congress, and to a certain extent the Justice Department, that competition will improve service and reduce costs. Nowhere is this truer than in data communications. AT&T, although belatedly, has managed to keep reasonable pace with modem development and most recently has been called to task for network diagnostics. Their answer is called Dataphone II, a centrally located control system used in conjunction with newly announced 2400, 4800, and 9600 bps modems. IBM also announced a new series of modems for delivery in the summer/fall of 1980. Both of these new offerings will be controlled by central site hardware, are capable of remote back-up, and do not require operators at the distant end. The AT&T version can be controlled by several different models of hardware at the central site; IBM's system will be under control of software resident in the front end. It should be noted that both of these systems require the use of the same companies' modems. This is also true of the larger, more familiar independents: Codex, Racal-Milgo, and Paradyne. General DataComm is believed to be the only major manufacturer of diagnostic management systems that can operate with virtually any standard modem.

Although network management systems are comprehensive diagnostic and control items, they are very expensive and as such are generally suited only to large networks where the cost savings and improved up-time make them worthwhile. However, between the simple manual patch panel and a network management system, a wide range of hardware is available for almost any application and budget. The remainder of this report focuses on this hardware, how it is used, what it costs, and who provides it.

TYPES OF TESTS AND EQUIPMENT

Loopbacks

This is the most common form of testing and entails connecting the transmitter and receiver elements of the device under test by mechanical or electronic means, under either local or remote control. The data source may be any device capable of generating a particular data pattern and comparing the received result against the transmitted standard. Loopback can be at either the analog or digital interface and in protracted systems can usually occur at multiple levels. Several levels of loopback are shown in the accompanying diagram. (Like loopbacks are possible for all elements shown.)

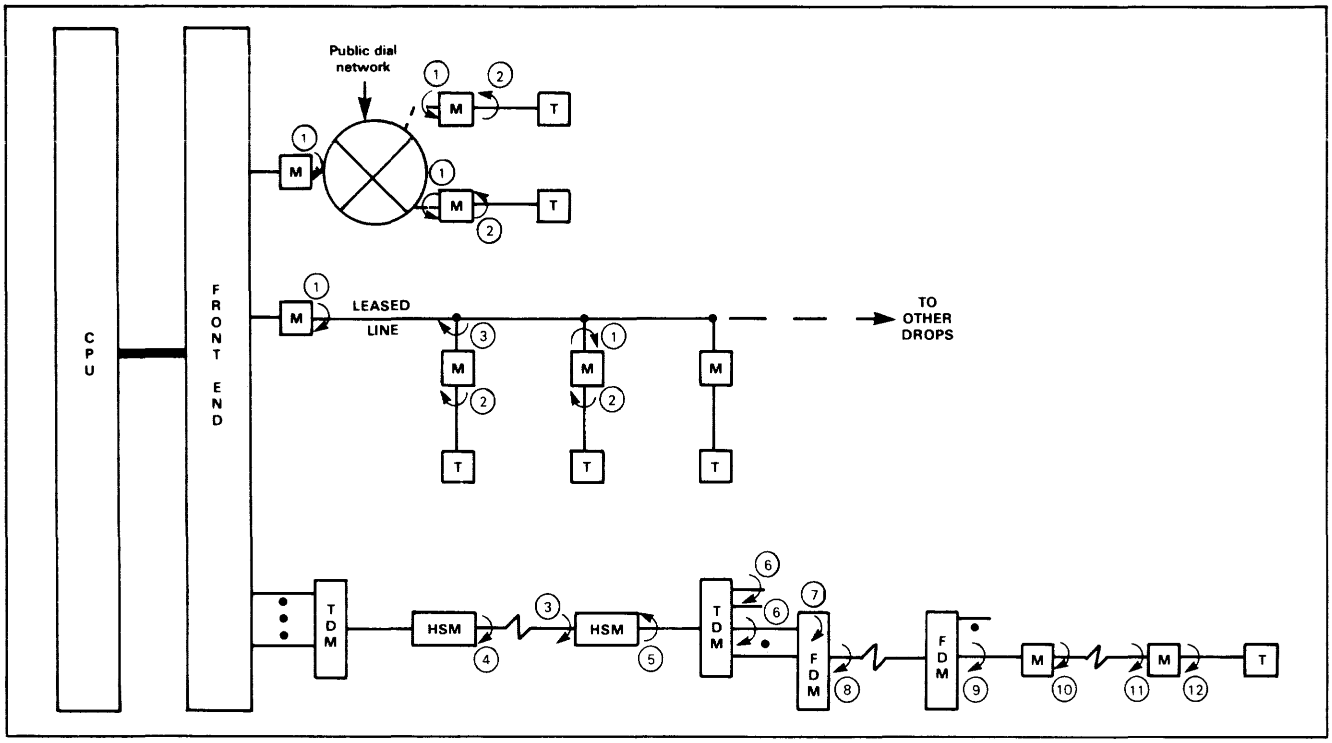
Local Analog Loopback. Shown in the diagram as ①, this loops the modem transmitter back to its own receiver. This is usually accomplished via local control and only tests the near-end modem.

Remote Digital Loopback. Shown in the diagram as ②, this test loops the modem receiver back to its transmitter for data analysis at the distant end. When initiated manually at the distant end, it is similar to the TEST function that was incorporated in the early AT&T modems. The BELL test board would be in voice contact with the site that was complaining about service and the customer would throw the TEST switch; the test board operator would then transmit a test pattern to the remote site where it would be returned for analysis. This saved the BELL repair people unnecessary service calls if the trouble was really in the terminal. Today's modems have an automatic feature to control this loopback. This test obviously disrupts traffic; the terminal is usually terminated in its characteristic impedance while the test is under way.

Line loopback. Shown in the diagram as ③, this test terminates the distant modem and turns the line around (through any necessary impedance/level matching equipment) so that operation over the transmission facility can be verified exclusive of the remote modem. This is also a disruptive test.

High-speed Analog loopback. Shown in the diagram as ④, this test loops back the composite output from a time division multiplexer (TDM) through a high-speed modem at the analog interface. This test checks only the high speed modem at the computer site and is disruptive of all channel traffic going through the TDM.

Network Diagnostic Tools: An Equipment Vendor Survey



Loopback Test Schematic

High-speed Digital Loopback. Shown in the diagram as ⑤, this test loops back the composite stream at the digital interface of the distant high-speed modem. This test is disruptive of all TDM channel traffic.

Remote Low-Speed Channel Loopback. Shown in the diagram as ⑥, this test confirms operation of only the channel under test at the digital interface of that channel in the remote city. This test is disruptive to that channel's traffic only.

Remote Analog Loopback-Level One. Shown in the diagram as ⑦, this test verifies operation of the particular channel involved at the near end Frequency Division Multiplexer (FDM) before all the channels are combined for transmission over the Level Two line. This test is only disruptive to that channel's traffic.

Remote Composite Analog Loopback. Shown in the diagram as ⑧, this test loops the composite voice frequency signal eminating from the near end FDM. All FDM and associated TDM channels are disrupted.

Remote Digital Loopback-Level Two. Shown in the diagram as ⑨, this test is similar to test 2 , except that it occurs at the digital interface of the affected FDM channel, which disrupts traffic on that FDM/TDM channel only.

Remote Analog Loopback-Level Three. Shown in the diagram as ⑩ , this test checks the modem shown and is similar to test ⑦. Only that channel is affected.

Remote Line Loopback-Level Three. Shown in the diagram as ⑪ , this test is similar to test 3 , and only affects the traffic on that channel.

Remote Digital Loopback-Level Three. Shown in the diagram as ⑫ , this test is similar to test ②and only that channel is affected.

Obviously tests of this extent require intelligence all the way through the system as control signals are passed with integrity through or to each element. In some cases, intelligence could reside in the remote terminal, which upon recognizing the appropriate address and control signal, will cause the associated modem to enter the requested test mode. Systems as complex as this often require outboard control units whose sole function is to interpret commands and cause the necessary action to take place, as in the more elaborate network management systems.

Line Analysis

Strictly speaking, these tests are performed using analog measuring/generating equipment and are conducted on audio facilities—most commonly telephone lines. Tests of this nature measure all types of line characteristics that the reader may or may not be familiar with (e.g., frequency response, signal/noise ratio, envelope delay distortion, amplitude delay distortion, hits, dropouts, fades, frequency translation, impulse noise, etc.). The hardware required for these tests is sophisticated and the operator must be skilled at interpreting the results. Similar hardware is located at various carrier test centers and should only be required at a users site when extremely troublesome lines are encountered and proof is needed to convince the carrier that action need be taken. Usually, digital instrumentation, while not providing data accurate enough to pinpoint the problem, is sufficient to identify that there is a problem. If some audio capability is required, a low-cost line monitor with a speaker, test tone generator, and power level meter will usually suffice.

Breakout Boxes

These extremely useful devices are absolutely necessary in troubleshooting and in identifying interface/timing incompatibilities. This capability is part of the majority of more sophisticated devices but the breakout box can be purchased separately. It usually consists of two connectors (one for connection to the business machine equipment (DTE)) and one for connection to the communications facility (DCE). The device is connected in series between the two elements with the status of the various leads visible through indicators (usually Light Emitting Diodes, LED's). Some of the units use tri-state LED's (red, green, off) to show if the interface lead is high, low, or off. In addition, the leads are generally brought out to pin jacks to allow connection to external test equipment. Most of these units derive power from the attached devices and are non-interfering with data and control signal flow between the DTE and DCE.

Error Rate Testers

The two most commonly used devices of this type are Bit Error Rate Testers (BERT) and Block Error Rate Testers (BLERT). They usually generate a known bit or block pattern (with most that are currently available, user selected test data can be generated as well as standard "fox" messages and 63, 511, and 2047-bit pseudorandom patterns). The device monitors what is returned to it (usually through one of the loopbacks previously described) and provides a numerical readout of the BER or BLER over a specified time frame. This is a good overall test of the transmission facility and hardware. It gives a reasonable indication that the problem is random, solid, or occurs periodically in bursts.

Data Line Monitors

These devices require some knowledge on the part of the user but an engineering degree is certainly not required. The function of these units is to bridge the connection between a DTE and DCE on a non-interfering basis to display data and events as the connected devices interact. The display is usually a CRT and an optional printer port is sometimes provided (not the printer). These units always have internal storage and many are equipped with an integral tape unit, or the means to connect to an external unit, in addition to semiconductor memory. Many of these devices can operate full-duplex with the send and receive data differentiated on the screen by reverse video, underlining, or some other method. Most of these units are capable of operating on a trap (a selected character or group of characters in the data stream or a control lead event). This conserves the storage media and allows only the data/events of interest to be captured for analysis. The unit can usually capture a predefined amount of occurrences/data both before and after the trap trigger to assist in analysis.

Data Line Monitor/Simulators

This equipment has all the attributes of the monitors previously described plus the ability to emulate a DTE or a DCE for off-line, interactive, testing of the counterpart. These units are particularly useful when installing new hardware, adapting to new control formats, and debugging software. The operation of the device can be made true or false for the application to allow determination of the behavior of the unit under test. They are invaluable aids when new protocols, such as SDLC or X.25, are to be implemented because all of the anticipated difficulties can be simulated and the results observed (e.g., improper zero-bit insertion; errored Frame Sequence Check, wrong packet length, incompatible timings, erroneous address or control fields, etc.). Since all testing occurs off-line, problems can be corrected without wasting valuable resources on-line only to discover a "glitch."

Network Control Systems

As discussed briefly earlier in this report, these systems are the most sophisticated offerings available today in the data communications test field. Many are capable of controlling/monitoring hundreds of lines with 50 or more drops per line. Each point in the network carries its own address, which is periodically scanned for proper operation. Each of the remote points has electronics associated with it that are devoted to the diagnostic function. Signalling back to the central site occurs over a secondary channel, usually out-of-band to the primary data path but normally sharing the same transmission facilities. The status of the EIA interface between the remote modem and terminal is continuously monitored and reported back to the central site as are conditions such as terminal failure, modem failure, line degradation, signal quality, transmit and receive levels, etc. A failed terminal or modem

Network Diagnostic Tools: An Equipment Vendor Survey

USER'S RATINGS OF DATA COMMUNICATIONS TEST, MONITOR, AND CONTROL EQUIPMENT

Manufacturer and Model	Number of User Re-sponses	Number of Units in Use	Overall Satisfaction					Ease of Use					Hardware Reliability					Maintenance Service					Software/ Firmware Flexibility (if applicable)				
			WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P	WA	E	G	F	P
Atlantic Research—DTM	3	4	3.3	1	2	0	0	2.3	0	1	2	0	4.0	3	0	0	0	3.3	1	2	0	0	3.0	1	1	1	0
Intershake	3	3	3.7	2	1	0	0	2.7	1	0	2	0	3.7	2	1	0	0	3.3	2	0	1	0	4.0	3	0	0	0
Others & unspecified	9	10	3.8	7	2	0	0	3.6	6	2	1	0	3.6	5	3	0	0	2.8	1	3	2	0	2.7	1	1	0	1
Subtotals	15	17	3.7	10	5	0	0	3.1	7	3	5	0	3.7	10	4	0	0	3.1	4	5	3	0	3.2	5	2	1	1
Digi-Log, all models	5	6	3.2	1	4	0	0	4.0	5	0	0	0	2.6	0	3	2	0	3.0	0	5	0	0	2.0	0	0	1	0
Digitech, Pacer	3	3	3.7	2	1	0	0	3.0	1	1	1	0	3.0	0	3	0	0	3.3	1	2	0	0	3.0	1	1	1	0
Epicom, all models	3	3	4.0	3	0	0	0	4.0	3	0	0	0	3.7	2	1	0	0	3.0	0	1	0	0	—	—	—	—	—
Halcyon—803	5	5	3.8	4	1	0	0	3.4	2	3	0	0	3.6	3	2	0	0	3.5	2	2	0	0	3.5	2	2	0	0
Others & unspecified	3	4	2.7	1	1	0	1	2.7	1	1	0	1	2.3	0	1	2	0	1.7	0	0	2	1	3.0	1	0	1	0
Subtotals	8	9	3.4	5	2	0	1	3.1	3	4	0	1	3.1	3	3	2	0	2.7	2	2	2	1	3.3	3	2	1	0
Hewlett-Packard, 1640	3	6	4.0	3	0	0	0	3.3	1	2	0	0	4.0	3	0	0	0	3.0	1	0	1	0	3.3	1	2	0	0
International Data Sciences— 1310	3	3	4.0	3	0	0	0	4.0	3	0	0	0	4.0	3	0	0	0	3.3	1	2	0	0	—	—	—	—	—
Others & unspecified	3	3	4.0	3	0	0	0	3.5	1	1	0	0	3.5	1	1	0	0	3.0	0	2	0	0	—	—	—	—	—
Subtotals	6	6	4.0	6	0	0	0	3.8	4	1	0	0	3.8	4	1	0	0	3.2	1	4	0	0	—	—	—	—	—
Racal-Milgo, 220	3	5	3.3	1	2	0	0	3.3	1	2	0	0	3.7	2	1	0	0	3.5	1	1	0	0	—	—	—	—	—
Sierra, all models	3	4	3.7	2	1	0	0	4.0	3	0	0	0	3.7	2	1	0	0	4.0	1	0	0	0	—	—	—	—	—
Spectron—301	4	5	3.3	1	3	0	0	3.5	2	2	0	0	3.3	1	3	0	0	2.8	1	1	2	0	2.0	0	1	0	1
501	9	10	3.6	5	4	0	0	3.4	4	5	0	0	3.4	4	5	0	0	3.0	3	2	3	0	3.0	3	2	1	1
502	16	17	3.7	11	5	0	0	3.5	9	6	1	0	3.3	8	5	3	0	2.8	3	8	0	0	3.5	7	7	0	0
601	14	32	3.6	8	6	0	0	3.5	8	5	1	0	3.0	4	6	4	0	3.0	5	4	3	1	2.8	2	0	1	1
Others & unspecified	15	18	3.6	10	4	1	0	3.6	10	4	1	0	3.3	9	4	0	2	2.5	0	10	1	3	2.3	1	0	1	1
Subtotals	58	82	3.6	35	22	1	0	3.5	33	22	3	0	3.3	26	23	7	2	2.8	12	25	9	7	3.1	13	10	3	4
All others	26	62	3.8	19	6	0	0	3.7	19	6	1	0	3.6	17	7	1	0	3.2	10	4	5	1	3.7	8	3	0	0
GRAND TOTALS	133	203	3.6	87	43	1	1	3.5	80	41	10	1	3.4	69	47	12	2	3.0	33	49	20	9	3.2	31	20	7	5

LEGEND: E—Excellent; G—Good; F—Fair; P—Poor; WA—Weighted Average based on a weighting of 4 for Excellent, 3 for Good, 2 for Fair; and 1 for Poor.

can be removed from the polling table and operation can continue; a streaming terminal can be disabled; a failed modem can be backed up automatically by remote control; a failed line can be dialed around, etc.

The remote electronics can be part of the modem circuitry, a separate card in the modem enclosure, or a completely separate standalone device. The central site hardware usually consists of a minicomputer with control console, storage, backup facilities, and master modems. The power of these systems is enormous and prices are falling periodically.

USER EXPERIENCE

In September 1979, Datapro included a Reader Survey Form in the monthly supplement to DATAPRO REPORTS ON DATA COMMUNICATIONS. By the editorial cutoff date of November 15, we had received over 130 usable responses covering over 200 pieces of equipment. It is our normal practice to include these survey forms with the DATAPRO 70 supplements as well as mailing to Data Communications subscribers. This dual mailing naturally results in greatly increased circulation and a correspondingly high return in the number of responses. However, Datapro felt that the particular nature of this equipment could best be reported on by communications subscribers, at least on the first edition of this report; therefore, DATAPRO 70 readers were not solicited this time. Considering the limited exposure that this survey received, and also the fact that this is a brand new report, we are gratified with the response. We hope that once this report is disseminated, next year's results will be even better.

Usage Patterns

In addition to the equipment ratings that we requested of our subscriber base, Datapro also used the survey form to ask some questions regarding usage. A question concerning the length of time the equipment had been in service revealed the following:

Less than one year .... 48 responses.
One to two years.................................. 48 responses.
Two to four years .... 30 responses.
Over four years.....................................7 responses.

When asked if there were any plans to replace this equipment with newer, more sophisticated hardware within the next 12 months, 93 respondents said No and 33 replied Yes. A query regarding cost justification for the item met with the following results:

Fully cost justified .... 75 responses.
Essential, regardless of cost .... 49 responses.
Marginally satisfactory .... 4 responses.
Expectations unfulfilled .... 1 response.

The majority of those responding indicated that the equipment was used on their own systems as opposed to customer or client systems. Responses were equally divided as to whether the equipment was used at more than one location. The capability to monitor remote locations using centrally located hardware was indicated by 91 users while another 33 operated locally only. The type of facility the equipment was associated with can be broken down as follows:

leased—118; switched—65; DDS—27; other—14 (some were using more than one type of facility; hence the apparent disparity in the totals).

Of those responding, 112 indicated that the chief use of the equipment was specific problem solving and 77 said that the primary function was monitoring; only 20 users identified preventative maintenance as the main usage. Another question concerned the applications on which the hardware was used, with results as follows:

Monitoring/recording .... 108 responses.
DTE/DCE simulation .... 46 responses.
Patching/control .... 25 responses.
Distortion analysis .... 21 responses.
Backup switching/control .... 4 responses.

Several other uses were mentioned, including hardware/software design and debugging. The specific tests conducted included Bit Error Rate (43), Block Error Rate (13), and distortion (17). The information garnered from these proceedings included: data transmission performance (105), control event timing and interaction (84), and protocol adherence and sequencing (89). The majority of users (88) employed their own data as source material while another 19 used the standard "fox" message.

User Reaction

The accompanying chart lists the ratings assigned by those responding to this survey. Where three or more responses were obtained on any particular model, that model is identified. Where less than three responses were received on any particular model, the equipment was lumped and identified as "others or unspecified" under the vendor's name. When a manufacturer did not receive at least three responses, the ratings were added and identified under the All Others category.

COMPARISON CHARTS

The charts at the end of this report present the key characteristics of over 100 test devices. The information contained in these charts was supplied and/or verified by the manufacturers during November and December 1979; their cooperation with the Datapro staff is greatly appreciated.

Datapro sent repeated requests for information to over 50 firms known or believed to be in the data communications test equipment market. The absence of any company from the charts means that the company either failed to respond, was unknown to us, or chose not to be listed. There are many manufacturers of components, such as patch panels, meters, EIA A/B switches, etc. that are not listed in the charts because their offerings do not constitute complete test entities. These vendors are no less important to the field but listing them here is not practical due to sheer volume. The vendor that you choose to implement a turnkey diagnostic system will be aware of these suppliers and no doubt will use some of them in assembling the hardware. It would be wise to ask for the identification of these "second tier" suppliers if you decide to purchase/lease a turnkey system. A listing, complete with address and telephone number, of all vendors represented in the charts immediately precedes the charts.

The comparison charts cover a wide range of equipment and therefore the "stub" entries or characteristics are not applicable to all hardware. Most of these stubs are self explanatory; however, abbreviations have been used freely due to space limitations. The following defines the stub entries and the more commonly used abbreviations found in the charts.

Functional Description The information found here gives the configuration of the device (e.g., rack mount, portable, standalone, etc.) and defines the general use of the equipment.

DTE/DCE emulation refers to the capability to simulate a Data Terminal Equipment or a Data Communications Equipment device to the counterpart under test. On-line, "bridged" monitoring means that the device can be connected between the DTE/DCE to monitor event and data flow without interfering with the prime data stream or the interaction between DTE and DCE.

When "fox" message is referred to, it means the standard message traditionally used to test telegraph/data terminals and transmission links: The quick brown fox jumps over the lazy dog 0123456789 (notice that all letters of the alphabet are included in the sentence).

Reversals generation refers to alternating Mark and Space tones or binary 1010's (RYRY characters in Baudot code or consecutive U characters in ASCII).

Network Diagnostic Tools: An Equipment Vendor Survey

A trap trigger is either a specific character or sequence of characters in the data stream, or a selected event on an interface control lead, that causes the test device to store, display, capture, or otherwise react to the occurrence in a predefined manner.

BERT (Bit Error Rate Test) and BLERT (Block Error Rate Test) refer to the capability of generating/recognizing specific bit/block patterns and providing a means to read out the numerical performance results in errored bits/blocks per specified time period.

User programmable, as normally defined in the data processing industry, refers to a general purpose computer and the ability of the user to create general purpose software including alterations to the operating system. As used in the data communications test environment, user programmable means that the user can select certain actions, routines, sequences, timing, etc. to occur under selected conditions in the order specified; the routines are usually contained in ROM and as such are not alterable. The equipment has been designed specifically for data communications test applications and is therefore not general purpose. Because the equipment was procured only for the test environment and the user has control over sequencing, timing, etc., most individuals consider the hardware user programmable even though the operating system is fixed.

Facilities used on refers to either the facility to be tested (e.g., 3002 voice grade telephone channel) or the point of connection for the test (e.g., RS-232-C digital interface).

Display type details the type of indicators provided as integral parts of the device; e.g., CRT, analog meter, speaker, LED (Light Emitting Diode), printer port (interface port only, customer must supply printer), etc.

Microprocessor based is answered yes or no: when yes, the model number of the component is usually provided.

Maximum bit rate generated (measured) indicates the highest speed, in bits per second, at which a digital device can operate; async. (asynchronous) or sync. (synchronous) mode is specified.

Frequency response for analog devices only; given in Hz (Hertz or cycles per second).

Power requirements specifies internal/external operating power.

Interfaces supported specifies the connecting arrangement by industry standard, specification, or recommendation; loop refers to current driving power at the interface in mA (milliamperes).

TEST, MONITOR, AND CONTROL EQUIPMENT VENDORS

For your convenience in obtaining additional information, the following list contains the full names, addresses, and telephone numbers of the vendors whose products are listed in the comparison charts.

Astrocom Corporation, 120 West Plato Boulevard, St. Paul, MN 55107. Telephone (612) 227-8651.

Atlantic Research Corporation, 5390 Cherokee Avenue, Alexandria, VA 22314. Telephone (703) 642-4000.

Codex Corporation, 20 Cabot Boulevard, Mansfield, MA 02048. Telephone (617) 364-2000.

Columbia Data Products, Inc., 9050 Red Branch Road, Columbia, MD 21045. Telephone (301) 992-3400.

Com/Tech Systems, 44 Beaver Street, New York, NY 10004. Telephone (212) 425-0733.

Datacomm Management Sciences, 181 Main Street, Norwalk, CT 06851. Telephone (203) 838-7183.

Dataproducts New England, Inc., Barnes Park North, Wallingford, CT 06492. Telephone (203) 265-7151.

DEI Teleproducts, 563 North Citracado Parkway, Escondido, CA 92025. Telephone (714) 743-8344.

Digi-Log Systems, Inc., Babylon Road, Horsham, PA 19044. Telephone (215) 825-9550.

Digitech Data Industries, Inc., 66 Grove Street, Ridgefield, CT 06897. Telephone (203) 438-3731.

Dynatech Data Systems, 7644 Dynatech Court, Springfield, VA 22153. Telephone (703) 569-9000.

Epicom, Inc., 592 North Douglas Avenue, Altamonte Springs, FL 32701. Telephone (305) 869-5000.

Gandalf Data, Inc., 1019 South Noel Avenue, Wheeling, IL 60090. Telephone (312) 541-6060.

General DataComm Industries, Inc., One Kennedy Avenue, Danbury, CT 06810. Telephone (203) 797-0711.

Halcyon, 1 Halcyon Plaza, 2121 Zanker Road, San Jose, CA 95131. Telephone (408) 293-9970.

Hekimian Laboratories, Inc., 15825 Shady Grove Road, Rockville, MD 20850. Telephone (301) 948-8855.

Hewlett-Packard Corporation, 690 East Middlefield Road, Mountain View, CA 94042. Telephone (415) 969-0880.

Infotron Systems, Cherry Hill Industrial Center, Pin Oak Lane and Olney Avenue, Cherry Hill, NJ 08003. Telephone (609) 424-9400.

Network Diagnostic Tools: An Equipment Vendor Survey

International Data Sciences, Inc., 7 Wellington Road, Lincoln, RI 02865. Telephone (401) 333-0640.

Intertel, Inc., 6 Vine Brook Park, Burlington, MA 01803. Telephone (617) 273-0950.

Navtel Limited, 8481 Keele Street, Unit 12A, Concord, Ontario L4K 1B1. Telephone (416) 669-9918.

Paradyne Corporation, 8550 Ulmerton Road, Largo, FL 33540. Telephone (813) 536-4771.

Questronics, Inc., 3565 South West Temple #5, Salt Lake City, UT 84115. Telephone (801) 262-9923.

Spectron Corporation, 344 New Albany Road, Moorestown, NJ 08057. Telephone (609) 234-5700.

T-Bar Incorporated, 141 Danbury Road, Wilton, CT 06897. Telephone (203) 762-8351.

Teleprocessing Products, Inc., 4565 East Industrial Street, Building 7K, Simi Valley, CA 93063. Telephone (805) 522-8147.

TRAN Telecommunications Corporation, 2500 Walnut Avenue, Marina Del Rey, CA 90291. Telephone (213) 822-3202.

Universal Data Systems, 4900 Bradford Drive, Huntsville, AL 35805. Telephone (205) 837-8100.□

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Astrocom Corp. Minichek	Astrocom Corp. Maxichek	Atlantic Research Corporation (ARC) Intershake Model DTM-2	Atlantic Research Corporation (ARC) Interview II (SAM-1/SAM-2)
FUNCTIONAL DESCRIPTION:	Portable test set, capable of generating/recognizing: 511 (2047 opt.) pseudo-random pattern, steady mark, steady space, and reversals. The interface can be monitored via test points	Portable data analyzer. Capable of generating/recognizing: 511 pseudo-random pattern (with/without start/stop bits), "fox" message in ASCII (EBCDIC opt.) reversals, or up to 50 8-bit user sel. char. Data trap on sel. event-HEX readout	Portable or rack mounted configuration; used as on-line monitor and diagnostics device or as off-line DTE/DCE emulator; interactive testing at the digital interface	Portable or rack mounted configuration; used as an on-line monitor and diagnostics device; monitors send and receive traffic with trap triggers and freeze features; 1024-character display; highlights and reverse video; compatible w/most protocols
Facilities used on:	Digital interface connecting to modem (DCE)	Digital interface connecting to modem (DCE)	Digital interface connecting to DTE, DCE, or bridged between	Digital interface in a bridged connection between DTE and DCE elements
Display type:	LED	LCD (4-position numerics)	LED's; optional CRT, printer and video ports	CRT; LED's; optional printer and video ports
Microprocessor based:	No	Yes (1802)	No	No
Maximum bit rate generated:	2400 bps, async; 24K bps, sync.	9600 bps, async; 19.2K bps, sync.	64K bps, async./sync.	9600 bps, async./sync.
Maximum bit rate measured:	2400 bps, async; 24K bps, sync.	9600 bps, async; 19.2K bps, sync.	256K bps, async./sync.	9600 bps, async., 62K bps sync.
Frequency response:	—	—	—	—
Power requirements:	Two internal batteries	Two internal batteries	115/230 VAC, 50/60 Hz, (Int. battery to support RAM)	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C; V.24	RS-232-C; V.24	RS-232-C; V.24; MIL-188. Options for RS-366; RS-449; AT&T 300 series; V.35; loop current	RS-232-C; V.24
TYPICAL CURRENT USER'S APPLICATION:	Field service; end-user maintenance; used by telephone companies	Field service; end-user maintenance; used by telephone companies	Field service; end-user diagnostics; particularly useful in large network performance evaluation and troubleshooting	Interpretation of hardware/software problems in large networks (particularly in a polled environment)
PRICING AND AVAILABILITY: Purchase price:	$295-$357	$895	$16,000 (depends on configuration	$4,900
Lease price, including maintenance:	Not offered	Not offered	Third party	Third party
Warranty & maintenance policy:	180 days; factory repair/return	180 days; factory repair/return	1 yr., factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	14	14	30	30
Serviced by:	Astrocom Corp.	Astrocom Corp.	ARC	ARC
No. of U.S. service locations:	Two	Two	Factory	Factory
Date of first delivery:	—	—	1976	1978
Number installed to date:	—	—	500	300
COMMENTS:	Astrocom also manufactures modems, couplers, & modem eliminators	Astrocom also manufactures modems, couplers, modem eliminators	Can be used independently or in conjunction with ARC's Interview offerings	Can be used independently or in conjunction with ARC's Intershake offering

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Atlantic Research Corporation (ARC) Interview 3000	Atlantic Research Corporation (ARC) DQA-10	Atlantic Research Corporation (ARC) TMG-10	Atlantic Research Corporation (ARC) DMS-3/4
FUNCTIONAL DESCRIPTION:	Portable or rack mounted configuration; used as an on-line monitor and diagnostics device; similar to Interview II except 3000 is microprocessor based with more power & storage	Portable or rack mounted configuration; generates 5, 6, 7, or 8 level async. traffic; provides in-service measurement of distortion, speed errors, parity errors	Portable or rack mounted configuration; generates 5, 6, 7, or 8 level async. test messages consisting of two characters established by front-panel switch settings with controlled distortion to 45%	Portable test instrument for adjusting loop current, bias, clutch, etc. on mechanical teleprinters
Facilities used on:	Digital interface in a bridged connection between DTE and DCE	20/60 mA neutral current loops; 10-100 MA polar loops; RS-232-C; MIL-188	Dry, neutral loops standard (up to 260 VDC, 100 mA); others optional	20-150 mA neutral loops
Display type:	CRT; LED's; optional printer and video ports	Analog meter	— (Front panel switches establish op. mode)	Analog meter
Microprocessor based:	Yes (three)	No	No	No
Maximum bit rate generated:	19.2K bps async./sync.	—	2400 bps, async; 9 standard speeds (up to 50K bps opt.)	—
Maximum bit rate measured:	64K bps async./sync.	50K bps, async. (9 standard speeds)	—	1200 bps, async.
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	Internal battery; 115 VAC, 60 Hz opt.
Interfaces supported:	RS-232-C; V.24; MIL-188. Options (future) for RS-449; AT&T 300 series; V.35; X.21; X.25	RS-232-C; V.24; MIL-188; 0-100 mA polar loops; 20/60 mA neutral loops	Dry, neutral current loops; RS-232-C; MIL-188; TTL	20/60 mA neutral loops; RS-232-C; MIL-188
TYPICAL CURRENT USER'S APPLICATION:	— (Projected: operation with ARC's Intershake to provide add'l programming capability—storage)	Low-speed telegraph-type loop signal distortion analysis; field or central site	Low-speed, telegraph-type loop signal generation with controlled distortion introduced	Field adjustment of mechanical teleprinters
PRICING AND AVAILABILITY:				
Purchase price:	$7,900	$1,600	$1,750	$625
Lease price, including maintenance:	Third party	Third party	Third party	Third party
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	60 (after availability)	30	30	30
Serviced by:	ARC	ARC	ARC	ARC
No. of U.S. service locations:	Factory	Factory	Factory	Factory
Date of first delivery:	January 1980 (projected)	1967	1967	1967
Number installed to date:	30 (back orders)	2,800	3,200	2,800
COMMENTS:	Can be used independently or in conjunction with ARC's Intershake offering; optional tape unit available	Can be used independently or in conjunction with ARC's TMG-10 offering	Can be used independently or in conjunction with ARC's DQA-10 offering	Can be used independently or in conjunction with ARC's TMG-3/4 offerings

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Atlantic Research Corporation (ARC) TMG-3/4	Atlantic Research Corporation (ARC) DATA TECH 9600	Atlantic Research Corporation (ARC) PMK04	Atlantic Research Corporation (ARC) IFT-680
FUNCTIONAL DESCRIPTION:	Portable test instrument that generates any two 5, 6, 7, or 8 level characters—independent of each other. Byte pattern established by front-panel switches	Portable or rack mount configuration; used as an on-line monitor bridging the DTE/DCE interface; includes generator and analyzer sections; capable of bit/block error rate testing	Portable test instrument; used to test async. DTE—particularly terminals; full duplex, split speed (gen. and rec. sections can operate at different speeds	Portable or rack mount configuration; used as an on-line monitor and diagnostics device; pulse traps detect and store transitions
Facilities used on:	Low-speed, neutral current loops	RS-232-C; V.24; MIL-188; current loop interface at up to 100 mA @ 260 VDC	Digital interface to DTE under test	Digital interface between DTE and DCE; can be interactive or a non-interfering bridge
Display type:	—	Analog meter	LED's	LED's
Microprocessor based:	No	No	No	No
Maximum bit rate generated:	600 bps, async.	9600 bps, async.	9600 bps, async. (21 standard rates)	—
Maximum bit rate measured:	—	9600 bps, async.	9600 bps, async.	56K bps, sync.
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	Dry-contact keying (110 mA, 300 VDC max.) neutral loops; RS-232-C; V.24	RS-232-C; V.24; MIL-188; current loop	RS-232-C; V.24; 20 mA current loop	CCITT V.35
TYPICAL CURRENT USER'S APPLICATION:	Field service testing of low-speed asynchronous devices	DTE/DCE simulator and exerciser; signal quality monitoring; AT&T 914 compatible error measurements	Terminal exerciser (DCE emulator); noted for DEC terminals; others possible	On site test of V.35 interface between DTE and DCE
PRICING AND AVAILABILITY:				
Purchase price:	$400	$2,500	$930	$1,175
Lease price, including maintenance:	Third party	Third party	Third party	Third party
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30	30	30	30
Serviced by:	ARC	ARC	ARC	ARC
No. of U.S. service locations:	Factory	Factory	Factory	Factory
Date of first delivery:	1967	1975	1977	Oct. 1979
Number installed to date:	2,800	1,100	860	10
COMMENTS:	Can be used independently or in conjunction with ARC's DMS-3/4 offerings			

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Atlantic Research Corporation (ARC) NCS-100	Atlantic Research Corporation (ARC) Terminal Controller	Atlantic Research Corporation (ARC) LAS	Codex Corporation (Sub. Motorola) ACQMS
FUNCTIONAL DESCRIPTION:	Rack mounted Network Control System; custom design per application; analog and digital monitoring, patching, control, reconfiguration	Rack mounted; central housing for network test apparatus; provides access, monitor/test, and reconfiguration capabilities	Rack mounted; central housing for network test apparatus; provides access, monitor/test, and reconfiguration capabilities	Rack mount central site Automatic Circuit Quality Monitoring System for use with Codex LSI Series modems; monitors up to 64 lines to 12 parameters with alarms and reporting
Facilities used on:	3002 voice grade; wide-band; T1 carrier; digital; RS-232-C	3002 voice grade; wide-band; T1 carrier; digital; RS-232-C	3002 voice grade; wide-band; T1 carrier; digital; RS-232-C	As determined by the Codex modems used (usually 3002 voice grade or equal)
Display type:	CRT; LED's; printer port	CRT; LED's printer port	CRT; LED's printer port	Async. control terminals; LED's; opt. audio monitor, patch panel, and X-Y oscilloscope
Microprocessor based:	Yes	Yes	Yes	Yes
Maximum bit rate generated:	256K bps, sync.	—	—	—
Maximum bit rate measured:	1.544M bps	9600 bps, async., 1.544M bps sync.	9600 bps async., 1.544M bps sync.	9600 bps, sync.
Frequency response:	20 Hz-40K Hz	200 Hz-4K Hz	40 Hz-20K Hz	Data pattern and line char. (300-3000 Hz
Power requirements:	115/230 VAC; 50/60 Hz	115/230 VAC 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C, V.24; V.28; V.35; AT&T 300 Series; T1 carrier	RS-232-C; V.24; V.28; V.35; AT&T 300 Series; T1 carrier	RS-232-C; V.24; V.28; V.35; AT&T 300 Series; T1 carrier	RS-232-C to control printer; non-std. in interface to Codex modems
TYPICAL CURRENT USER'S APPLICATION:	Large, diversified data networks	Projected for use in large data networks, regardless of application	Projected for use in large data networks regardless of application	Any point-to-point (usually multi-circuit) leased line application
PRICING AND AVAILABILITY:				
Purchase price:	Dependent upon configuration—contact vendor	Dependent upon configuration—contact vendor	Dependent upon configuration—contact vendor	$8,750 (four lines) plus $3,150 each add'l 16 lines
Lease price, including maintenance:	Third party	Third party	Third party	Base; per month: $450-1 yr.; $300-2 yr.; $250-3 yr.
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 year
Availability (days ARO):	90-180	180	180	60
Serviced by:	ARC	ARC	ARC	Codex
No. of U.S. service locations:	Factory	Factory	Factory	50 cities, nationwide
Date of first delivery:	1975	Oct. 1980 (projected)	Mar. 1980 (projected)	1978
Number installed to date:	250	—	—	—
COMMENTS:				Tests central site modems (no remotes) for BERT, line conditions, etc. according to user-defined thresholds

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Codex Corporation (Sub. Motorola) MNCS	Codex Corporation (Sub. Motorola) Tech. Control	Columbia Data Products, Inc. Model DB-16	Columbia Data Products, Inc. Model 300C
FUNCTIONAL DESCRIPTION:	Rack mount Multipoint Network Control System; uses out-of-hand secondary channel for diagnostics and control of up to 8 lines with 30 drops/line; most use Codex LSI FP modems	Rack mount central monitoring, switching, and diagnostics; integration of all necessary audio and digital test facilities as required by application	Portable device for field service line testing, monitoring, and storage; generates "fox" message in ASCII; 16K bytes of RAM (battery supported)	Portable or rack mount configuration; used for line testing, monitoring and storage; generates "fox" message in ASCII; digital storage on 3M tape, 1.5M bytes std., 2.25M bytes opt.
Facilities used on:	As determined by Codex modem used; usually 3002 voice grade, or equal	As required	As required by transmission device	As required by transmission device
Display type:	LED's; additional async. terminal	LED's; terminal	LED	LED's
Microprocessor based:	Yes	No	Yes (8080A)	Yes (8080A)
Maximum bit rate generated:	—	—	Std. rates from 110 bps to 19.2K bps, async., ASCII	Std. rates from 110 bps to 19.2K bps, async., ASCII
Maximum bit rate measured:	9600 bps, sync.	9600 bps, sync.	Std. rates from 110 bps to 19.2K bps, async., ASCII	Std. rates from 110 bps to 19.2K bps, async., ASCII
Frequency response:	Data pattern and line characteristics (300-3000Hz)	Data spectrum as required	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50-400 Hz	115/230 VAC, 50-400 Hz
Interfaces supported:	Special interface to central site Codex LSI-FP modems; remotes have special built-in card	All standard; RS-232-C; V.24; AT&T 300 series; 2/4-wire 3002	RS-232-C; V.24; 20 mA loop	RS-232-C; V.24; 20 mA loop
TYPICAL CURRENT USER'S APPLICATION:	Multipoint and point-to-point circuits (usually polled applications)	Multipoint and point-to-point networks	Local, low-speed dial-in data gathering; code and speed conversion for later XMSN to central	Local, low-speed dial-in data gathering; speed conversion for later XMSN to central; remote program loading (IPL)
PRICING AND AVAILABILITY:				
Purchase price:	$4,950	Per configuration; contact vendor	$995	$1,995
Lease price, including maintenance:	Base, per month: $225—1 yr.; $165—2 yr.; $155—3 yr.	Per configuration; contact vendor	Contact vendor	Contact vendor
Warranty & maintenance policy:	1 year	1 year	90 days; factory repair/return	90 days; factory repair/return
Availability (days ARO):	60	60	30	30
Serviced by:	Codex	Codex	Columbia Data Products	Columbia Data Products
No. of U.S. service locations:	50 cities, nationwide	50 cities, nationwide	Factory	Factory
Date of first delivery:	1976	1971	1978	1976
Number installed to date:	150	Over 500	Over 1,000	Over 3,500
COMMENTS:			Auto answer capability for local dial in or remote dump	Auto answer capability for local dial in or remote dump

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Com/Tech Systems LINK/CHEK 202	Datacomm Management Sciences (DMS) ANMACS	Datacomm Management Sciences (DMS) ARM	Datacomm Management Sciences (DMS) FI-Comm 3600
FUNCTIONAL DESCRIPTION:	Two configurations; central site (master) and remote locations (slaves); stand-alone or rack mount; generates/compares 1023-bit pseudo random test pattern with up to 1000 automatic repeats	Rack mount semi-Automatic Network Management And Control System for bridging/breaking VF (3002) or RS-232-C connections; central housing for all test gear	Rack mount semi-automatic Tech Control for monitoring system performance; bridges EIA interfaces; provide central housing for system test equipment	Central site and remote units for diagnosing faults and by-passing failed hardware automatically (from the central site) on IBM 3600 Financial Loop Systems
Facilities used on:	Used in conjunction with modems operating over dial-up, or point-to-point, multi-point leased lines	Per network; 3002 voice grades; current loops	Per network; 3002 voice grades	Analog, leased/private line series loop 3002 network
Display type:	LED's	LED's; associated CRT with keyboard	LED's; associated CRT	LED's; audible alarm
Microprocessor based:	No	Yes (8080, 8085)	No	No
Maximum bit rate generated:	19.2K bps, async., sync.	—	—	—
Maximum bit rate measured:	—	—	—	—
Frequency response:	—	—	—	—
Power requirements:	115 VAC, 60 Hz	115 VAC, 60 Hz	115 VAC, 60 Hz	115 VAC, 60 Hz (central and remote)
Interfaces supported:	RS-232-C; V.24	RS-232-C; 2/4-wire 3002; 20/60 mA loop	RS-232-C	4-wire 3002 (back-up via two dial connections)
TYPICAL CURRENT USER'S APPLICATION:	On-line/off-line BERT testing; EIA interface monitoring; leased or switched lines	Large private network (operates as a Tech Control)	Small to medium networks (operates as a Tech Control)	IBM-3600 Financial Systems (banking, point-of-sale, etc.)
PRICING AND AVAILABILITY:				
Purchase price:	Stand alone: master—$715-1090; slave—$480-630	Depends on configuration; contact vendor	Depends on configuration; contact vendor	Depends on configuration; contact vendor
Lease price, including maintenance:	Not offered	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30	90-120	90-120	90-120
Serviced by:	Com/Tech Systems	DMS	DMS	DMS
No. of U.S. service locations:	Factory	Factory	Factory	Factory
Date of first delivery:	1976	1979	1977	Oct. 1979
Number installed to date:	Over 150	1,000 (channels)	1,500 (channels)	—
COMMENTS:	Testing from central site with remote unattended	Modular components, patch arrangements; alternate routing and monitoring; network status reports; CKT's controlled by KAD	Modular components; touch-pad console	Dial backup for failed components with automatic restoration upon completion of corrective action

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Datacomm Management Sciences (DMS) RTA-327	Datacomm Management Sciences (DMS) SP-25 MV	Datacomm Management Sciences (DMS) TS-5	Datacomm Management Sciences (DMS) TS-6
FUNCTIONAL DESCRIPTION:	Portable or rack mount Response Time Analyzer; used in IBM 3270 and NCR 501 environments; transparently measures component/application delays selectively; ASCII or EBCDIC standard	Rack mount unit to allow complete analysis of RS-232-C interface; non-interfering (bridged) monitoring of all 25 pins by two-color, tri-state LED's and pin jacks; voltmeter included	Portable unit for taking audio power measurements from -50 to +3 dbm; 600/900 ohm impedance (20K ohm-bridging)	Portable unit for taking audio power measurements from -50 to +3 dbm; test tone generator (five frequencies); 600/900 ohm impedance (20K ohm—bridging)
Facilities used on:	IBM 3270 (or equivalent) applications; RS-232-C digital interface (bridged)	RS-232-C interface	Analog (3002 or equal)	Analog (3002 or equal)
Display type:	LED's; printer port	LED's; LCD numeric (voltmeter)	LED; analog meter	LED; speaker; analog meter
Microprocessor based:	Yes (8085)	No	No	No
Maximum bit rate generated:	—	—	—	—
Maximum bit rate measured:	19.2K bps, sync.	20K bps	—	—
Frequency response:	—	—	Voice band	300-3000 Hz
Power requirements:	115/230 VAC, 50/60 Hz	115 VAC, 60 Hz	Internal battery	Internal battery (rechargable)
Interfaces supported:	RS-232-C; printer output port operates to 1200 bps async.	RS-232-C	Two-wire, audio	Two-wire, audio
TYPICAL CURRENT USER'S APPLICATION:	Response time analysis in IBM 3270 (or equivalent) networks	Important element in Tech Control centers	Level measurements; 2-wire private or leased lines; modem output	Level measurement and test generation for testing/aligning VF facilities
PRICING AND AVAILABILITY:				
Purchase price:	$4,875	$1,150	$150	$300
Lease price, including maintenance:	Contact vendor	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30	30	30	30
Serviced by:	DMS	DMS	DMS	DMS
No. of U.S. service locations:	Factory	Factory	Factory	Factory
Date of first delivery:	1978	1977	1977	1977
Number installed to date:	30	20	150	200
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Datacomm Management Sciences (DMS) TS-366	Dataproducts New England, Inc. STEP 21	DEI Teleproducts SAM	Digi-Log Systems, Inc. DLM II
FUNCTIONAL DESCRIPTION:	Portable device for non-interferring monitor of the RS-366 interface between CPU/front end and AT&T's 801 Automatic Calling Unit (ACU) or equivalent	Portable or rack mount device used as a monitor or DTE/DCE simulator; user programmable; HEX keyboard; generates 63, 511, 2047 patterns	Portable voice frequency monitor; speaker, test oscillator, and db meter included	Portable or rack mount unit for passive monitoring of RS-232-C data and control leads
Facilities used on:	Non-interfering bridge on RS-366 interface	RS-232-C; MIL-188; BISYNC, SDLC, HDLC environments	2/4-wire 3002	RS-232-C non-interfering bridge monitor
Display type:	LED's	LED's; CRT; dot matrix	Speaker, analog meter	LED's; CRT
Microprocessor based:	No	Yes (Z80)	No	No
Maximum bit rate generated:	—	64K bps, sync.	—	—
Maximum bit rate measured:	—	64K bps, sync.	—	9600 bps
Frequency response:	—	—	300-3800 Hz	—
Power requirements:	Internal battery (rechargable)	115 VAC, 60Hz	115 VAC, 60 Hz	115 VAC, 60 Hz
Interfaces supported:	RS-366	RS-232-C; MIL-188	2/4-wire 3002	RS-232-C; 20/60 mA loop
TYPICAL CURRENT USER'S APPLICATION:	Field service	Field service; part of Tech Control	Airlines; service bureaus	Field service
PRICING AND AVAILABILITY:				
Purchase price:	$600	$6,400	$695	$3,245
Lease price, including maintenance:	Not offered	Base, per month: $435—1 yr.; $310—2 yr.	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	90 days; factory repair/return
Availability (days ARO):	30	90	30	30
Serviced by:	DMS	Dataproducts, New England	DEI	Digi-Log
No. of U.S. service locations:	Factory	Factory	Factory	Two
Date of first delivery:	1977	1979	1976	1974
Number installed to date:	10	—	—	2,000
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Digi-Log Systems, Inc. DLM III	Digi-Log Systems, Inc. NSM	Digi-Log Systems, Inc. DS/40	Digi-Log Systems, Inc. MBC-1200
FUNCTIONAL DESCRIPTION:	Portable or rack mount monitor and DTE/DCE simulator; BERT tester; 511, 2047, "fox", user defined messages generated; menu selection and video prompting	Rack mount central site diagnostic, switching, reconfiguration, monitoring, and alarm systems	Rack mount configuration; used for RS-232-C interface switching for back-up or resource selection	Rack mount modules for "bus" controller function
Facilities used on:	RS-232-C	Digital and analog	RS-232-C interface	RS-232-C
Display type:	LED's; CRT	LED's; CRT port	LED's	Illuminated switch
Microprocessor based:	Yes (Z80)	Yes (Z80)	No	No
Maximum bit rate generated:	19.2K bps, async., sync.	9600 bps, async., sync.	—	—
Maximum bit rate measured:	19.2K bps, async., sync.	19.2K bps, async., sync.	—	19.2K bps, async., sync.
Frequency response:	—	—	—	—
Power requirements:	115 VAC, 60 Hz	115 VAC, 60 Hz	Operating power derived from NSM or other housing	115 VAC, 60 Hz
Interfaces supported:	RS-232-C	RS-232-C	RS-232-C	RS-232-C
TYPICAL CURRENT USER'S APPLICATION:	Field service; central site	Network monitor to reduce down time	Back-up and resource switching	Part of Tech Control center
PRICING AND AVAILABILITY:				
Purchase price:	$3,000	Configuration dependent (approximately $300/channel)	Configuration dependent (approximately $300/switch module)	$375 (approx.)
Lease price, including maintenance:	Not offered	Contact vendor	Contact vendor	Contact vendor
Warranty & maintenance policy:	90 days; repair/return	1 yr.; repair/return	1 yr.; repair/return	1 yr.; repair/return
Availability (days ARO):	90	90	90	90
Serviced by:	Digi-Log	Digi-Log	Digi-Log	Digi-Log
No. of U.S. service locations:	Two	Two	Two	Two
Date of first delivery:	1980 (projected)	1975 (Europe)	1975 (Europe)	1975 (Europe)
Number installed to date:	—	50	800	25
COMMENTS:		Also manufactures line drivers, modem eliminators, modem sharing units, etc.	Also manufactures other Tech Control and transmission equipment	Used with Digi-Log SAM modules

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Digi-Log Systems, Inc. SAM/1	Digitech Data Industries ENCORE-100	Digitech Data Industries PACER-103	Dynatech Data Systems Dyna-Test 1500
FUNCTIONAL DESCRIPTION:	Rack mount modules for monitor, control and alarm functions	Monitor; DTE/DCE simulator	Monitor; DTE/DCE simulator	Portable or rack mount data line monitor
Facilities used on:	RS-232-C	Packet-switching environment; HDLC, X.25, and other bit-oriented protocol applications	—	RS-232-C
Display type:	LED's	LED's; CRT	Gas discharge; LED's	CRT
Microprocessor based:	No	Yes (8085)	Yes (8080)	Yes (8085)
Maximum bit rate generated:	—	19.2K bps, sync.	20K bps, async., sync.	—
Maximum bit rate measured:	19.2K bps, async., sync.	19.2K bps, sync.	20K bps, async., sync.	100K bps, sync., half duplex
Frequency response:	—	—	—	—
Power requirements:	Derived from NSM or other housing	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115 VAC, 60 Hz
Interfaces supported:	RS-232-C	RS-232-C std. others available	RS-232-C; V.35; 20/60 mA	RS-232-C; V.24; MIL-188; V.35; AT&T 300 Series
TYPICAL CURRENT USER'S APPLICATION:	Component of Tech Control systems	Banking; airlines; utilities; universities; multi-drop polled networks; dial-up applications	Banking; airlines; utilities; universities; government; multi-drop polled networks; dial-up applications	Data link monitoring
PRICING AND AVAILABILITY:				
Purchase price:	Approximately $300/ channel	$19,500	$8,500	$5,100
Lease price, including maintenance:	Contact vendor	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	90	30	30	90
Serviced by:	Digi-Log	Digitech	Digitech	Dynatech
No. of U.S. service locations:	Two	Factory	Factory	Factory
Date of first delivery:	1975 (Europe)	June 1979	1976	—
Number installed to date:	500	20	825	—
COMMENTS:	Used in conjunction with other Digi-Log offerings		Pacerscope optional	

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Dynatech Data Systems Dyna-Test 2000	Epicom, Inc. EPIVIEW 100/101	Epicom, Inc. EPIVIEW 110/111	Epicom, Inc. EPITAPE 200/201
FUNCTIONAL DESCRIPTION:	Portable or rack mount interactive test device	Portable or rack mount line monitor	Portable or rack mount line monitor	Desk top or rack mount diagnostic recording unit; captures full duplex data and up to six control lead events per byte
Facilities used on:	RS-232-C	RS-232-C	RS-232-C	RS-232-C
Display type:	CRT	LED's; CRT	LED's; CRT	LED's
Microprocessor based:	Yes (8080, 8085)	No	No	No
Maximum bit rate generated:	9600 bps	—	—	—
Maximum bit rate measured:	9600 bps (full duplex; 19.2K bps (half duplex)	100K bps; (50 to 9600 bps standard)	60K bps; (50 to 9600 bps standard)	19.2K bps; (50 to 9600 bps standard)
Frequency response:	—	—	—	—
Power requirements:	115 VAC, 60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C; MIL-188	RS-232-C; others optional with external adapters	RS-232-C; others optional with external adapters	RS-232-C; others optional with external adapters
TYPICAL CURRENT USER'S APPLICATION:	Simulation and monitoring	Line monitoring; full duplex, asynchronous or byte synchronous	Line monitoring; full duplex, asynchronous, byte orbit synchronous (SDLC)	Data and control recording independent of line discipline or code
PRICING AND AVAILABILITY:				
Purchase price:	$9,950	$3,500	$4,500	$5,750
Lease price, including maintenance:	Not offered	2-yr. lease $175/mo.	2-yr. lease $225/mo.	2-yr. lease $285/mo.
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return; $200 yr. after warranty	1 yr.; factory repair/return; $200 yr. after warranty	1 yr.; factory repair/return; $300 yr. after warranty
Availability (days ARO):	60	30-60	30-60	30-60
Serviced by:	Dynatech	Epicom, Inc.	Epicom, Inc.	Epicom, Inc.
No. of U.S. service locations:	Factory	Factory	Factory	Factory
Date of first delivery:	1978	1977	1978	1977
Number installed to date:	—	250	150	450
COMMENTS:		Model 100 is desk-top unit; Model 101 is rack mount	Model 110 is desk top unit; Model 111 is rack mount	Model 200 is desk top unit; Model 201 is rack mount

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Epicom, Inc. EPISOLVER 400	Epicom, Inc. EPIBERT 320	Gandalf Data Systems TTS 400	Gandalf Data Systems TTS 400C
FUNCTIONAL DESCRIPTION:	Stand alone data communications analyzer for asynchronous byte or bit synchronous networks	Stand alone multi-function error rate tester; BERT, BLERT, % distortion; RTS/CTS delay; XMT/RCV db levels; clock rate measurements	Portable Bit Error Rate Tester	Portable Bit Error Rate Tester
Facilities used on:	RS-232-C	RS-232-C; 2/4-wire 3002 analog	RS-232-C	RS-232-C
Display type:	LED's; CRT; 6-digit counter	LED's; 12 segmented numerical displays	LED digital readout	LED digital readout
Microprocessor based:	No	No	No	No
Maximum bit rate generated:	—	100K bps (50 to 9600 bps standard)	9600 bps, async. (2047-bit pattern)	19.2K bps, async.
Maximum bit rate measured:	100K bps; (50 to 9600 bps standard)	100K bps; (50 to 9600 bps standard)	56K bps, sync.; 9600 bps async.	56K bps sync.; 19.2K bps async.
Frequency response:	—	3002 (300-3000 Hz)	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115 VAC, 60 Hz	115 VAC, 60 Hz
Interfaces supported:	RS-232-C; others optional with external adapters	RS-232-C; others optional	RS-232-C; V.24	RS-232-C; V.24
TYPICAL CURRENT USER'S APPLICATION:	Full duplex monitoring of ASCII EBCDIC (plus two optional codes) networks	Simultaneous measurement and display of 12 different line functions (both analog and digital)	Checkout of all RS-232-C compliant data links	—
PRICING AND AVAILABILITY:				
Purchase price:	$5,250	$3,200	$785	To be determined
Lease price, including maintenance:	2-yr. lease $265/mo.	2-yr. lease $160/mo.	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return; $200/yr. after warranty	1 yr.; factory repair/return; $150/yr. after warranty	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30-60	30-60	30	30 (after availability)
Serviced by:	Epicom, Inc.	Epicom, Inc.	Gandalf	Gandalf
No. of U.S. service locations:	Factory	Factory	Five	Five
Date of first delivery:	Sept. 1979	June 1979	1975	1980 (projected)
Number installed to date:	50	25	300	—
COMMENTS:	Optional 64K bit memory			

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	General DataComm Industries, Inc. BERT-901	General DataComm Industries, Inc. NETCON-2	General DataComm Industries, Inc. NETCON-5	Halcyon 520B2/521A
FUNCTIONAL DESCRIPTION:	Portable or rack mount data transmission test set for bit or block error rate testing	Rack mount central site hardware; stand-alone remote hardware or integral to GDC modems	Rack mount central site hardware; stand-alone remote hardware; can be used in multi-level distributed processing networks	Portable or stand-alone analog test set for checking parameters specified by AT&T Pub 41009
Facilities used on:	Synchronous digital facilities	Single level polled modem networks	Single and multi-level polled modem networks	3002 voice grade switched or leased
Display type:	LED's, three sets of numerics	LED's	CRT, LED's	Dual numeric display; CRT
Microprocessor based:	No	Yes (6800)	Yes (8086, 8039)	No
Maximum bit rate generated:	1.544M bps, sync.	1800 bps async.; 9600 bps sync.	1800 bps async.; 9600 bps sync.	—
Maximum bit rate measured:	1.544M bps, sync.	1800 bps async.; 9600 bps sync.	1800 bps async.; 9600 bps sync.	—
Frequency response:	—	—	—	100 Hz-4K Hz
Power requirements:	115 VAC, 60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C; V.24; V.28; V.35; AT&T 300 Series; TI carrier; MIL-188	Digital-RS-232-C, V.24/V.28; analog—4-wire 3002 voice grade	Digital—RS-232-C, V.24/V.28; analog—4-wire 3002 voice grade	2-/4-wire 3002 voice grade
TYPICAL CURRENT USER'S APPLICATION:	Error performance measurement for equipment, systems and facilities	Diagnostic testing in multi-drop polled network	Surveillance and test in multi-level, multi-drop, polled network	AT&T DATEC testing
PRICING AND AVAILABILITY:				
Purchase price:	$2,400	Central site with 32 lines and 300 drops $52,000	Central site with 32 lines and 300 drops $225,000	$10,000
Lease price, including maintenance:	Not offered	3-yr. lease $2,100/mo.	Described above, 3-yr. lease $9,900/mo.	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30-60	45-60	90-120	30
Serviced by:	General DataComm	General DataComm	General DataComm	Halcyon
No. of U.S. service locations:	12	12	12	8
Date of first delivery:	1974	1977	1977	1975
Number installed to date:	500	12 systems	25 systems	2,000
COMMENTS:		In-band diagnostics channel; used with GDC modems	Out-of-band, non-interferring diagnostic channel, used with any modem	Model 521A is CCITT version

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Halcyon 545A	Halcyon 546A	Halcyon 547A	Halcyon 701A/702A
FUNCTIONAL DESCRIPTION:	Portable phase/amplitude jitter test set; measures phase/gain hits, dropouts, impulse noise	Portable 107 type signal source	Portable 107 type line test set (measuring unit)	Portable transmission test set; level, frequency, noise, and notched noise measurements
Facilities used on:	3002 voice grade switched or leased	DDD and private switched networks	DDD and private switched networks	3002 voice grade leased
Display type:	LED's; digital meters	—	Dual numeric	Dual numeric
Microprocessor based:	No	No	Yes (6800)	No
Maximum bit rate generated:	—	600 bps, async. (internal modem)	—	—
Maximum bit rate measured:	—	600 bps, async. (internal modem)	—	—
Frequency response:	300-3000 Hz	100-4000 Hz	100-4000 Hz	50-20,000 Hz
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	Internal battery; 115 VAC 60 Hz
Interfaces supported:	2-/4-wire 3002 voice grade	2-wire voice grade; RS-232-C for modem	2-wire voice grade	2-/4-wire 3002 voice grade
TYPICAL CURRENT USER'S APPLICATION:	AT&T DATEC testing	AT&T DATEC testing	AT&T DATEC testing	Telex and private leased or switched networks
PRICING AND AVAILABILITY:				
Purchase price:	$3,395	$1,500	$4,595	$1,395
Lease price, including maintenance:	Not offered	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30	30 (after release)	30 (after release)	30
Serviced by:	Halcyon	Halcyon	Halcyon	Halcyon
No. of U.S. service locations:	8	8	8	8
Date of first delivery:	Dec. 1979	1Q, 1980	1Q, 1980	1976
Number installed to date:	—	—	—	2,500
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Halcyon 704A	Halcyon 802A	Halcyon 803A	Halcyon 804A
FUNCTIONAL DESCRIPTION:	Portable transmission test set; level, frequency, noise, notched noise, and 3-level impulse noise measurements	Portable data monitor with terminal polling capability; block, frame and character error testing; 19 counters, 2 timers	Portable, programmable monitor and DTE/DCE simulator; block frame and character error testing; 511 bit pseudo random pattern generator fox message generator trapping capability	Portable data recorder; protocol and code transparent; records on selected pattern matches, or under EIA lead control, or by time of day
Facilities used on:	3002 voice grade switched or leased; wideband	RS-232-C	RS-232-C	RS-232-C
Display type:	Dual numeric	LED's; CRT	LED's; CRT; printer port	LED's; CRT
Microprocessor based:	No	Yes (6800)	Yes (6800B)	Yes (6800B)
Maximum bit rate generated:	—	19.2K bps async., sync.; independent DTE & DCE rates	19.2K bps async., sync.	19.2K bps async., sync.
Maximum bit rate measured:	—	56K bps sync.; 19.2K bps async.	19.2K bps async., sync.	56K bps async., sync. HDX; 19.2K bps async., sync. FDX
Frequency response:	50-110,000 Hz	—	—	—
Power requirements:	Internal battery, 115 VAC, 60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	2-/4-wire 3002 voice grade; wideband	RS-232-C; V.24; X.21; 20/60 mA loop; optional RS-449	RS-232-C; V.24; 20/60 mA loop	RS-232-C; V.24; 20/60 mA loop
TYPICAL CURRENT USER'S APPLICATION:	Telco and private leased or switched networks	Field service and end users	Field service and end users	Field service and end users
PRICING AND AVAILABILITY:				
Purchase price:	$2,245	$6,895	$10,995	$6,795
Lease price, including maintenance:	Not offered	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30	30-45	30-45	30-45
Serviced by:	Halcyon	Halcyon	Halcyon	Halcyon
No. of U.S. service locations:	8	8	8	8
Date of first delivery:	1977	Dec. 1979	1977	Sept. 1979
Number installed to date:	1,000	—	350	50
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Hekimian Laboratories, Inc. Model 3900	Hewlett-Packard Corporation Model 1645A	Hewlett-Packard Corporation Model 3551A	Hewlett-Packard Corporation Model 4940A
FUNCTIONAL DESCRIPTION:	Portable or rack mounted analog test facility to verify transmission performance to AT&T Pub 41004, 41009	Portable or rack mount data error analyzer; simultaneously runs BERT and BLERT tests and checks for carrier loss, clock slips, skew, and jitter	Portable transmission test set; measures levels, noise and frequency on voice grade circuits	Portable or rack mount transmission impairment measuring set; tests levels, frequency, noise, delay, jitter, hits, dropouts, etc. per AT&T Pub 41009
Facilities used on:	3002 voice grade	Multipoint and point-to-point leased lines or dial-up 3002	3002 voice grade	3002 voice grade
Display type:	Dual digital displays	LED's	LED's	LED's
Microprocessor based:	No	No	No	No
Maximum bit rate generated:	—	9600 bps async.; up to 5M bps sync. with ext. clock	—	—
Maximum bit rate measured:	—	5M bps sync.	—	—
Frequency response:	40 Hz—20K Hz	—	40 Hz—60K Hz (opt. to 80K Hz)	4100 Hz
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz; internal battery	115 VAC, 60 Hz
Interfaces supported:	RS-232-C; V.24; V.28; V.35; 2-/4-wire 3002	RS-232-C; V.24; V.35; AT&T 300 Series; MIL-188; future opt. RS-449	2-/4-wire 3002 voice grade	4-wire 3002
TYPICAL CURRENT USER'S APPLICATION:	Telephone companies and end users	Field service and end users	Field service and end users	Used by carriers and end users
PRICING AND AVAILABILITY:				
Purchase price:	$11,000	$2,800	$2,100	$9,900
Lease price, including maintenance:	Not offered	Contact vendor	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; HP service center repair/return	1 yr.; HP service center repair/return	1 yr.; HP service center repair/return
Availability (days ARO):	90	60	45	70
Serviced by:	Hekimian	Hewlett-Packard	Hewlett-Packard	Hewlett-Packard
No. of U.S. service locations:	Factory	10	10	10
Date of first delivery:	1976	1973	1974	1974
Number installed to date:	—	—	—	—
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Hewlett-Packard Corporation Model 4943A	Hewlett-Packard Corporation Model 4944A	Infotron Systems TE620	International Data Sciences, Inc. Model 60
FUNCTIONAL DESCRIPTION:	Portable transmission impairment measuring set; tests levels, frequency, noise, delay, and jitter	Portable transmission impairment measuring set; tests levels, frequency, noise, delay and jitter; tests all C and D level conditioning parameters	Portable or rack mount monitor and DTE/DCE simulator; BERT, BLERT testing; pseudo random pattern generator; bias measurement to 49 percent	Portable, hand-held EIA breakout panel and monitor
Facilities used on:	3002	3002	RS-232-C	RS-232-C
Display type:	LED	LED	LED	LED's
Microprocessor based:	Yes (6800)	Yes (6800)	No	No
Maximum bit rate generated:	—	—	19.2K bps async.; sync.	—
Maximum bit rate measured:	—	—	19.2K bps async., sync.	—
Frequency response:	To 4100 Hz	To 4100 Hz	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	Internal battery, 115 VAC 60 Hz
Interfaces supported:	4-wire 3002	4-wire 3002	RS-232-C, V.24, V.28	RS-232-C, V.24
TYPICAL CURRENT USER'S APPLICATION:	Carriers and end users	Carriers and end users	End user	Field service, end user
PRICING AND AVAILABILITY:				
Purchase price:	$7,700	$7,700	$2,495	$240
Lease price, including maintenance:	Not offered	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; HP service center repair/return	1 yr.; HP service center repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	70	70	90	30
Serviced by:	Hewlett-Packard	Hewlett-Packard	Infotron Systems	IDS
No. of U.S. service locations:	10	10	Factory	Factory
Date of first delivery:	1978	1978	Sept. 1979	1974
Number installed to date:	—	—	50	20,000
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	International Data Sciences, Inc. Model 65/60	International Data Sciences, Inc. Model 70	International Data Sciences, Inc. Model 80	International Data Sciences, Inc. Model 1310
FUNCTIONAL DESCRIPTION:	Portable hand-held bit error rate tester and EIA breakout box; 63, 511 or 2047 bit patterns, steady mark; steady space; reversals	Portable interface monitor for AT&T 300 series modems	Portable interface monitor for CCITT V.35	Portable test set & BERT BLERT testing with 63, 511, or 2047-bit pseudo random pattern generator
Facilities used on:	RS-232-C	AT&T 300 Series	V.35	RS-232-C; V.24; V.35; AT&T 300 Series; MIL-188
Display type:	LED's	LED's	LED's	LED's; NIXIE'S
Microprocessor based:	No	No	No	No
Maximum bit rate generated:	1200 bps, async., 56K bps sync.	—	—	9600 bps, async.; sync.
Maximum bit rate measured:	1200 bps async., 56K bps sync.	—	—	200K bps sync.
Frequency response:	—	—	—	—
Power requirements:	Internal battery, 115 VAC, 60 Hz	Internal batteries	Internal batteries	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C; V.24	AT&T 300 Series	V.35	RS-232-C; V.24; V.35; AT&T 300 Series; MIL-188
TYPICAL CURRENT USER'S APPLICATION:	Field service and end user	Field service and end user	Field service and end user	Field service and end user
PRICING AND AVAILABILITY:				
Purchase price:	$850	$1,265	$1,265	$3,325
Lease price, including maintenance:	Not offered	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30	60	60	30
Serviced by:	IDS	IDS	IDS	IDS
No. of U.S. service locations:	Factory	Factory	Factory	Factory
Date of first delivery:	1979	1977	1977	1975
Number installed to date:	600	50	60	1,800
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	International Data Sciences, Inc. Model 1700	International Data Sciences, Inc. Model 3000D	International Data Sciences, Inc. Model 4010	Intertel, Inc. EMS-ONE
FUNCTIONAL DESCRIPTION:	Portable test set to measure polling performance and inbound, outbound errors	Central site hardware for BERT testing of high speed synchronous links, pseudo-random pattern generation of 511, 2047, 32,767 and 65,535 bits	Portable test set used for interactive trouble-shooting; DTE/DCE simulation	Central site network control console to monitor up to 160 lines with up to 40 drops/line; constantly monitors modem and line parameters; alarms on fault
Facilities used on:	RS-232-C	RS-232-C; V.24; MIL-188; V.35; AT&T 300 Series	RS-232-C; V.24; MIL-188	RS-232-C; V.24
Display type:	LED	LED's	LED's; CRT	Gas discharge panel
Microprocessor based:	Yes (8035)	No	Yes (8080, 8039)	Yes (Z80)
Maximum bit rate generated:	9600 bps, async., sync.	56K bps sync.	19.2K bps async., sync.	1200-9600 bps async., sync.
Maximum bit rate measured:	9600 bps async., sync.	56K bps sync.	19.2K bps async., sync.	9600 bps async., sync.
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C; V.24	RS-232-C; V.24; V.35; MIL-188; AT&T 300 Series	RS-232-C; V.24; MIL-188 Series	RS-232-C; V.24
TYPICAL CURRENT USER'S APPLICATION:	Multidrop, polled networks	Test of satellite communications channels	End user testing	Point-to-point and multipoint networks
PRICING AND AVAILABILITY:				
Purchase price:	$1,800	$7,985	$7,500	Base system with display and printer $16,675
Lease price, including maintenance:	Not offered	Not offered	Not offered	2-yr. lease $770/mo.
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	60	60	60	30-60
Serviced by:	IDS	IDS	IDS	Intertel
No. of U.S. service locations:	Factory	Factory	Factory	150
Date of first delivery:	1979	1976	Sept. 1979	1979
Number installed to date:	50	50	25	—
COMMENTS:				Intertel also manufactures a line of modems

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Intertel, Inc. NCS 4000	Navtel Datatest-25	Navtel Datacheck-25	Navtel Datacheck-VF
FUNCTIONAL DESCRIPTION:	Central site network control console to monitor up to 100 lines with up to 40 drops per line; constantly monitors modem and line parameters; alarms, under fault condition	Portable BERT tester and EIA breakout box, 511 and 2047 pseudorandom pattern generator	Portable breakout box with pulse traps	Portable breakout box with VF monitor and pulse traps
Facilities used on:	RS-232-C; V.24	RS-232-C; V.24	RS-232-C; V.24	RS-232-C; V.24
Display type:	LED's	LED's	LED's	LED's
Microprocessor based:	No	No	No	No
Maximum bit rate generated:	1200-9600 bps, async., sync.	2400 bps async., sync.	—	—
Maximum bit rate measured:	9600 bps, async., sync.	2400 bps async., sync.	—	—
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	Internal battery	Internal battery	Internal battery
Interfaces supported:	RS-232-C; V.24	RS-232-C; V.24	RS-232-C; V.24	RS-232-C; V.24
TYPICAL CURRENT USER'S APPLICATION:	Point-to-point and multipont networks	Field service	Field service	Field service
PRICING AND AVAILABILITY:				
Purchase price:	Base system with integral display $7,900	$729	$249	$359
Lease price, including maintenance:	2 yr. lease $318/mo.	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return
Availability (days ARO):	30-60	45	30	30
Serviced by:	Intertel	Navtel	Navtel	Navtel
No. of U.S. service locations:	150	Two	Two	Two
Date of first delivery:	1976	Nov. 1979	Jan. 1979	Oct. 1979
Number installed to date:	Over 150	—	350	35
COMMENTS:	Intertel also manufactures a line of modems			

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Navtel DTS-102	Navtel SDA-103	Navtel TMG-303	Paradyne 4410
FUNCTIONAL DESCRIPTION:	Portable BERT BLERT tester with four selectable pseudorandom patterns with block sizes from 10^2 to 10^7	Portable distortion analyzer; measures bias and end distortion as well as total, early, and late peak	Portable test message generator; fox message with controlled distortion to 45 percent	Central site teleprocessing system that can accommodate up to 39 lines with up to 96 drops/line; monitors channel performance over a secondary non-interfering channel
Facilities used on:	RS-232-C; V.35; AT&T 300 Series	RS-232-C; V.24	RS-232-C; V.24	RS-232-C; V.24; V.27; MIL-188; 4 wire 3002
Display type:	4 digits; LED's	2 digits; LED's	LED	Alphanumeric LED's, CRT
Microprocessor based:	No	No	No	Yes (9980)
Maximum bit rate generated:	19.2K bps async. 500K bps sync.	—	9600 bps async., sync.	110 bps-diagnostic channel
Maximum bit rate measured:	19.2K bps async. 500K bps sync.	9600 bps async., sync.	—	110 bps-diagnostic channel
Frequency response:	—	—	—	—
Power requirements:	115 VAC, 60 Hz	115 VAC, 60 Hz	115 VAC, 60 Hz	115 VAC, 60 Hz
Interfaces supported:	RS-232-C; V.24, V.35; AT&T 300 Series	RS-232-C; V.24; 20/60 mA loop	RS-232-C; V.24; neutral loops to 100 mA	RS-232-C; V.24; V.27; MIL-188; 4-wire 3002
TYPICAL CURRENT USER'S APPLICATION:	Field service	Field service	Field service	Point-to-point and multipoint networks; multilevel
PRICING AND AVAILABILITY:				
Purchase price:	$1,529	$1,459	$1,479	$36,600; configuration dependent; contact vendor
Lease price, including maintenance:	Not offered	Not offered	Not offered	2-yr. lease; configuration dependent, contact vendor
Warranty & maintenance policy:	1 yr.; factory repair/return	1 yr.; factory repair/return	1 yr.; factory repair/return	90 days: factory repair/return
Availability (days ARO):	45	45	30	90
Serviced by:	Navtel	Navtel	Navtel	Paradyne
No. of U.S. service locations:	Two	Two	Two	50
Date of first delivery:	1975	1975	1974	1978
Number installed to date:	275	725	835	Over 50
COMMENTS:				Used with Paradyne modems

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Questronics, Inc. TRTM	Questronics, Inc. 300	Questronics, Inc. 500	Spectron Corporation D-301 Datascope
FUNCTIONAL DESCRIPTION:	Portable response time monitor; computes and displays minimum, maximum, average, and last transaction response time; measures from 0.1-9,999 seconds	Portable response time monitor; computes and displays minimum, maximum, average, and last transaction response time; measures from 0.1-9,999 seconds	Central site performance monitor; modules available for async., bisync., SDLC applications, provides printout of response time, line utilization, time/date stamps, etc.	Portable line monitor with integral 262K bit electronic buffer
Facilities used on:	Optical input coupling permits use with any device that has a light that goes on/off during operation	Optical input coupling or RS-232-C bridged connection	Optical input coupling or RS-232-C bridged connection	RS-232-C for transparent monitoring
Display type:	Four-digit LED	Four-digit LED	Hard copy	LED; 5-inch CRT
Microprocessor based:	Yes (8748)	Yes (8748)	Yes (8748)	Yes
Maximum bit rate generated:	—	—	—	—
Maximum bit rate measured:	—	9600 bps, async., sync.	9600 bps, async., sync.	9600 bps, async.; 72K bps, sync.
Frequency response:	—	—	—	—
Power requirements:	115 VAC, 60 Hz	115 VAC, 60 Hz	115 VAC, 60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	Any device with a light source that goes on/off to indicate mode of operation	Optical coupling, RS-232-C bridged connection	Optical coupling, RS-232-C bridged connection	RS-232-C; others optional
TYPICAL CURRENT USER'S APPLICATION:	Response time measurement for interactive terminals	Response time measurement for async., bisync., multi-drop terminal networks	Response time printout; line utilization statistics, Histogram records	Field service
PRICING AND AVAILABILITY:				
Purchase price:	$950	$1,500	$3,600 to $13,000	$4,900
Lease price, including maintenance:	Not offered	Not offered	Not offered	Contact vendor
Warranty & maintenance policy:	90 days; factory repair/return	90 days; factory repair/return	90 days; factory repair/return	1 yr.; regional center repair
Availability (days ARO):	30	30	60	30
Serviced by:	Questronics, Inc.	Questronics, Inc.	Questronics, Inc.	Spectron
No. of U.S. service locations:	Factory	Factory	Factory	Four
Date of first delivery:	1975	1978	March 1979	1978
Number installed to date:	300	200	25	—
COMMENTS:				Available with tape option

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Spectron Corporation D-501B Datascope	Spectron Corporation D-502B Datascope	Spectron Corporation D-600/D-601 Datascopes	Spectron Corporation D-901 Datascope
FUNCTIONAL DESCRIPTION:	Portable line monitor and analyzer; user programmable	Portable monitor and analyzer; interactive DTE/DCE simulator; user programmable	Portable line monitor; D-601 equipped with integral tape unit	Rack mount or stand-alone central site device used as a monitor, data analyzer, interactive DTE/DCE simulator; storage device; user programmable
Facilities used on:	RS-232-C for transparent monitoring	RS-232-C for transparent monitoring or interactive testing	RS-232-C for transparent monitoring/recording	Digital facilities, compatible with bit-oriented protocols
Display type:	LED's; 5-inch CRT	LED's; 5-inch CRT	LED's; 9-inch CRT	LED's; 9-inch CRT
Microprocessor based:	Yes	Yes	No	Yes (multiple)
Maximum bit rate generated:	—	9600 bps, async., sync.	—	1M bps, sync.
Maximum bit rate measured:	9600 bps, async., 80K bps sync.	9600 bps, async., 80K bps sync.	2000 bps, async. 80K bps, sync.	1.6M bps, sync.
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60Hz
Interfaces supported:	RS-232-C; others optional	RS-232-C; others optional	RS-232-C; others optional	RS-232-C; V.24; V.35; AT&T 300 Series; 20/60 mA loop; X.21; RS-449 (optional)
TYPICAL CURRENT USER'S APPLICATION:	Field service or central site data line monitoring	All data networks, not restricted to industry application or line configuration	All data networks, not restricted to industry applications or line configuration	Advanced networks; bit-oriented protocols serviced
PRICING AND AVAILABILITY:				
Purchase price:	$7,500	$10,800 $12,500 with keyboard $14,900 with tape unit	$4,000 (D-600) $8,000 (D-601)	$24,800
Lease price, including maintenance:	Contact vendor	Contact vendor	Contact vendor	Contact vendor
Warranty & maintenance policy:	1-yr.; regional repair	1-yr.; regional repair	1-yr.; regional repair	1-yr.; regional repair
Availability (days ARO):	30	30	30	60
Serviced by:	Spectron	Spectron	Spectron	Spectron
No. of U.S. service locations:	Four	Four	Four	Four
Date of first delivery:	1976	1977	1975/1974	1979
Number installed to date:	—	—	—	—
COMMENTS:	Available with tape option	Available with tape option; ASCII, EBCDIC or IPARS keyboard also optional		Equipped with keyboard and dual flexible diskette drives

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Spectron Corporation T-511 Tape Unit	Spectron Corporation RASP	Spectron Corporation Conventional Tech Control Center	Spectron Corporation Line Monitor System
FUNCTIONAL DESCRIPTION:	Portable high-speed tape unit capable of storing full-duplex data	Central site Remote Access Switching and Patching system; manual or semi-automatic unattended network control	Central site manually actuated network management system; analog/digital reconfiguration	Central site line selection system
Facilities used on:	Digital facilities for data storage	All common analog and digital facilities	All common analog and digital facilities	RS-232-C
Display type:	LED's	LED's; CRT; optional printer port	LED's	LED's
Microprocessor based:	No	Yes (multiple)	No	No
Maximum bit rate generated:	—	—	—	—
Maximum bit rate measured:	9600 bps, async., 56K bps, sync.	—	—	—
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C	RS-232-C; V.24; V.35; AT&T 300 Series	RS-232-C; V.24; V.35; AT&T 300 Series; RS-449 planned	RS-232-C
TYPICAL CURRENT USER'S APPLICATION:	Standalone data recorder or can be used with Datascope	Remote control management, and reconfiguration of analog and digital facilities	Network control and management	Used in conjunction with Datascope
PRICING AND AVAILABILITY:				
Purchase price:	$5,900	$200-$800/line	$85-$200/line	$230/line (16 line system)
Lease price, including maintenance:	Contact vendor	Contact vendor	Contact vendor	Contact vendor
Warranty & maintenance policy:	1-yr.; factory repair/return	Up to 5-yr; regional repair	Up to 5-yr.; regional repair	1-yr.; regional repair
Availability (days ARO):	30	60-90	30	30
Serviced by:	Spectron	Spectron	Spectron	Spectron
No. of U.S. service locations:	Four	Four	Four	Four
Date of first delivery:	1977	1979	1973	1977
Number installed to date:	—	—	—	—
COMMENTS:	DC-100 tape cartridge	Provides automatic alarms, hard copy of system activity	Turnkey installation	

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Spectron Corporation BDTS	Spectron Corporation D-201 Printer/Monitor	Spectron Corporation ITU-HH24	T-Bar, Inc. Explorer Model TD-12
FUNCTIONAL DESCRIPTION:	Portable Buffered Data Transmission Simulator; programmable test message generator; up to eight patterns selectable to 1024 bytes	Portable async. printer	Portable, hand-held EIA breakout box	Portable or rack mount data link monitor, tester, DTE/DCE simulator; Bit Error Rate Tester; Block error rate tester; "fox" generator
Facilities used on:	RS-232-C	RS-232-C	RS-232-C	RS-232-C
Display type:	LED's	Hard copy	LED's	LED's; CRT
Microprocessor based:	No	Yes	No	Yes (Z80)
Maximum bit rate generated:	9600 bps, async., sync.	—	—	19.2K bps
Maximum bit rate measured:	—	9600 bps, async.	—	19.2K bps
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	Internal battery	115 VAC, 60 Hz
Interfaces supported:	RS-232-C	RS-232-C	RS-232-C	RS-232-C; others available
TYPICAL CURRENT USER'S APPLICATION:	Exercising lines, modems, and controllers	Standalone printer or complement to Datascope	Field service	Monitoring; DTE/DCE exercising; pseudo-random (511) pattern generation
PRICING AND AVAILABILITY:				
Purchase price:	$1,850	$4,500	$195	$5,990—monitor $7,990—monitor & emulator
Lease price, including maintenance:	Contact vendor	Contact vendor	Contact vendor	Not offered
Warranty & maintenance policy:	1 yr.; regional repair	1 yr., regional repair	1 yr.; regional repair	1-yr.; factory repair/return
Availability (days ARO):	30	30	30	30
Serviced by:	Spectron	Spectron	Spectron	T-Bar, Inc.
No. of U.S. service locations:	Four	Four	Four	Factory
Date of first delivery:	1976	1977	1979	—
Number installed to date:	—	—	—	—
COMMENTS:				

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	T-Bar, Inc. Model 5911	T-Bar Inc. DAP-16	Teleprocessing Products, Inc. TP-260	Teleprocessing Products, Inc. TP-270
FUNCTIONAL DESCRIPTION:	Rack mount voice frequency circuit monitor; db level measurement	Rack mount EIA signal alarm panel	Portable data line tester; generates a 1004 Hz tone for testing audio facilities	Central site polled network analyzer for response time measurements
Facilities used on:	Analog 3002 voice grade lines	RS-232-C	Analog; audio	RS-232-C
Display type:	Analog meter and speaker	LED's	Analog meter; speaker	LED's; alphanumeric
Microprocessor based:	No	No	No	Yes (8085)
Maximum bit rate generated:	—	—	—	9600 bps, sync.
Maximum bit rate measured:	—	—	—	9600 bps, sync.
Frequency response:	300-3000 Hz	—	20 Hz to 30K Hz	—
Power requirements:	115 VAC, 60 Hz	115 VAC, 60 Hz	Internal battery	115 VAC, 60 Hz
Interfaces supported:	2-/4-wire 3002 lines, or equal	RS-232-C	2-/4-wire audio facilities	RS-232-C
TYPICAL CURRENT USER'S APPLICATION:	Analog monitor	Part of Tech Control facility	Data line testing	IBM 3270 network testing
PRICING AND AVAILABILITY:				
Purchase price:	$590	Configuration dependent; contact vendor	$280	$2,195
Lease price, including maintenance:	Not offered	Not offered	Not offered	Not offered
Warranty & maintenance policy:	1-yr.; factory repair/return	1-yr.; factory repair/return	1-yr., factory repair/return	1-yr.; factory repair/return
Availability (days ARO):	30	30	14	30
Serviced by:	T-Bar, Inc.	T-Bar, Inc.	TPI	TPI
No. of U.S. service locations:	Factory	Factory	Factory	Factory
Date of first delivery:	—	—	1974	May 1979
Number installed to date:	—	—	300	Over 50
COMMENTS:				Optional printer

Network Diagnostic Tools: An Equipment Vendor Survey

MANUFACTURER AND MODEL	Tran Telecom-munications Corp. M350 Checktran	Tran Telecom-munications Corp. M361 XPRT	Universal Data Systems COMTEST	Universal Data Systems COMTEST 100
FUNCTIONAL DESCRIPTION:	Portable transmission test set; generates 63, 511, 2047 pseudo-random patterns; RTS/CTS delay measurement	Standalone or rack mount protocol tester for X.25, Level 2	Portable or central site monitor, DTE/DCE simulator; "fox" message generator; user entered messages	Portable or central site monitor
Facilities used on:	RS-232-C; V.24; V.35; AT&T 300 Series; MIL-188	X.25 level 2 packet switching	RS-232-C; V.24; 20 mA loop	RS-232-C; V.24; 20 mA loop
Display type:	LED's; optional hard copy	LED's; CRT	CRT	CRT
Microprocessor based:	No	Yes (8070)	Yes	Yes
Maximum bit rate generated:	250K bps async., sync.; external clock to 2M bps, sync.	4800 bps, sync.	9600 bps, async., sync.	—
Maximum bit rate measured:	250K bps async., sync.; external clock to 2M bps sync.	4800 bps, sync.	9600 bps FDX; 19.2K bps HDX; async., sync.	19.2K bps, sync., async., FDX
Frequency response:	—	—	—	—
Power requirements:	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 Hz	115/230 VAC, 50/60 HZ	115/230 VAC, 50/60 Hz
Interfaces supported:	RS-232-C; V.24; V.35; MIL-188; AT&T 300 Series	RS-232-C; V.24	RS-232-C; V.24; 20 mA loop	RS-232-C; V.24; 20 mA loop
TYPICAL CURRENT USER'S APPLICATION:	Transparent monitoring; end-to-end testing	Hardware and software debugging for X.25 applications	Monitor/Simulation	Monitoring
PRICING AND AVAILABILITY:				
Purchase price:	$1,500-up	$15,000	$7,950 with 3500 byte buffer	$4,400; $4,850 with CRT
Lease price, including maintenance:	1-yr.—$105/mo. 3-yr.—$67/mo.	1-yr.—$1,000/mo. 3-yr.—$620/mo.	Not offered	Not offered
Warranty & maintenance policy:	1-yr.; incl. in lease or factory repair	1-yr.; incl. in lease or factory repair	1-yr.; factory repair/return	1-yr.; factory repair/return
Availability (days ARO):	30	60	30	30
Serviced by:	TRAN	TRAN	UDS	UDS
No. of U.S. service locations:	12	12	Factory	Factory
Date of first delivery:	1973	July 1979	1977	Oct. 1979
Number installed to date:	600	10	500	—
COMMENTS:				

Topic Index—Section CS50

General Topic	Report Title	Report No. CS50-	See Also Report
System Performance Evaluation	—How to Measure and Evaluate Computer Response and Turnaround Times in an Interactive Network	-120-101	
	—A Communications System Performance Checklist	-130-101	
Line Optimization	—Differences in Line Optimization Approaches to Short- and Long-Line Systems	-220-101	CS10-310-101; CS10-320-101; CS15-320-101; CS40-410-101;
Special Systems Management	—Communications System Management in a Research and Development Environment	-310-101	
	—Institutional Control of Computing Funds and Resources in a Networking Environment	-320-101	
	—Using a Minicomputer as a Communications/System Management Tool	-330-101	
Applications Management	—Distributed Processing—Before	-510-101	CS20-510-101; CS60-170-101
	—Distributed Processing—After	-510-201	
International Operations	—How to Cut Costs and Improve Service of Your International Telecommunications	-810-101	CS40-510-101
	—Nationalism and Data Communications	-815-101	CS20-150-101

Using a Minicomputer as a Communications System Management Tool

Problem:

If you have ever considered the possibility of using a minicomputer as a management tool for controlling a complex communications system you no doubt will find some of your ideas repeated in this report, which describes the use of an IBM Series 1 processor for just that purpose.

The report describes an actual implementation project. In addition to stating what was done and how it was done, the author discusses the type of environment wherein this type of communications control (use of a "sidestream" processor) is likely to be most effective.

Solution:

Today, numerous advancements in teleprocessing are resulting from rapid technological breakthroughs. However, these positive enhancements in selected areas such as functional capability and lower component costs are often accompanied by such negative factors as more costly skilled labor and management complexity. Thus, it is important to exercise good management discipline in the areas of system planning, system development, and system operation. Technology can also benefit the management system through automated management functions and lower costs for management tools. The use of such management tools almost becomes mandatory if a stable situation is desired in a continually growing telecommunications environment.

One can generally characterize the teleprocessing environment of today as inadequate in real management discipline. However, successful management systems have existed since the early 1960s (e.g., SAGE, Ballistic Missile Early Warning System, NASA Manned Spaceflight Systems, FAA Air Traffic Enroute System, multiple Department of Defense Command and Control Systems, multiple Airline Reservation Systems, etc.) in the teleprocessing area without benefit of today's technology. The difference between then and now appears to be one of clear recognition of objectives and a willingness to pay the associated price.

The prototype management tool, which we call a "sidestream" processor tool (based on use of an IBM Series/1) as discussed in this report, has benefited from past experience. It is also an outgrowth of a study between IBM and a customer where the management objectives and associated price were clearly understood. The management tool was planned, developed, and tested in concert with the application system. Thus it had its own project nature and emerged as an operational tool at the same time as the application system became operational. This tool was designed to automate many of the labor-intensive management activities associated with a large teleprocessing environment. Heretofore, many of these activities

"A sidestream approach using a small processor as a tool for managing communication systems" by J.R. Leach of the IBM Western Region Project Office and R.D. Campenni of the IBM Data Processing Division. From *IBM Systems Journal* Volume Nineteen, Number 1, 1980. Reprinted by permission from *IBM Systems Journal.* © 1980 by International Business Machines Corporation.

Using a Minicomputer as a Communications System Management Tool

were done manually by the data processing organization. The tool was also designed to complement a centralized management organization in a network management center environment, other tools such as digital and analog communications test facilities, product-oriented problem determination procedures, and other generalized operational procedures. This total management approach has put the customer in the position of being an experienced system manager and fully confident that he can satisfy the expanding service level objectives of his corporate organization.

The shortcomings in today's management systems and the realization that automation of management styles may not be solved with a single universal tool is clear from past experience where user requirements demanded diverse hardware products, multiple operating systems, multiple data base access methods, and multiple data communications access methods. As a result, IBM has tested a number of different approaches to communications management tools. Some of these tools are based on the traditional host computer application environment (i.e., mainstream). Another approach is the sidestream processor tool prototype being discussed here. This prototype includes functions as discussed in the following section.

PROTOTYPE DESIGN

A total management system has to address the system planning, system development, and system operation processes. Within these distinct processes there must be components to address problem management, change management, installation management, project management, performance management, accounting management, security management, recovery management, operations management, etc. With the sidestream small processor tool we are focusing on an integrated approach to problem, change, and installation management relative to the systems operation process (Figure 1). These particular areas have special significance because (1) they are major cost elements today for any teleprocessing servicer and (2) there has not been an effective integrated approach to problem and change management.

It should also be noted that a sidestream approach is complementary to a mainstream approach (Figure 2). In this illustration those management functions that are best integrated into the host (i.e., network problem determination application, display exception monitoring facility (DEMF), network performance application, interactive problem control system, etc.) are shown complementing the sidestream approach to form an overall system management approach.

The specific philosophy which guided the design of the prototype is summarized as follows:

1. Addresses all components of a communication system—A communication system consists of all system-related components. We define a component as anything that, when not functioning properly, impacts the end user. Included are terminals, modems, lines, multiplexers, storage devices, processors, system software, and application software used to support a network. Also included are the sources of power, air-conditioning, water chillers, and numerous operational procedures because their failures can also impact the end user.

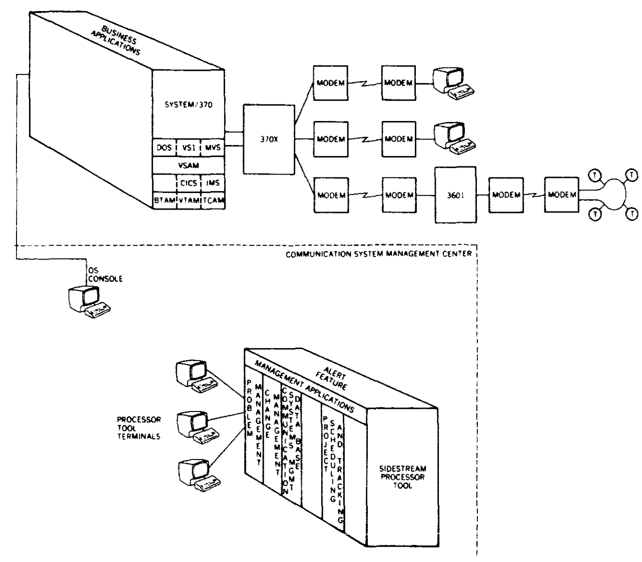


Figure 1. Sidestream approach

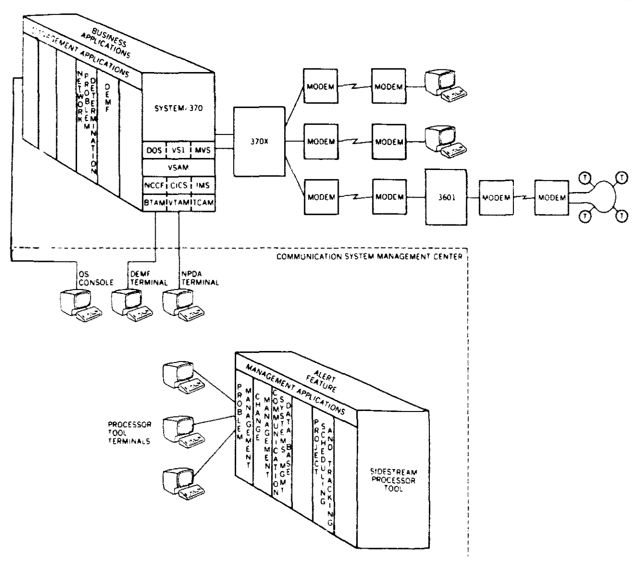


Figure 2. Sidestream approach complementary to mainstream approach

Using a Minicomputer as a Communications System Management Tool

2. Anticipation of problems—Problems must be approached realistically with the realization that they will occur. Procedures should be in place to minimize their impact. This, of course, is true for any computer system. However, when the computer system is contained in one location and a select group of computer operations personnel (e.g., user help, host control, network control, master terminal operator) interface to it, outages can be more easily managed, and the end user can be shielded from a number of problems.

The span of control is greatly increased when terminals are attached to the computer. The computer operation is more visible, and it is more obvious to the end user when things are not working properly.

3. Oriented toward centralized management—The third principle of the design philosophy of the prototype is centralized management. All user problems must be called to one central location. This implies that it is necessary to accept anything the end user wants to categorize as a problem (this is consistent with principle 1). It also implies that the prototype is able to promptly collect the data that the user provides on the problem. The process of recording all problems in one centralized location helps to solve problems. For example, if a telephone line fails, and all the users connected to that line call the central location, the cause of the problem will be fairly evident.

4. Separate system—The problem management system is implemented on a computer separate from the system being managed. By being separate, the sidestream processor tool can be inserted into the communication system operational process without any disruption to the applications. The tool is not dependent on host hardware or software configurations, so the customer can change applications, hardware, or control software without disrupting the tool. The tool can in fact be used to manage such changes. It is available during most system outages. Installation and maintenance of the processor tool are not competing for resources used to run the communication system. Contention for system resources and critical programming skills is avoided. Physical network facilities (lines and modems) can be installed, tested, or reconfigured prior to attachment to the operational host CPU or to a communications multiplexer that has been properly generated for the system. This requires additional test equipment which will be discussed later under Network Control. This capability enables flexible installation/reconfiguration plans, with multiple activities being completed in parallel.

There is another subtle benefit to a separate system. On-line networks frequently have customer service, quality control, or other departments outside the data processing organization responsible for network stability. A separate system provides a tool totally controlled by that organization.

Lastly, there is an element of simplicity associated with a separate system approach. The sidestream processor tool becomes stable after installation and provides reliable service for long periods before there is a component failure. By being separate, it also frees management from addressing interactions that might occur with the business application system.

5. Single data base—The prototype maintains a data base reflecting the current state of the communication system including all problems and changes affecting the communication system. Any user group, customer group within the data processing organization, supplier, or contractor has access to relevant data from this data base. This will prevent each group from having a separate list of problems and priorities.

6. User advocate—A final and key principle of the design philosophy is that the communication system management staff must view itself as a user advocate as opposed to a data processing center advocate. Many of the sidestream processor tool functions are oriented to this attribute. Flexible formats, alerts, scripts, and data base are major examples of the end-user advocate.

SIDESTREAM PROCESSOR TOOL PHYSICAL CONFIGURATION

In order to achieve the sidestream approach design objectives of environment independence, establishment of organizational structure independence, and high tool availability, it was decided to implement the basic management functions on a physically separate and dedicated processor. To achieve the objectives of cost effectiveness, it was decided to use a small processor with expansion capabilities in processor speeds, storage sizes, I/O peripherals, etc. Figure 3 illustrates one potential configuration with three operator stations; however, up to 14 operator stations can be attached to a single processor, if so required. Also to achieve customer productivity through use of lower skill levels (e.g., fewer technicians and no programmers), a turnkey package of control and application software was integrated into diskette format. Customer access to flexible function is achieved via utility functions which allow easy format specification for problem management, change management, configuration management, report generation, alert values, authorizations, etc. Lastly, a predefined starter system was packaged on optional diskettes to allow tool installation and production use within a short period. For example, with the prototype we were able to install a system in as little time as one hour.

Using a Minicomputer as a Communications System Management Tool

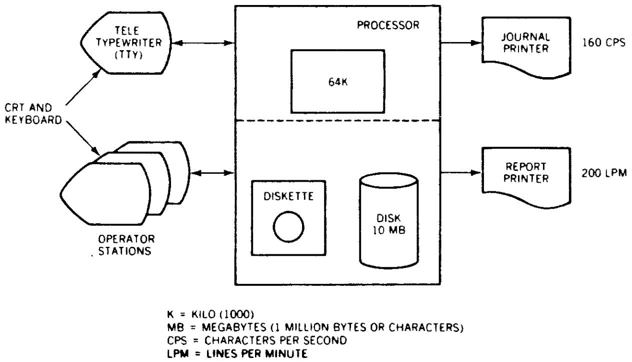


Figure 3. Processor tool components

THE COMMUNICATION SYSTEM MANAGEMENT CENTER

A focal point where all problems are managed is of primary importance to problem recognition. Ideally, a communication system management center is the entity in an organization that provides this focal point. The small processor containing the software for the sidestream processor tool is physically located in this center. Communication system problems, changes, and projects will be managed within this entity. One way to organize this center for communication system management is to have the primary host console and the various system consoles that apply to the communication system located in the center.

The organization and staffing is hierarchical and reflects the four key functions of the center. See Figure 4.

1. User Help is the entry point for user-reported problems. People in this function understand the operation of the end users' terminals and possess good verbal communication skills. They have responsibility to log the reported problems and perform first-level problem determination. They resolve operational problems or identify the correct problem resolver capable of further problem determination. It is helpful if these people are technically conversant but they need not necessarily be technicians. Experience has shown that a large percentage of network problems can be handled by the user help function without assistance from technicians. This relieves the host operator and the teleprocessing experts from the routine phone work they have traditionally done, thereby allowing these highly skilled people to focus on problems requiring their talents.

2. Host Control is the operator or operators who are actually running the on-line systems. Locating this group in the center allows easy interface to all people

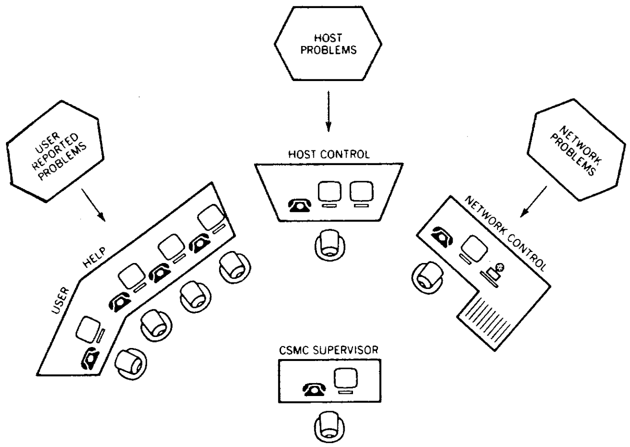


Figure 4. Problem flow in the communication system management center

involved with the communication system. This is important to sustained operation and the coordination of problems.

3. Network Control is responsible for the physical communications network. This function will be discussed in detail later.

4. The Center Supervisor oversees the process of resolving problems. The supervisor is also the destination of management alerts and a safety valve for user complaints.

These functions do not necessarily represent individuals. The number of people in a management center will depend on the size of the communication system and the organizational structure of the data processing center.

PROCESSOR TOOL IMPLEMENTATION

The processor tool is implemented through four major elements:

1. Problem management
2. Change management
3. Project scheduling and tracking
4. Network control

Problem Management

Problems in the communication system are divided into two categories: (1) major outages and (2) service degradation.

The first category is the loss of a major facility, such as the outage of a telephone line, multiplexer, or

Using a Minicomputer as a Communications System Management Tool

CPU. People in the organization usually have the expertise to manage major outages. What is often ignored are individual minor "irritants" that collectively contribute to service degradation. Such irritants include problems with the operation of the hardware and software, and poor performance by operators and by suppliers of various services.

Service degradation and major outages have equal impact on user satisfaction, and evaluating the effectiveness of a communication system exclusively from reports and statistics generated by host programs may result in missing what is actually happening. An evaluation of the performance of all factors involved with system operation must be based on how the end user views the communication system. System degradation is addressed by the processor tool just as effectively as it addresses major outages.

We now discuss four components essential to the processor tool problem management system. They are : (1) a structured approach, (2) alerts, (3) assisting the center operator, and (4) a complete record of problem activity.

When the user calls to report a problem, the user help operator opens a problem record on the processor tool and selects the appropriate problem category. Categories may include host hardware, host software, remote hardware, facilities, line, modems, procedures, the tool itself, etc.

These categories were previously defined by the organization managing the communication system. Unique express formats can also be designed to enhance quick and meaningful data capture. At the time these categories were originally generated, unique data fields were defined for each category. We also note here that decisions can be made to specify attributes for each data field, namely length, use, required versus optional, fixed versus variable length, numeric versus alphanumeric, and default values. These fields will now be displayed to the user help operator to prompt for relevant data.

After recording the data needed from the user, the operator then examines the data base for additional data needed to complete the problem record. If sufficient information is available at this point, the operator will assign the problem to a specific problem resolver. If the correct problem solver is not known at this time, the problem will be assigned to another individual in the management center for further problem determination or to the center supervisor if management action is required.

The steps just discussed could be carried out in a manual system. However, in a manual system paper can get misplaced, problems can go unresolved, and schedules can be missed. To assist in the tracking of problems, the processor tool provides an automatic, three-stage alert facility. The alert is a notification to the center operator and ultimately to management that some situation requires attention.

The first-stage alert notifies the operator that it is time to follow up on the progress of a problem. An alert does not create havoc in managing the communication system. It simply notifies operators that it is time to review the progress of a problem.

A second-stage alert escalates notification to the operator's supervisor if no action was taken by an operator after a first-stage alert. Management is now involved in time to act rather than react. In similar fashion, an automatic third-stage alert escalates notification to another management hierarchical level if no action was taken by either the operator or the next supervisor level. Alerts are also recorded in the problem record history to allow for audit trail analysis.

There are three conditions that will cause alerts: (1) time exception, (2) reopened problem, and (3) excessive reassignment.

A time exception occurs if no action is taken within a designated period of time. If the management center calls for service and the supplier fails to respond within the established time period, a first-stage alert is issued. A second-stage alert is issued if no corrective action is taken within a subsequent designated period of time. Finally, if there is still no corrective action, a third-stage alert is issued.

When a problem is logged in the processor tool as resolved, it is closed. If the same problem reoccurs, it may be reopened. Reopening a closed problem causes a second-stage alert because from the end user's perspective, it is an aged problem, and it should also be worthy of management review. Management attention will help to minimize abuse of the alert feature by premature operator closing action to avoid the alert function.

Excessive reassignment occurs when no one problem solver will take responsibility for a situation. An alert is issued if a problem is reassigned four times.

The gathering of relevant data for problem management is accomplished through system prompts to the operator. If the operator is unsure of what to enter in the data collection fields or who should receive the problem, additional operator prompting is available through a feature called scripts. Scripts are user-defined programs written in a simple English-like language, which run in the

Using a Minicomputer as a Communications System Management Tool

processor tool. When an operator calls a script, text is displayed or questions are presented on the center operator's display terminal. The operator's response determines the path to the next question. More text may be displayed or another question asked. Answers to script-prompted questions may be automatically inserted in the appropriate data collection fields in the problem data collection format. A script may also contain information that could be used to circumvent or even solve the caller's problem. In this mode, the script functions as a problem determination procedure and/or a problem recovery procedure. Scripts may also present a menu of options. The operator selects one of the options, which results in other scripts being invoked. New center operators are apt to become productive more quickly through the use of scripts.

All problems are logged and tracked. No problem record can be deleted. Updating a problem record always adds to the record; it never rewrites any part of the record. The complete record provides data for basic status reports. In addition, the data is gathered and formatted for convenient processing by the host computer. Host data reduction programs can process the information to produce reports for management evaluation and trend analysis.

Change Management

A specific change may or may not be desirable, but any change can be catastrophic if unmanaged. Every installation has a procedure for handling change. The procedure may be an informal one in which there is an oral communication that a change needs to be made and that it should be done by a specific time. If there are no objections, the change is implemented. Most data processing organizations have a change management system with more structure than the informal one just mentioned. Included with a request for a specific change may be an estimation of its priority and the schedule for its implementation. The proposed change is then reviewed. This review process may involve one person or many people at a regularly scheduled meeting.

Change management as implemented in the side-stream processor tool will enhance an existing change management system, not replace it. If a formalized change management system does not exist, the processor tool provides an automated change management system. The functions addressed are:

1. Change Request—A person requesting a change enters the request through the tool. The change request is assigned to a specific category, and the reviewers associated with that category are automatically assigned and notified. The user-defined formats will capture all relevant change data (e.g., description prerequisites, impact, backout procedures, etc.).

2. Prioritization and Scheduling—When the change request is entered in the tool, key operational dates are recorded. These dates indicate when each stage (review, test, implementation) should be completed. Reports will be automatically generated for distribution. Text of specific changes will be generated for individuals responsible for reviewing changes of a given category. A calendar of all planned and implemented changes will be printed for general distribution. This calendar is used to identify conflicts in the schedule and communicate the status of changes in the communication system.

3. Change Review—Reports provided by the processor tool enable changes to be individually reviewed. Approvals and disapprovals are entered into the system. The change requester is automatically notified of disapprovals. If all reviewers approve the change request, those individuals and departments affected by the change will receive printed notification.

4. Implementation—Implementing the change also involves testing and planning fallback procedures. This data should be entered into the change request record. When a change has been tested or implemented, those affected are notified.

5. Alerts—If the review, test, or implementation date specified passes without indicating to the processor tool that the phase was successfully completed, an alert will be automatically issued. These dates may be altered to reflect changing conditions. Thus, alerts tend to establish discipline for the change process.

6. Problem Management and Changes—There is a clear relationship between changes and problems in communication systems management. Resolving problems often entails change. However, changes can cause problems. By maintaining a history of implemented changes, the change management facility becomes an integral part of problem management. A display of selected changes from the change history file is available during problem determination. If a change appears to have caused a problem, the backup or bypass procedures included in the change request record could be an interim solution. A field may be defined in the change record to indicate that this change is a result of a specific problem. A field may also be included in the problem record to indicate that this problem is a result of a specific change. Reports can be generated to indicate the quality of the change review process and cost of particular changes.

Using a Minicomputer as a Communications System Management Tool

Some benefits of the automated approach are that it:

- Provides a tool to facilitate communication to concerned parties.
- Provides a preview of changes to expedite the review process.
- Provides a vehicle to detect potential conflicts in scheduled changes.
- Enhances problem management by providing a central source for information related to changes in the communication system.
- Provides a means to record the effect of changes on the communication system.

Project Scheduling and Tracking

Some communication system changes are simple, one-step procedures; others involve a series of tasks or steps that must be completed before the change is implemented. Changes requiring a series of tasks are projects. Project scheduling and tracking is a tool to assist in more effective project management.

Project management in the general sense may be divided into planning and execution phases. Several tools are available to aid in planning the project. Technicians are available who understand how to execute the plan. What is sometimes poorly handled is the process that connects these two phases: scheduling project tasks and tracking their implementation. This process involves assigning resources and monitoring the tasks to avoid conflicts as discussed below.

Establish a plan—A project plan is developed using the traditional tools. The project tasks are identified, and the duration of each task is estimated. This information is entered into the processor tool. Once the project plan has been entered, a project start date is specified. The system then creates a working plan, identifying the various tasks to be completed by specific dates.

Project control point—A project that involves installing new equipment in the network may require electrical, heating, air-conditioning, and telephone changes, all involving different suppliers. Data indicating the schedule and sequence of these activities is carried in the project record.

When a subcontractor or supplier finishes a task, that information is called to the communication system management center and recorded in the project record. Everyone working on the project will have the most current data available, thus preventing conflicts. If completions are not recorded within the specified time, an alert is automatically issued.

Repetitive projects—Some projects are repetitive. For example, if a new terminal is to be installed in several locations within the communication system or a new version or release of a control problem is to be installed in every communications controller in the communication system, the same project plan may be reused. A master plan is developed and entered in the system as before. Multiple projects can now be created simply by entering a different start date for each new working plan.

The purpose of project scheduling and tracking is to enhance management control. This system will automate the recordkeeping function and provide a single point for contact. The schedules, alerts, and the periodic progress reports generated will give management better control over the progress of a project.

Network Control

As stated earlier, network control is responsible for the physical communications components. In order to clarify this responsibility, it is necessary to make a distinction between physical problems and logical problems. Physical problems are those directly associated with physical components of the network. Examples include problems associated with the CPU, modems, lines, terminals, controllers, etc. Logical problems are those associated with program communication over the functioning physical paths.

Problem determination can be greatly simplified if the physical network integrity has been established. Over the years, many tools have been used to accomplish this. Some of these tools are audio panels, decibel meters, data line test sets, modem test sets, oscilloscopes, Electronic Industry Association (EIA) interface displays, data link character displays, and digital or voice frequency patch panels. Recently, the introduction of intelligent modems has helped remove some of the difficulties in network problem determination.

An intelligent modem has the ability to decode an address and thus recognize messages sent to it, to recognize control data, act upon commands received, and transmit specific status information about itself. Additionally, they typically have self-test capability and are capable of being looped back to the data terminal and/or the communication channel.

Communication with these intelligent modems may be accomplished in one of two ways normally referred to as mainstream or sidestream. Mainstream communication is accomplished in the normal data communication channel. Data intended for the intelligent modem is inserted directly into the flow of data normally intended for the business machine. This data insertion is accomplished by momentarily interrupting the flow of data from the business machine, or by inserting the modem data into gaps in the business machine data. The first technique requires additional code in the communication control program, whereas the second technique requires the ability in the modem to recognize gaps according to the communication protocol being used.

Sidestream communication is accomplished outside of the normal data communication channel. Rather than use techniques that implant the modem status data within the business machine data, analog multiplexing techniques (i.e., frequency multiplexing) are used that allow the modem data and the business machine data to coexist on the same communication carrier. Coexistence is accomplished by utilizing a portion of the bandwidth that is available on a telephone line not required for the business machine data. Sidestream communication usually requires additional hardware such as a microprocessor to control the flow of data (sidestream communication cannot be used in digital transmission systems such as Dataphone® Digital Service).

The data available from these intelligent modems is of immense value when used as a problem determination tool. To list a few examples, it is possible to determine if a terminal is powered on or off by examining the "data terminal ready" lead on the EIA interface that is the interface between the business machine and the modem. By an examination of the "request to send" and "clear to send" leads it is possible to determine if a modem or a terminal is causing a streaming condition. (Streaming is a condition that exists when a multipoint terminal is transmitting continuously. A streaming terminal blocks out communication with all other terminals on a multipoint line.) The state of the "data set ready" lead can indicate whether or not a modem is in a test condition.

Additional problem determination capability exists in the ability to transmit commands to the modems. This ability can cause a modem to self-test and report the results, to transmit and receive predetermined test patterns and detect errors, to switch to stand-by modems, or to loop back either the phone line or the terminal. In the case of a streaming modem or terminal, it is possible to prevent transmission on that modem, thus restoring the rest of the line. Furthermore, all of these tests are controlled in the communication system management center and require no involvement of personnel at remote locations. The fact that these tests can be executed from the management center is even more significant when a remote site is either unattended or staffed by nontechnical personnel.

The recognition of a problem utilizing the intelligent modems, coupled with the problem tracking structure of the processor tool system, has the effect of making both functions more valuable than they are when used separately.

INFORMATION TO SUPPORT THE SIDESTREAM PROCESSOR TOOL

The processor tool maintains a data base reflecting the current state of the network. This data base consists of four files. The content of these files is summarized in Table 1.

In addition to the data required by the tool, each file provides space for information defined by the management of the center. Examples of this information are shown in Table 2.

These files are related by a set of pointers. This relationship is indicated schematically in Figure 5.

The four data base files provide information to support the problem management effort. Each particular file has to be accessible from the reference point of the person who needs the information. Therefore, the locations and circuits are known to the system by more than one name. These names are defined by the communication system management staff.

The location file points to the inventory items at the location, the suppliers who service the location, and the circuits which are connected to the location. The circuit and inventory records point to their associated locations.

If the location is known, any inventory, supplier, or circuit associated with the location can be determined. If the circuits are known, any location on that circuit can be found. If an inventory item is known, it is easy to find the suppliers servicing it by examining

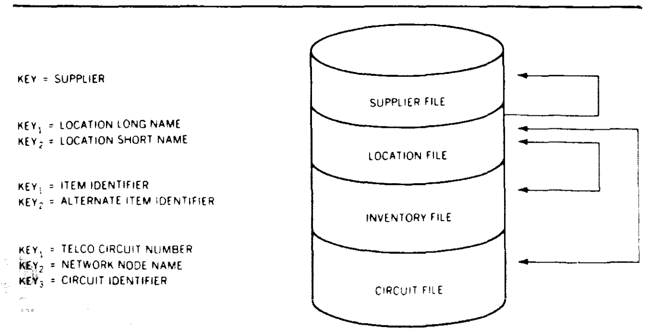


Figure 5. Relationships of files

Using a Minicomputer as a Communications System Management Tool

File	Description
Location	Names each physical location in the communication system
Inventory	Identifies communication system components to be managed
Supplier	Lists the suppliers who service the communication system
Circuit	Describes the communication lines in the communication system

Table 1. Files to support sidestream processor tool

File	Possible Customer-Supplied Information
Location	Primary contact, address, hours of operation, security or after hours phone numbers, etc.
Inventory	Microcode level, diskette volume identification, available back level, etc.
Supplier	Phone numbers, hours of service, dept/person to contact, service reporting procedure etc.
Circuit	Alternate dial backup, line speed, times of availability, protocol, etc.

Table 2. Customer-provided data

the related location record and then the suppliers that service the location.

CONCLUDING REMARKS

Prototype testing experiences have led the authors to the following observations in the specified areas.

1. Management objectives—This is the key element of any management system. Data processing management must clearly state their objectives relative to delivery of data processing services to end users. When these objectives are clearly known, then monitoring for deviations can be performed effectively. Thus, a management system must be capable of recording appropriate information required for management decision making. This requirement necessitates a need for tool flexibility to cater to unique customer requirements. This flexibility is required in the areas of problem management, change management, communication system data base, and report generation. Such flexibility should allow user definition of categories, reports, data fields (i.e., names, sequences, length, required versus optional, numeric versus alphanumeric, fixed length versus variable length).

2. Management discipline—automated tools can contribute to improved management discipline. There are benefits from automatic notification (i.e., alerts) about management events that should have taken place. An automatic notification system contributes to increased problem management discipline because of its increasing hierarchical visibility as it escalates. This escalation activity also contributes to shorter problem resolution times because it minimizes the errors of omission that may result in an undisciplined environment. Lastly, an active notification system contributes to increased data processing credibility with the end-user community. This credibility results from the end user's awareness that the management center operator is working the problem without benefit of additional end-user complaints. In the change management area, active notification will prevent additional costs resulting from poorly coordinated changes in a distributed processing or similar environment. Similarly, an active notification for project scheduling and tracking elements can reduce costs with unnecessary rescheduling associated with suppliers, building contractors, etc.

3. Separate systems—The definition of "sidestream" as discussed in this report is based on physical separation of the management tool from the business application process. It allows data processing management to choose this approach when they might have the following environment:

- An organizational structure that desires complete control of the management tool
- A requirement to have tool availability when the host processing environment is unavailable
- A requirement to have management function without regard to communication system component dependencies (i.e., no hardware/software prerequisites)
- A requirement to have the capability to modify management function without supplier dependencies (e.g., functional enhancements via new software releases from supplier)

4. Productivity—Customer productivity can be considerably enhanced by a management tool that

- Installs quickly and without requirements for highly skilled technical personnel (e.g., system programmers)

Using a Minicomputer as a Communications System Management Tool

- Allows for clerical skill levels and technical skill levels in realistic ratios
- Can be dynamically updated by the management organization without depending on suppliers

5. Cost Effective Tools—Management tools must be cost effective to receive proper consideration. They must be considered relative to the potential teleprocessing growth requirements of the enterprise. Trade-off analyses must be made between manual but labor-intensive efforts and automated tools. Data processing management must be able to articulate the value relative to the communication system management requirements and do the necessary planning relative to the future business environment. Too often this approach is not taken, and automated tools are pursued on an expedited basis only in crisis or disaster mode. In the absence of a crisis or a proper future plan, management usually finds it difficult to justify funding of automated management tools. For example, our prototype testing forced a trade-off analysis between the traditional manual approach and an automated tool as described in this paper which cost less than the services of one person on a life-cycle basis. It is our opinion that this may not be a difficult position for a growth-oriented enterprise that wishes to enhance management discipline.

These observations were consistent with the design philosophy of the sidestream approach. The inherent flexibility of the tool allowed for dynamic adaptation as a result of operational experience with negligible impact to the business application process. This same flexibility allowed for subsequent creation of an IBM predefined starter system for a short installation period at a subsequent test facility with the prototype. Also, the above activity was accomplished without the need for any system programmers or technical knowledge of the sidestream processor. Lastly, this automated tool has helped prototype facilities to achieve the objectives of productivity, management discipline, and resultant end-user satisfaction.

Dataphone is a registered trademark of AT&T.

CITED REFERENCES

[1]R.A. Weingarten, "An integrated approach to centralized communications network management," *IBM Systems Journal* **18**, No. 4, 484-506 (1979).

[2]R.A. Bird and C.A. Hofmann, "Systems management," *IBM Systems Journal* **19**, No. 1, 140-159 (1980, this issue).

GENERAL REFERENCES

These items are available through the local IBM branch office.

Change Tracker, General Information, GH19-6028, IBM Corporation.

Communication Systems Management—Basic Concepts, G320-5840, IBM Corporation (September 1978).

Data Processing Operations—Planning and Control, GG22-9016, IBM Corporation.

DOS/VS Incident Reporting and Tracking, Field Developed Program 5798-CNP, IBM Corporation.

J.B. Ford, "Enhanced problem determination capability for teleprocessing," *IBM Systems Journal* **17**, No. 3, 276-289 (1978).

J.L. Forman, *Problem Management in a Complex Systems Environment,* Technical Bulletin, GG22-9150, IBM Corporation (March 1978).

IBM 3270 Information Display System—IBM 3278 Display Station Problem Determination Guide, GA27-2839, IBM Corporation.

IBM 3270 Information Display System 3274 Control Unit Problem Determination Guide, GA27-2850, IBM Corporation.

IBM 3872 Modem Problem Determination, GA27-3062, IBM Corporation.

Improving Stability in Large Systems (A Management System), GG22-9051, IBM Corporation (June 1978).

Installation Management Brief: MCAUTO Hierarchical Communications in DP Management, GK200864, IBM Corporation (August 1975).

Installation Management Brief: Recovery Structured Walk Through at Marine Midland Bank, GK201120, IBM Corporation (June 1978).

An Introduction to 3790 Network Management, GC20-1853, IBM Corporation (April 1977).

Management Control for Data Processing, GF20-0006, IBM Corporation (April 1975).

Managing the Data Processing Organization, GE19-5208, IBM Corporation (October 1976).

Organizing the Data Processing Activity, GC20-1622, IBM Corporation (October 1973).

OS/VS Online Problem Tracking, Field Developed Program 5798-CNJ, IBM Corporation.

Problem and Change Management in Data Processing—A Survey and Guide, GE19-5201, IBM Corporation (August 1976).

SNA Problem Determination Guide, G320-6016, IBM Corporation.

Tracking and Resolving Problems and Coordinating Changes, Technical Bulletin, GG22-9000, IBM Corporation (June 1975).

VSE/Interactive Computing and Control Facility (VSE/ICCF), General Information, GC33-6066, IBM Corporation. □

CS99-001-051

datapro Communications Management Vol. 2, No. 5, May 1980

Newsbriefs

Let's Hear It For The FCC

The country and the world are going to discover whether or not a giant comfortable monopoly, which had effectively been guaranteed its profits by the federal government; can organize, staff, and operate a profit making arms-length subsidiary in competition with other profit making companies.

We refer of course, to the recent FCC decision permitting both AT&T and GTE to establish such subsidiaries for the purpose of providing customers with data processing equipment and services. Both companies have been given until March 1, 1982 to restructure their activities to conform to the FCC mandate.

Aside from the 2-year restructuring requirement, everything set forth in this decision has been the subject of extensive media coverage, water cooler conferences, and seminar discussions for the past year. Consequently we are somewhat taken aback by the extremely conservative response of both GTE and AT&T. A GTE spokesman said that the company would be unable to assess its impact until they had an opportunity to study the full text of the decision. AT&T expressed concern about the degree of separation required between Bell operating companies, Bell Labs and Western Electric and the "all-too-brief transition period imposed by the action."

Considering the amount of press coverage given this subject we had expected AT&T to have Plans A, B, C, D, etc. ad infinitum in readiness. Instead, it appears that both of these giants had decided to "wait and see" what the FCC would do, and then begin to make plans. Such an approach is not surprising for AT&T in its role as a monopoly; however, if the company had wanted to impress its potential customer base for data communications services, it could have had one or more plans ready, and upon hearing the FCC decision, proceeded to implement the most attractive one, consistent with the commission's decision.

We assume that AT&T recognizes that the new subsidiary will be a totally new kind of business for the company and that it will be operating in a new kind of market. We believe that the degree of success of the AT&T data processing subsidiary will depend upon who constitutes the management of this subsidiary and how much freedom that management is given by the parent company. If the management is drawn directly from AT&T, we believe the degree of success will be less than if it is drawn from experienced executives in the data processing industry.

It will be very interesting to watch how AT&T organizes and staffs this new subsidiary. In recent years there have been numerous examples of large companies expanding into other fields in diverse ways. For example, there was the Northern Telecom acquisition of Sycor and Data 100, both profitable companies which put Northern Telecom in the data processing equipment business. Then we saw ITT acquire Courier, Qume and Shugart which provided the carrier with a complete terminal manufacturing capability. Acquisition is not the only way to do this, however, it is probably the fastest method and the surest when venturing into a new business area. The GTE acquisition of Telenet was another fortunate event for data communications users.

The method used by Exxon to enter new business areas has been to find new high technology entrepreneurs and provide them with venture capital. Some of these organizations literally started out in garages with two or three people. Exxon would spoon feed these operations with what they needed in the way of funds and little more. It was up to the entrepreneur and his hand picked organization to plan, develop, build and sell the product with virtually no interference from Exxon. Three organizations that started in this way are Qwip (facsimile) Vydec (Word processors) and Qyx (Intelligent typewriters). Exxon concluded (and rightly so) that the inhabitants of its diamond studded headquarters tower, were in no position to advise an entrepreneur with the vision of a new office or data product. Exxon's track record using this approach has been excellent.

The acquisition route to forming a subsidiary may not be open to AT&T, although an argument can be made that it should be, in the case of a true arms-length subsidiary. However, regardless of how the subsidiary is constituted,

Note to subscribers: After reading this issue of Newsbriefs and routing it to other interested parties within your organization, please file it under Newsbriefs tab—No. CS99—at the back of Volume 2 of your Communications Solutions service.

we hope that the management takes an aggressive marketing stance as an entity apart from AT&T. Attempts to exploit the relationship with Ma Bell as opposed to innovative service offerings and competitive pricing could, we believe, be disastrous.

One of the more interesting elements of the FCC decision is that "advanced services" (services other than basic communications service) would have to be provided by the subsidiary. Furthermore, the subsidiary will have to acquire its transmission facilities under tariff from a common carrier. This element of the decision requires that Bell's proposed Advanced Communications Service (ACS), be marketed by the subsidiary rather than by Bell. We believe that this is a wise decision by the FCC, and will benefit all users of data communication, but it increases the pressure on AT&T to organize and operate a profit making subsidiary in an open market.

Data processing equipment and services must be *sold* to business. AT&T does not *sell* telephone service to businesses, businesses *buy* it because it is essential to the operation of the business. They buy it from AT&T because they can't get it from someone else. (Where a business *can* get it from someone else, the business buys it from the carrier who provides the best service for the money.) AT&T has never had to sell to business. The "phone-power" and "long distance" campaigns are effective business appeals but they basically consist of institutional advertising—they certainly would not be effective in drawing a company away from a competitor.

There is nothing in the history of the Bell System to indicate that it knows how to sell data communications services. We have heard it said by many individuals on repeated occasions that "AT&T doesn't *really* understand data communications." The implication of that statement is that the management thinking within AT&T is dominated by considerations of voice communications. This is understandable considering that data communication is barely 25 years old and voice contributes so much more to the earnings of AT&T, there are so many more customers for it, and they are so much less demanding than customers for data communications. For the last 10 years however, data communications has been growing by leaps and bounds. If AT&T does not provide the kinds of service required by users of data communications, *someone* is going to. And that someone is going to make a lot of money in the process.

PROCESSORS

● NCR Comten has announced its new 3400 Link Processor System (LPS)

The 3400 is a microprocessor-controlled adaptive multiplexing system designed to extend remote concentration functions into areas of the network where a 3600 Processor would not be economically justifiable. Capabilities include error checking, data compression, statistics and performance monitoring, remote auto-dialing and remote auto-answer, and diagnostics. The system is comprised of the 3400 LPS software, a 3401 Link Controller, and one or more 3410 Link Processors. The 3401 Link Controller is mounted in a Comten 3600 FEP or REP or in a 36X1 Module Controller connected to the 3600 and provides one to four link circuits to remotely located 3410 Link Processors. Each circuit can handle one to eight 3410 Link Processors, and each 3410 can provide concentration capabilities for up to 31 terminal line interfaces; however, the maximum number of terminals that can be supported per 3400 LPS is 31. The 3401 Link Controller may service one 3600 Processor, or be switched to interface between two 3600s. The 3400 LPS is supported under (and requires) Comten's Emulation Processing (EP), Partitioned Emulation Processing (PEP), Network Control Program (NCP), Communications Networking System (CNS), and Data Switching System (DSS) software resident in the 3600 processor. Software for the 3400 LPS is generated under Comten's Network Definition Procedures (NDP) and requires Comten 3600 System Control Software release 62.0.

● Memorex Corporation has introduced a new model in its 1270 Family of hardwired terminal control units

The 1270 Model 8 is designed for small communications users and provides control unit functions which allow an IBM 360/370, 303X, 4300, or equivalent mainframe to communicate with a variety of local and remote data communication terminals. It attaches directly to the mainframe's byte multiplexer channel, and provides eight-line functional replacement for an IBM 270X, or for an IBM 370X or 4331 Communications Adapter operating in 270X emulation mode. Features include automatic polling, synchronous transparency, automatic speed detection and protocol selection for asynchronous lines, and support for ASCII, IBM 2741, and IBM BSC (EBCDIC and ASCII) protocols. Three submodels vary only in the number of 56K bps wideband BSC lines provided (each wideband line replaces a standard 9600 bps line): Model 81 provides none; Model 82, one; and Model 83, two. Optional features include an alternate channel switch, automatic dialing, code conversion, IBM 2260 support, and integral asynchronous modems.

● Honeywell's Information Systems Group has announced its Datanet 6661 Front-end Network Processor (FNP)

The new processor is an enhanced version of the Datanet 6641 and 6651 front-ends, which it replaces, and is based on Honeywell Level 6 minicomputer technology. It can be

Datapro Communications Management Newsbriefs provides comprehensive reporting of and insight into the implications of office-related news for Communications management by the publishers of Datapro Communications Solutions, Datapro Reports on Data Communications, Datapro Reports on Retail Automation, Datapro Reports on Banking Automation, Datapro 70, Datapro Reports on Minicomputers, Datapro Reports on Office Systems, Datapro Reports on Automated Office Solutions, Datapro Reports on Word Processing, Datapro Reports on Copiers and Duplicators, and Datapro Directory of Software. Datapro Communications Management Newsbriefs is supplied to all subscribers to Datapro Communications Solutions without additional charge, and is offered to nonsubscribers for $42.00 per year.

ısed with Honeywell's DPS 8, Level 66/DPS and Level ›8/DPS computers. The basic Datanet 6661 provides the ame capabilities and configuration as an entry-level)atanet 6641, including 64K bytes of memory, a ommunications console, a direct interface adapter for ıost connection, and support for up to 32 communica-ions lines. Options provide for expansion of the memory o a maximum of 512K bytes (more than double the ›651's maximum capacity) support for up to 96 ommunications lines, and inclusion of a cache memory.)ther enhancements include 16K MOS memory echnology, new bulkhead connectors designed to ease nstallation and configuration of data cables, and better ›rice/performance than the 6641/6651. The new ›rocessor is logically compatible with system software .nd user-generated programs of the Datanet 66 family of ⁼NP's, including support for Honeywell's Remote [erminal Supervisor-II (GRTS-II), Network Processing ›upervisor (NPS), and Multics Communication System MCS).

›OFTWARE

▶ Datapoint Corporation has released ARCCOM ›270

[his is a software product that permits multiple ›rocessors in an Attached Resource Computer (ARC) ıetwork to perform interactive data communications vith a remote mainframe. The ARCCOM software ·esides in a 6010 or 6020 ARC communications ›rocessor, which emulates an IBM 3271 control unit, and ·equires no revision in host hardware or software. The ›010 or 6020 is connected to the ARC network by the nterprocessor bus, and to the mainframe by the Multifunction Communications Adapter (#9481), a ›ynchronous modem, and leased lines. With ARCCOM ›270, Datapoint users can write high-level applications ›rograms for local processing that include operator-ransparent interaction with the mainframe, using)atapoint's Interactive COBOL, RPGPLUS, BASIC-PLUS, or DATABUS programming languages. Termi-ıals in the network emulate IBM 3277 keyboard/displays ınd IBM 3280 printers, and may access the system in one ›f two ways: by connection to the 6010/6020 communica-ions processor itself or by connection to an ARC ıpplications processor interconnected with the communi-:ations processor on the interprocessor bus. A Model ›010 processor can support up to eight Model 8200 ‹eyboard/displays and workstation printers, a Model ›020 processor can support up to 32 workstations. ARCCOM emulator utilities loaded into ARC applica-ions processors, such as another 6010 or 6020 processor ›r a 3800, 6600, 5500, or 1800 system, provide these ›ystems with IBM 3277/3280 emulation. Operator and ›rogram requests from these systems to the host and ·esponses from the host are sent via the ARC :ommunications processor with complete transparency to :he user. The ARCCOM software is available for no :harge when ordered with a new Datapoint system; for :xisting Datapoint users, the one-time license fee is $500. Monthly maintenance is $15 per month.

PROGRAMMABLE TERMINALS

● Burroughs adds to BMT product line

The new terminals include two new series of general purpose display-based systems, MT 700 and MT 900, and additional models in the MT 300 printer-based banking terminal line. The MT 700 Series terminals are user programmable and provide up to 96K bytes of random access memory. Applications programs are written in Burroughs' new Transaction Programming Language (TPL). TPL is a transaction-oriented, high-level language that provides the ability to create or modify Burroughs standard program products or user-written programs. Programs are developed interactively using the key-board/display and a micro-cassette unit, and compiled on the mainframe. The three MT 700 models released differ only in the size of the display screen: Model MT 755 has a five-inch screen; the MT 785, a 12-inch screen; and the MT 795, a nine-inch screen. Three different keyboard styles are offered. A choice of six printers designed primarily for banking applications, a micro-cassette storage unit with a 100K-byte capacity, a personal identification number keypad, and a magnetic stripe reader, all of which are newly announced BMT peripherals, may be added to the basic system.

The MT 900 Series terminals are designed to replace the Burroughs TD 830 Series. The MT 900s operate in TD 830 emulation mode and provide a nine- or 12-inch screen and a choice of three keyboards. Peripherals include a journal printer, 80K or 160K bytes of micro-diskette storage, and a magnetic stripe reader. The MT 300 Series of validation/journal printer terminals, which previously included the MT 355 and the MT 325 models, has been joined by the MT 357 and MT 327. Otherwise identical to the 355/325 units, the BMT 327/357 devices incorporate a RAM memory that can be programmed completely to user specifications through the use of the TPL compiler. The MT 300 Series is currently supported for commer-cial/banking applications only. All of the new terminals operate with Burroughs B 6700, B 6800, and B 6900 computers. In their initial versions they are compatible only with Burroughs communications protocols.

● Raytheon adds SNA compatibility

Raytheon Data System has expanded its PTS 100 line of IBM 3270-compatible products to include an SNA-compatible clustered terminal system. The new system emulates IBM's 3274 terminal controller and provides a 16-bit processor, direct access memory, and support for up to 32 3277/3278-type keyboard/displays and workstation printers. Raytheon has equipped its PTS-100 with 3270-compatible functions and a 3278-type keyboard. Because the PTS-100 terminal system is user-programmable, data storage, processing, and printing operations may be performed locally without host involvement. Basic user memory capacity is 16K bytes, expandable to 128K bytes. Two 3274-type configurations are currently available, one of which supports remote

BSC operation and the other of which supports remote SNA/SDLC operation. A third configuration that emulates the 3274-1B, which provides for local host connection, is scheduled for release before July, and a fourth configuration that supports 300-lpm printers in local or remote configurations will be ready sometime in mid-1980. Previous PTS-100 terminal systems operating in IBM 3271/2 emulation mode may be upgraded to 3274 operation.

DISPLAY TERMINALS

● ECS Microsystems has introduced the ECS 4000

This is a multifunction and multiprotocol display terminal that can interface with all major mainframes. The ECS 4000 terminal offers multiprotocol options for IBM, Burroughs, Honeywell, DEC, Sperry Univac, and NCR, among others. The ECS Microsystems 4000 is designed for distributive information networks and as a companion to the ECS 4500 for distributive processing and network operations, where the ECS 4500 functions as a local cluster controller. The ECS 4000 features synchronous or asynchronous transmission at rates from 75 to 19,200 bps via an RS-232-C interface. Screen management features include reverse video and reverse video cursor, four intensity levels, underline, blinking, double-width characters, protected fields, and character insert/delete. The screen can display 25 lines of 80 characters each and uses an 8 x 7 dot matrix for character formation. The ECS 4000 is based on a Z-80 microprocessor and includes 16K line/display buffer, a 256 programmable character set, full modem controls, parallel interface options, and diagnostics. The keyboard is typewriter style and has numeric and program function keys. The ECS 4000 can be interfaced with printers from Centronics, Diablo, DEC, NEC, Teletype, and similar manufacturers.

TELEPRINTERS

● GE unveils new printers

General Electric's Data Communications Products Department has announced the introduction of the TermiNet 2000 impact matrix printers. The TermiNet 2000 teleprinters feature a new 7 x 9 dot matrix printhead mechanism which employs blade-mounted pins designed for quiet (less than 60 dBa) bidirectional printing. The TermiNet 2030 has a print rate of 30 cps and a catch-up rate of 60 cps. The TermiNet 2120 has a print rate of 120 cps and a catch-up rate of 150 cps. Both models offer underlining capability and selectable print density of 10, 13.2 and 16.5 characters per inch along a 13.2-inch prin line. The ribbon cartridge is stationary, and the ribbon i driven by a stepper motor. The ribbon cartridge als features a Mobius loop that extends ribbon life. A standard friction-feed platen is employed for pape handling. An optional tractor feed is available and ma be field installed. The printers can produce up to thre copies at various lines per vertical inch, ranging from 2 t 12 lines, and they can slew at the rate of five-inches-per second. The standard keyboard is an ANSI typewrite layout; an ANSI/APL layout and a numeric cluster ar available as options.

The new TermiNet 2000 series terminals weigh only 2 pounds. These new teleprinters require only 30 watt (TermiNet 2030) and 50 watts (TermiNet 2120) of power Both units also feature dual 8085 microprocessors, a received data buffer of 512 characters, non-volatil configuration memory, a 20-character answerback, and an RS-232-C interface. The TermiNet 2030 is capable o tranmission rates of 110 or 300 bps, while the TermiNe 2120 has transmission rates from 110 to 9600 bps. A 300 baud modem, an extended 16K data buffer, and a 32K edit buffer are some of the major options that will b available. The new TermiNet 2000 terminals will be sol to OEM's and through third-party distributors.

FACILITIES

● GTE Telenet is launching a nationwide electronic mail service on July 14

Unlike Telex, TWX or facsimile, which are machine and location-dependent, Telemail is designed as a universal person-to-person communications system. Each user is provided with an electronic "mailbox" which he can access from his home, his office or from out of town locations using the telephone and either a portable or desk-top terminal. Some Telemail features include electronic indexing and storage of messages, the ability to send multiple-address and broadcast messages to any number of people, and an electronic directory of subscribers nationwide.

Proposed rates for prime time Telemail are $13.25 per hour for local public dial charges and $25.00 per hour for nationwide in-WATS. Comparable non-prime time charges for the same service are $3.00 per hour and $17.00 per hour respectively. There is a delivery charge per message unit (1000 characters minimum) of $.12 for dial-out, $.05 over dedicated lines, $1.30 for TWX, $3.25 for Telex and $2.60 for Mailgram.

General Solutions to Common Communications Interface Problems

Problem:

From a nontechnical point of view, the entire question of communications often appears to be a trivial matter of connecting a wire from here to there. The goal of all communications engineering is to reduce every connection to the almost-transparent simplicity of a point-to-point wire, but the underlying technical realities of communications engineering constitute a substantial body of discrete problems whose solutions are far from trivial. The problems are divided into two general classes: compatibility and control.

A direct terminal-computer communications connection is the least complex in terms of compatibility because the terminal is invariably designed to accommodate the requirements of the computer, not the other way around, so that many of the technical solutions to the connection problems are preordained by the engineering makeup of the computer. The situation changes completely when the terminal is moved farther away from the computer and a third-party communications linkage facility (like the telephone network) is interposed between the terminal and the computer. Now the problems of compatibility turn into a three-way tug of war among dominating technologies of the computer and the communications facility and the less dominating but more cost-sensitive technology of the terminal. The problems of distribution and control are also aggravated because now the computer is frequently required to service upwards of a hundred terminals to gain the maximum possible leverage from investments in the computer and the communications facility.

The emerging perspective of the total set of communications problems consists of several levels. The immediate problems of inter-technology compatibility are generally handled by auxiliary hardware like modulators/demodulators (modems). The larger problems of distribution and control are attacked by combinations of hardware and software solutions. Distribution, for example, is optimized by devices like multiplexers. Concentrators further optimize the distribution solutions and provide some lower-level solutions to the control problems. Front-end and communicating processors are at the top of the distribution and control hardware solutions hierarchy. The software solutions also consist of several parts. The basic fabric of software control and coordination is provided by a multilayered structure of linguistic conventions called protocols. Protocols are basically special phrases that invariably bracket a communications dialogue of any type, that may be interspersed throughout *a dialogue, and that may be supplied on separate but related paths to the main dialogue path. Protocols consist of question and answer phrases based on coordinating queries "Who? What? When? How Long? Where? How much? Do you read me?" and so on. Less structured but even more complex software consists of telecommunications access methods, like IBM's TCAM; of teleprocessing (TP) monitors, which do the same job for a communications system that the operating system does for a computer; and of composite network control architectures, like IBM's SNA, that supply the final degree of coordinating "intelligence" for all the separate and separable components of a communications network. This report examines the basic problem/solution set for the terminal-to-transmission facility interface in more detail. Reports in this and later sections focus on these and other problems in even greater detail.*

Solution:

A teleprocessing system has three components that are incompatible in speed: the computer, the terminal, and the communication line linking them. These components are usually incompatible in other ways also. The computer and most terminals are digital devices designed to deal with bits, or square-edged pulse trains; but most communications lines are analog in operation, designed to transmit a continuous range of frequencies. The coding of characters used in most terminals differ from that in the computers they are connected to. The error rates encountered on most communication lines are higher than those generally accepted in computers.

This report is concerned with the engineering that is used to overcome these incompatibilities.

INCOMPATIBILITIES IN SPEED

Since its earliest days, the computer has operated much faster than its input and output units. Being an expensive machine, it cannot afford to wait for them. The problem was solved by using buffers. The computer dumps a quantity of output into a storage unit, or buffer, at full speed, and then the contents of this unit are printed slowly, at the speed of the printing mechanism. When the information has been printed, the computer again refills the buffer at its own speed. The opposite process takes place for input.

A communication line is handled in a similar way. The characters from the computer will normally be transmitted one bit at a time from a buffer that is refilled periodically.

TERMINAL BUFFERING

The terminal may also have a buffer. Suppose that a voice line transmits at 4800 bits per second from keyboard terminals. The operators cannot type information into the terminal at this speed; at best, they may type about three characters (21 bits) per second, but there will be lengthy pauses for thought and other activity. In this case, their keying might fill up a buffer of, say, 100 characters, the contents of which would be transmitted, in turn, over the line in one burst that would take less than one-fifth of a second. This process makes sense, however, only when the line has some other work to occupy its time when it is not transmitting from this particular terminal. If only one terminal is attached to the line, then there is no point in having a terminal buffer (at least for purposes of timing). The need for the buffer arises when more than one device is attached to the line. In practice, as we shall see, many devices can be attached to one communication line, and sometimes many terminals share the same buffer. The buffers are used both for transmissions to and from the computer.

Paper tape has traditionally been used as a cheap form of buffer. A message is punched into paper tape, and the tape is placed on a tape reader to be read at the computer's convenience. This scheme is used on many message-switching systems. A buffer allows messages to be composed fully before being transmitted to the computer. If the message is long or complicated or if it takes a long time to compose, it may be worthwhile checking to see that it is correct

---

Systems Analysis for Data Transmission by James Martin, Chapter 13, pp. 155-179. © 1972 Prentice-Hall Inc. Reprinted by permission of Prentice-Hall Inc., Englewood Cliffs, N.J.

General Solutions to Common Communications Interface Problems

before transmitting any of it. If many long messages are sent to the computer without buffering at the terminal and the operator enters characters slowly—perhaps because she is talking to a client on the telephone at the same time—then a large amount of computer memory will be tied up in buffering the messages at the computer end of the communication lines. Partially completed messages will remain in computer storage for a relatively long period before being processed. When buffering is used, no single terminal will occupy the line for a long period of time; neither will it tie up the computer for a long period.

A hardware buffer may be in the terminal itself, or it may be in a unit that controls several terminals. The logic associated with it usually carries out functions other than simple buffering.

The cost of buffering in a terminal was high once but has fallen substantially in recent years. Nevertheless, at the time of writing more unbuffered than buffered terminals are installed.

ERROR CONTROL

A different reason for having a buffer in a terminal concerns error control. It is one of the important features that a terminal may or may not have.

Errors in the transmission may be detected by using codes designed to reveal whether any bits have been changed. A block of data normally has some extra bits (or characters) added to it; these bits are obtained by performing a logical of mathematical operation on all the other bits. In this way, a unique checking pattern is generated at the transmitting machine. The same process is performed at the receiving machine; and if the pattern generated is identical, then it is assumed that no error has occurred in the transmission. The certainty of detecting an error depends on the effectiveness of the technique for generating the error pattern. Some rather complicated techniques make it highly unlikely for an error to slip through undetected.

Having detected an error, the receiving device may or may not take automatic action to correct it. The safest action is to request that the item be retransmitted. The device that originally sent the message receives an instruction to send it again. This step can be done only if the sending device still has the message. For a terminal, a buffer would be needed. The message would be retained in the buffer until it was known whether it had been received correctly or not.

*The name "Baudot code" has been commonly used to refer to the standard five-bit telegraph code (International Alphabet No. 2). It was, however, devised by Donald Murray and differs substantially from Baudot's original 5-bit code.

If a terminal has no buffer, error checking of messages from the terminal may still take place; but when an error is found, there can be no automatic retransmission. Instead, the operator is notified of the error, and she must initiate retransmission. The terminal must be able to send and receive the signals saying whether the data was received correctly or not.

An alternative to error detection and retransmission is error correction. Here a code is used that makes it possible to ascertain not only that there has been an error but also what that error was. The error is corrected without retransmission; thus this operation is referred to as "forward error correction."

SYNCHRONOUS VS. ASYNCHRONOUS TRANSMISSION

Data transmission can be either synchronous or asynchronous. Asynchronous transmission is often referred to as start-stop. With synchronous transmission, characters are sent in a continuous stream. A block of perhaps 100 characters or more may be sent at one time, and for the duration of that block the receiving terminal must be exactly in phase with the transmitting terminal. With asynchronous transmission, one character is sent at a time. The character is initialized by a START signal, shown in Figure 1 as a "O" condition on the line, and terminated by a STOP signal, here a "1" condition on the line. The pulses between these two give the bits of which the character is composed. Between characters, the line is in a "1" condition. As START bit switches it to 0, the receiving machine starts sampling the bits.

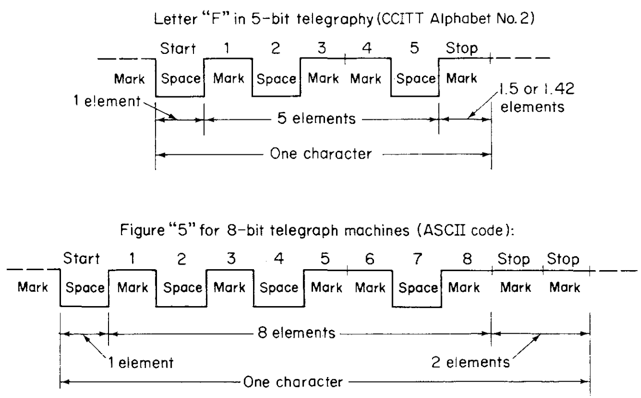


Figure 1. Typical character structures for START-STOP (asynchronous) transmission

ASYNCHRONOUS (START-STOP) TRANSMISSION

Figure 1 shows the form in which a character is sent with START-STOP transmission. The two most common types of character coding are illustrated, Baudot* code and ASCII code. Most non-American

General Solutions to Common Communications Interface Problems

teleprinters transmit Baudot code characters with five data bit plus the START and STOP elements (CCITT telegraph alphabet No. 2), as shown at the top of the figure. Most American terminals designed in recent years transmit ASCII characters, shown at the bottom of the figure, with eight data bits (of which one is often unused) plus the START and STOP elements.

START-STOP transmission is usually used on keyboard devices that do not have a buffer and on which the operator sends characters along the line at more or less random intervals as she happens to press the keys. The START pulse initiates the sampling; thus there can be an indeterminate interval between the characters. Characters are transmitted when the operator's finger presses the keys. If the operator pauses for several seconds between one keystroke and the next, the line will remain in the 1 condition for this period of time.

Start-stop machines are generally less expensive to produce than synchronous machines; for this reason, many machines that transmit card-to-card or paper tape-to-printer, card-to-computer, and so on, are also start-stop, although the character stream does not have the pauses between characters a keyboard transmission has.

The receiving machine has, in essence, a clocking device that starts when the START element is detected and operates for as many bits as there are in a character. With this, the receiving machine can distinguish which bit is which. The STOP element was made longer than the data bits in case the receiver clock was not operating at quite the same speed as the transmitter.

When this start-stop transmission is used, there can be an indeterminate period between one character and the next. When one character ends, the receiving device waits idly for the start of the next one. The transmitter and the receiver are then exactly in phase, and they remain in phase while the character is sent. The receiver is thus able to attach the correct meaning to each bit it receives.

When an automatic machine such as a paper-tape reader is sending START-STOP signals, the length of the STOP condition is governed by the sending machine. It is short, always 1.42 (1.5 or 2) times the other bits, so as to obtain the maximum transmission rate. When a typist uses the keyboard of a start-stop machine, on the other hand, the duration between her keystrokes varies. The transmission occurs when she presses each key, so the stop condition varies in length considerably. When a scientist uses a teleprinter on a time-sharing system, he may be doing work that involves a great amount of thinking. Occasionally there may be a very long pause between one character and another while he thinks or makes notes. The STOP "bit" will last for the duration of this period.

Teletype speeds in common use are listed in Table 1. Such speeds are often quoted in "words per minute." An average teletype word is considered to be five characters long. Because there is a space character between words, there are, then, six characters per word, and x words per minute = x 10 characters per second.

Speed in Bauds (bits per second)	Number of Bits in Character	Stop Bit Duration (in bits)	Information Bit Duration	Characters per Second	Words per Minute (nominal)
45.5	7.42	1.42	21.97	6.13	60
50	7.42	1.42	20	6.74	66
50	7.50	1.50	20	6.67	66
74.2	7.42	1.42	13.48	10	100
75	7.50	1.50	13.33	10	100
75	10	1.00	13.33	7.5	75
75	11	2.00	13.33	6.82	68
150	10	1.00	6.67	15	150

Table 1. Commonly used Teletype speeds

SYNCHRONOUS TRANSMISSION

When machines transmit to each other continuously, with regular timing, synchronous transmission can give the most efficient line utilization. Here the bits of one character are followed immediately by those of the next. There are no START or STOP bits and no pauses between characters. The stream of characters of this type is divided into blocks. All the bits in the block are transmitted at equal-time intervals. The transmitting and receiving machines must be exactly in synchronization for the duration of the block so that if the receiving machine knows which is the first bit, it will be able to tell which are the bits of each character (or words).

Devices using synchronous transmission employ a wide variety of block lengths. The block size may vary from a few characters to many hundreds of characters. Often it relates to the physical nature of the data medium. For example, in the transmission of punched cards it is convenient to use 80 characters as the maximum block length, for there are that many characters per card. Similarly, the length of print lines, the size of buffers, the number of characters in records, or some other such system consideration may determine the block size. Some time is taken up between the transmission of one block and the next; therefore, the larger the block length, in general, the faster the overall transmission.

With asynchronous transmission, the unit of transmission is normally the character. The operator of a teletype machine presses a key on her keyboard and one character is sent, complete with its START and STOP bits. It is independent in time of any other character. With synchronous transmission, the characters are stored until a complete block is ready to be sent. The block is sent from a buffer at the maximum speed of the line and its modems. There are no gaps between characters as there are when a teletype operator taps at her keyboard. Synchronous transmission is therefore of value when one communication line has several different terminals operating on it. In order to permit synchronous transmission, however, terminals must have buffers; consequently, they are more expensive than asynchronous devices.

The synchronization of the transmitting and receiving machines is controlled by oscillators on many systems. Before a block is sent, the oscillator of the receiving machine must be brought exactly into phase with the oscillator of the transmitting machine. This step is done by sending a synchronization pattern or character at the start of the block. If this were not done, the receiving device would not be able to tell which bit received was the first bit in a character, which the second, and so on. Once the oscillators at each end are synchronized, they will remain so until the end of the block. Oscillators do, however, drift apart slightly in frequency. This drift is very low if highly stable oscillators are used; but with those low enough in price to be used in quantity in input-output units, the drift is significant. Oscillators in common use in these machines are likely to be accurate to about one part in 100,000. If they are sampling the transmission 2500 times per second, say, then they are likely to stay in synchronization for a time of the order of 20 seconds. Most data processing machines resynchronize their oscillators every one or two seconds for safety. Synchronization can also be maintained by "framing" blocks and carrying timing information in the frames.

On some systems, this places an upper bound on the block length, but not always, because resynchronization characters may be set in the middle of a block. The IBM range of "binary synchronous" equipment, for example, inserts two synchronization characters into the text at one-second intervals. In the U.S. ASCII code with parity checking, these characters would be coded 01101000 01101000. The receiving station is constantly looking for the synchronization pattern and so ensures that the transmitter and receiver are in step.

BLOCK STRUCTURE

A block of bits sent by synchronous transmission must have certain features. It must, for example, start with the synchronization pattern or character. It will normally end with an error-checking pattern or character. The block length, as with other data records used by computers, may be of fixed length or variable length. Frequently it is the latter, for variable length usually allows better line utilization. It would be necessary on most systems to pad many blocks with blank characters if fixed-length blocks were used. If the block is of variable length, an end-of-block pattern must be used to tell the receiving machine to begin the actions needed when a block ends. This pattern will normally be sent immediately prior to the error-checking pattern.

Often data are sent in the form of characters or groups of (usually) 6, 7, or 8 bits. The above patterns can be 1, 2, or more characters. One transmission scheme, for example, uses six-bit characters. These are transmitted without parity checking, so that the whole block is divided up into groups of six bits. The block must start with the following characters: 111111, 111110 (in that sequence). This constitutes the synchronization pattern. A circuit in the receiving machine spends its life scanning the input for this pattern. When it finds it, then the receiving device knows that the next bit it receives is the first data bit. The synchronization pattern is unique. The coding of characters must be such that it could not occur anywhere else in the transmission.

The block ends with a six-bit error-checking pattern (one character), and immediately preceding that is the end-of-message character. When the text is being transmitted, the receiving device is generating its own error-checking pattern, which is computed from the characters received. At the same time, it is examining each character received to see whether it is the end-of message character. When this character is received, the machine knows that the next one is to be the transmitted error-checking pattern, and so it compares that with the pattern it has generated itself. If there is a difference, the receiving machine sends a message to the transmitting machine to demand a retransmission of that message.

Figure 2 shows the format of a block of text transmitted in this manner. It is designed for a line to which many input-output machines are attached. These machines are arranged in groups, and each group is connected to a control unit, which itself is connected to the line that transmits data to and from the computer. After the synchronization pattern in each block, comes the address of the control unit (one character) and the address of the input-output machine (one character) to which the message is going or from which the message has come. It is possible that messages transmitted to the computer may be longer than the maximum length of a block. In this case, they are divided into as many blocks as necessary, and a character is used as a segment identifier to link them. The control unit places this

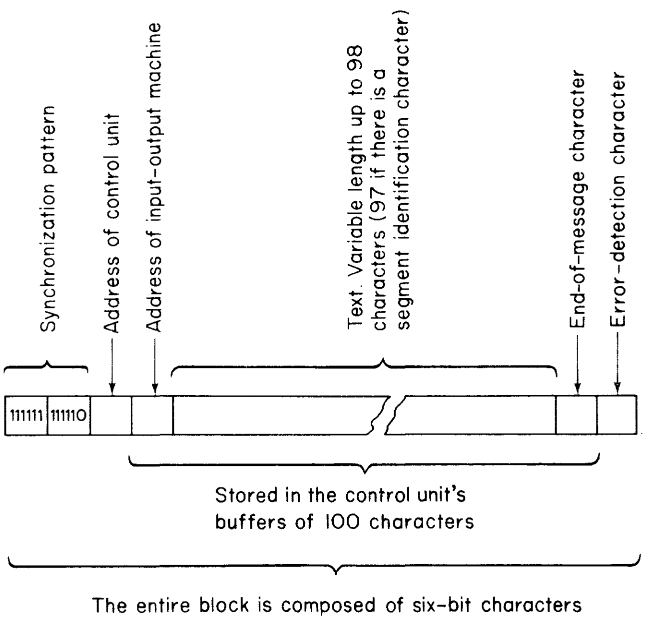


Figure 2. Typical format of a block of data for synchronous transmission on a heavily loaded line with many terminals where efficient line loading is important

identifier, if it is needed, in the block immediately before the text. The text itself is again in six-bit characters and can be of any length up to 98 characters. This maximum is imposed by the size of the buffers (100 characters) used in the control units.

There are many variations of this type of format. Sometimes one character is designated as the "synchronization character," and a stream of these characters is sent continuously between messages when the line would otherwise be idle. At least two such characters are necessary prior to a message to establish synchronization. The example in Figure 2 is perhaps noteworthy because it minimizes the number of bits transmitted. Only six-bit characters are used, which is satisfactory for many applications. On many synchronous transmission schemes, the number of bits needed has been allowed to grow much higher than in this illustration; consequently, it may be worthwhile to question the necessity for the excess bits.

ADVANTAGES AND DISADVANTAGES OF SYNCHRONOUS OPERATION

There are two conflicting desires: to make the terminal inexpensive and to use the communication lines efficiently by putting many terminals on one line. Up until now, low-cost terminals have been start-stop with no buffers. Buffered, synchronous terminals have been more expensive, but have given better line utilization. Where the lines are short and inexpensive (e.g., within one city), efficient line utilization is of little importance. When a dial-up line is used, normally there will only be one terminal on the line and start-stop operation will often be good enough. A somewhat higher character-transmission rate could be obtained with synchronous transmission.

The other main advantage of synchronous transmission is that the error rate can be less. Extremely good error control can be achieved with high-order error-detecting codes and a buffer in the sending machine so that retransmission can be automatically requested.

The disadvantage of synchronous operation, the fact that it is more expensive, is diminishing as the cost of logic circuitry drops. At the same time, the reliability of logic circuitry is substantially increasing. All the logic for a synchronous, buffered, error-checking terminal can now be constructed on one large-scale-integration chip, which can be low in cost if large quantities are mass-produced.

EDITING

The inexpensive start-stop terminal may contain no logic other than that for transmitting and receiving each character. At the other end of the scale, a terminal may contain a high level of "intelligence." The systems analyst must assess the value of placing intelligence in the terminal rather than elsewhere.

One of the functions that logic circuitry in the terminal can perform is editing. Messages from the computer may contain new line indicators and tab or column number indicators. These will cause the date to be laid out on the screen in a neat, readable fashion without having to transmit many blank characters within the message. Similarly, the operator may construct her message in a buffer, modify it, and correct any errors she may have made before it is transmitted. This operation may be more intricate if her responses consist of filling in or changing items on a screen already filled with data.

On a graphic terminal, there are many possible ways in which the image may be "edited." The number of bits needed to transmit a line drawing or, for that matter, any other image depends on the amount of logic in the terminal to assist in constructing the image.

LINE DISCIPLINE

It is often desirable to attach several terminals to one line. There are many ways to organize a network so that the cost of attaching many terminals to a computer is minimized. Where one line interconnects several terminals that transmit in turn, not simultaneously, some form of line discipline is needed.

General Solutions to Common Communications Interface Problems

When several devices all share one communication path, only one can transmit at once, although several or all points can receive the same information. Each terminal must have an address of one or more characters, and it must have the ability to recognize a message sent to that address. A line may, for example, have 26 terminals with addresses A to Z. The computer sends down the line a message that is to be displayed by terminals A, G, and H. The message is preceded by these three addresses, and each terminal has circuitry that scans for its own address. Terminals A, G, and H recognize their addresses and display the message simultaneously. The other terminals do not recognize their address and so ignore the message. The network may also have a "broadcast" code, which causes all terminals on a line to display those messages preceded by it.

POLLING

For transmission in the other direction, several terminals may wish to transmit at the same time. Only one can do so; the others must wait their turn. To organize this, the line will normally be polled. A polling message is sent down the line to a terminal, saying, "Terminal X, have you anything to transmit? If so, go ahead." If terminal X has nothing to send, a negative reply will be received and the next polling message will be sent, "Terminal Y, have you anything to transmit? If so, go ahead."

The device that does the polling—often a computer—will have in its memory a polling list giving the sequence in which the terminals should be polled. The polling list and its use determine the priorities with which terminals are scanned. Certain important terminals may have their address more than once on the polling list, and thus they are polled twice as frequently as the others. Any number of lines may be in use at one time.

Roll-call polling can sometimes degrade the response time obtained at the terminals. This is particularly true when the time taken to reverse the direction of transmission on the line is lengthy (line turnaround time). Nevertheless, there are many fast-response systems on which a large number of terminals are polled. Polling schemes other than roll-call polling can avoid many of the line turnarounds that occur in that process, thereby giving much improved response times.

There are other forms of line discipline in which a continuous stream of characters travel nonstop on the line. For any line discipline, however, appropriate logic must be built into the terminal.

FULL DUPLEX VS. HALF DUPLEX

Over a given physical line, the terminal equipment may be designed so that it can either transmit in both directions at once, full-duplex transmission, or it can transmit in either direction but not both at the same time—half duplex.

An input-output terminal or a computer-line adapter works in a somewhat different fashion, depending on which is used. Where full-duplex transmission is employed, it may be used either to send data streams in both directions at the same time or to send data in one direction and control signals in the other. The control signals govern the flow of data and are used for error control. Data at the transmitting end is held until the receiving end indicates that the data has been received correctly. If the data is not received correctly, the control signal indicates this fact, and the data is retransmitted. Control signals ensure that no two terminals transmit at once on a line with many terminals, and the signals organize the sequence of transmission.

Simultaneous transmission in two directions can be obtained on a two-wire line by using two separate frequency bands. One is used for transmission in one direction and the other for the opposite direction. By keeping the signals strictly separated in frequency, they can be prevented from interfering with each other.

The two bands may not be of the same bandwidth. A much larger channel capacity is needed for sending data than for sending the return signals that control the flow of data. If, therefore, data is to be sent in one direction only, the major part of the line bandwidth can be used for data. Some schemes thus permit a high bit-rate in one direction with a very low bit-rate return path. This transmission can usually be reversed in direction so that data can be sent either way. One modem, for example, permits transmission of data at 3600 bits per second in one direction and provides a simultaneous return path for control signals at 150 bits per second.

Many data processing situations are not able to take advantage of the facility to transmit streams of data in both directions at the same time. Consequently, where full-duplex transmission is used, it is often with data traveling in one direction only, the other direction being used for control signals.

Full-duplex lines are generally more expensive than half-duplex lines, commonly 10 percent more expensive in the United States.

General Solutions to Common Communications Interface Problems

CHARACTER CODING

A variety of different codes are used on transmission lines.[1] The most common ones are the 7-bit US ASCII code, the United States standard, and the 5-bit Baudot code used for international telegraphy.

All such codes use control characters to indicate start of message, end of message, error indication, and so on. The transmission cannot take place without some of these control characters. However, it is often desirable to transmit all of the six-, seven-, or eight-bit combinations that a computer or its peripheral device can store. A conflict therefore arises here.

The conflict is resolved by using a pair of characters for the control character, instead of one. For example, the DLE character (Data Link Escape) of the ASCII and similar codes may precede any control character, and this tells the receiving machine that the control character has its control meaning. The DLE character is regarded as not being part of the data. In order to transmit a DLE character and have the receiving machine accept it, it must itself be preceded by a DLE character.

This type of transmission is sometimes referred to as a transparent code or transmission in transparent text mode.

Some machines can switch backward and forward between transparent and normal text. Sequences of characters are needed for this operation, for example.

DLE STX: Initiate transparent text mode.

DLE ETB: Terminate transparent transmission.

DLE ITB: Terminate transparent text mode but continue transmission in normal mode.

Sometimes it is desirable to transmit more characters than there are combinations in the code. The five-bit Baudot code, for example, has $2^5 = 32$ possible combinations, but it is necessary to send the digits, letters of the alphabet, and punctuation with it. This step is done by an "escape mechanism," a character that changes the meaning of the following characters. In the case of the Baudot code, "letters shift" and "figures shift" characters indicate whether the characters following them are from a numeric set or an alphabetic set, like the use of the shift key on a typewriter.

[1]James Martin, Teleprocessing Network Organization, Chapter 2, Prentice-Hall, Englewood Cliffs, N.J., 1970.

CODE CONVERSION

It is desirable for devices that use different codes to be able to communicate. In order to do so, some form of code conversion must be used. Code conversion most commonly takes place in the central computer system. However, it may take place in a remote line-control computer or in the terminal control unit.

SECURITY

Several aspects of security may affect the interface between the terminal and the communication line. First, the terminal should be able to identify itself uniquely to the computer. This will prevent a person at a different terminal from contacting the computer and carrying out unauthorized operations. The substitute terminal could be connected to the computer by simple dialing in some cases and in other cases by wire-tapping at a private branch exchange. To identify itself, the terminal may, on interrogation, transmit a unique number, which is hard-wired into tamperproof circuiting. This step alone will not give protection from the determined intruder who has a high level of engineering capability. He could obtain the unique number by recording on a tapped wire and modify his terminal to transmit it. However, the unique terminal number combined with a related set of other measures will defeat most attempted breaches in security.

The second technique that may be used is cryptography. The user may have a magnetic cartridge of random numbers, and the computer has the same set on its file for the user. The terminal uses these sets for encrypting the data sent and for decrypting the data received. The cartridge is changed periodically. A variety of other cryptography techniques are possible.

The terminal may be equipped with a lock so that it cannot be used by persons without a key. Some terminals have the facility to read a user's identity card. A card the size of a credit card with data encoded on a magnetic stripe is used for this purpose. Lastly, the terminal may have the ability to inhibit printing or display when the user keys in his security code.

BANDWIDTH OF A VOICE CHANNEL

The signal-carrying capacity of communication links can be described in terms of the frequencies they will carry. A certain physical link might, for example, transmit energy at frequencies ranging from 300 to 150,000 hertz. [The word hertz (Hz) has replaced "cycles per second" in describing frequency and bandwidth. Their meanings are identical.] Above 150,000 and below 300 Hz, the signal is too much attenuated to be useful. The range of frequencies is

General Solutions to Common Communications Interface Problems

described as the bandwidth of the channel. The bandwidth is 149,700 (= 150,000 — 300) Hz (or cycles per second). In fact, the upper cutoff point is not as sharp as is suggested by this, and we would probably say a bandwidth of 150 kilohertz (kHz).

Figure 3 shows the attenuation of different frequencies on a typical voice channel. It will be seen that between about 300 and 3400 Hz different frequencies are attenuated roughly equally. Frequencies outside these limits are not usable, and therefore we would say that this channel had a bandwidth of 3100 Hz.

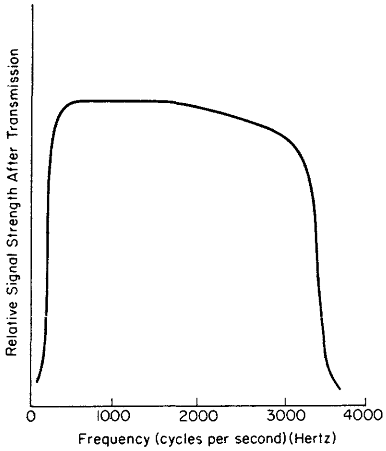


Figure 3. Variation in signal strength with frequency after transmission over a typical voice line

The quantity of data that can be transmitted over a channel is approximately proportional to the bandwidth.[2]

The frequencies transmitted in Figure 3 are not sufficient to reproduce the human voice exactly. They are, however, enough to make it intelligible and to make the speaker recognizable. This is all that is demanded of the telephone system. Hi-fi enthusiasts strive to make their machines reproduce frequencies from 30 to 20,000 Hz. If the telephone system could transmit this range, then we could send high-fidelity music over it. Sending music over the channel in Figure 3 would clip it of its lower and higher frequencies, and it would sound less true to life than over a small transistor radio.

The physical media used for telecommunications all have a bandwidth much larger than needed for one telephone conversation, so between towns one link is made to carry as many voice channels as possible. The bandwidth of one physical channel is electronically cut up into slices of 4000 Hz, and each of these slices becomes one voice channel.[3] The result is shown in Figure 3. The frequencies given here fit easily into the 4000-Hz slice.

In order to transmit data over the telephone line, then, we must manipulate it electronically so that it fits into the frequencies of Figure 3. This step is done by a modem, which will be discussed shortly.

Nonvoice channels have different bandwidths than shown in Figure 3—subvoice-grade channels are lower in bandwidth, and broadband channels are higher. If desirable, channels of extremely high bandwidth can be obtained.

MODULATION

Data entering or leaving data processing machines is normally binary in form and consists of rectangular pulses resembling those in Figure 4. It is necessary to convert these pulses so that they will travel over the range of frequencies shown in Figure 3, or whatever the frequencies of the line in question are.

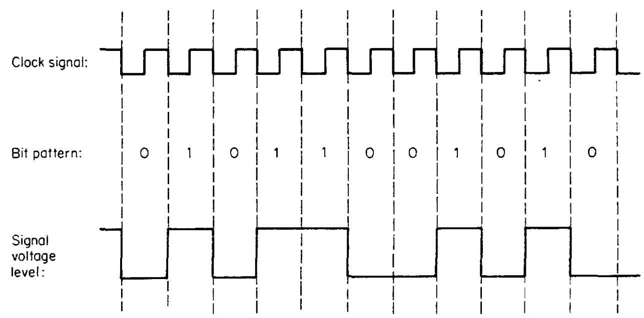


Figure 4 Pulse form of computer data

Two problems become apparent. First, the line represented by Figure 3 does not transmit dc current. Frequencies below 200 Hz are severely attenuated. A data pattern in which every bit is a "1," for example, would not be transmitted. Second, high frequencies are attenuated, and this fact alone would cause our square-edged pulses to become distorted. The faster the bit rate, the greater would be the distortion.

The square-edged pulse train is, therefore, manipulated electronically to make it fit as well as possible into the transmission frequencies of Figure 3. In a typical system, a "carrier" is used that is a single-frequency signal in the middle of the band available for transmission. The carrier is modified in some way by the data to be sent so that it "carries" the data. This process is referred to as "modulation."

As can be seen in Figure 3, a certain range of frequencies travels without much distortion over

[2]See James Martin, *Telecommunications and the Computer*, Chapter 10 and 11, Prentice-Hall, Englewood Clifss, N.J., 1969.
[3]*Telecommunications and the Computer*, Chapters 10 and 15.

General Solutions to Common Communications Interface Problems

telephone circuits. A frequency of 1500 Hz, for example, is near the middle of the human voice range. Modulation employs these voice frequencies to carry data that would otherwise suffer too much distortion. Thus, a sine wave of 1500 Hz may be used as a carrier on which the data to be sent is superimposed in the manner shown in Figure 5.

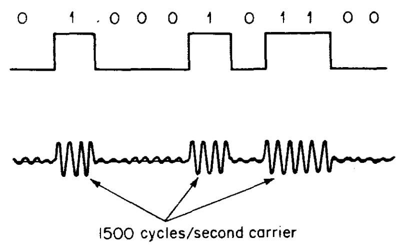


Figure 5. Amplitude modulation of a frequency at the center of the voice band

In addition to making it possible to send signals with a dc component over channels that will not transmit direct current, modulation achieves two ends: first, it reduces the effects of noise and distortion and, second, it increases the possible signaling speed. By using simple modulation devices, one can send computer data without undue distortion over the voice circuits and other communication lines of the world.

Figure 5 illustrates one type of modulation—amplitude modulation—in which a 1 is represented by a high-amplitude sine wave at the carrier frequency and a 0 is represented by a lower-amplitude wave of the same frequency. Other types of modulation are described in the later reports.

MODEMS AND DATA SETS

In order to achieve modulation, the binary output from the data processing machine must enter a "modulator," which produces the appropriate sine wave and modifies it in accordance with the data. This process produces a signal suitable for sending over voice circuits; and whatever manipulation the electronics do to the human voice, they can also do to this signal, and the data will still be recoverable. At the other end of the communication line, the carrier must be "demodulated" back to binary form. The circuitry for modulating and demodulating is usually combined into one unit, referred to by the abbreviated term modem.

The modem, a unit about the size of a domestic radio set, is connected to the data-processing machine, and the machine is then able to transmit data over normal telephone or other lines as shown in Figure 6.

Modems are made both by the computer manufacturers and by the telephone companies. They are sometimes also called data sets.

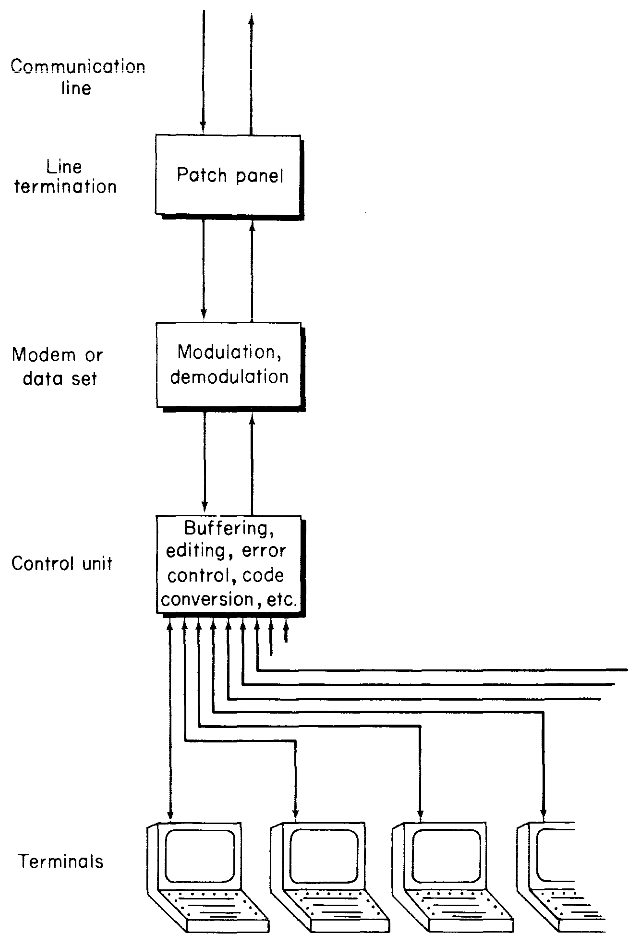


Figure 6 Terminal-to-communications-facility conversion sequence

The increasing choice of modems on the market and in the laboratory will give the systems engineer more scope in selecting between these criteria to suit his particular data processing environment.

TRANSMISSION WITHOUT MODEMS

Modems are not necessarily used on lines that are privately laid, although here, still, they can increase greatly the speed of transmission. Many computer users need to have data transmission lines within their own premises, as well as privately owned lines linking two buildings near each other. The terms in-plant and out-plant system are used. "Out-plant" system implies that common carrier lines are used. "In-plant" lines are normally a straightforward copper path, possibly coaxial cable, connecting the points in question. Private links of this type are often installed by a firm's own engineers. Sometimes they are also provided by telecommunication companies but external to any major telecommunication network.

General Solutions to Common Communications Interface Problems

Devices that use these lines often operate by the simple making and breaking of relay contacts, or the sending of rectangular pulse trains such as those shown in Figure 4. No modulation is needed. Over a wire pair a few miles in length, dc pulses can be sent at speeds up to about 300 bits per second. The distortion of the signals makes it impractical to send data in this form at speeds much higher than 300 bits per second over an ordinary pair of wires, except over short distances or unless closely spaced repeaters reconstruct the pulse stream—a very powerful technique to be discussed later. A speed of 300 bits per second, however, is a useful one for many computer applications. The speed could be increased greatly by using small coaxial cables rather than wire pairs. In many systems, a large number of typewriter-speed terminals within a localized area, say 3 miles across, could be connected to a time-sharing system or to a concentrator without modems. Although most common carrier lines require modems today, it is possible that the wire pairs that connect a central office to all locations with a telephone could be used over a limited area for dc signaling as in the earlier days of telegraphy. A low-cost private branch exchange for data signals used in this way has been developed.

A modem for low-speed transmission typically costs about $20 or $40 monthly rental. A time-sharing system with 500 low-speed terminals (only a few of which are in use at any one time) would be likely to pay, then, about $20,000 to $40,000 per month for modems. If dc signaling could be used, this high cost would be avoided.

TERMINAL COMPONENTS

All the mechanisms discussed in this report may be in one unit, or there may be several units. Figure 6 shows the units on a typical visual display system attached to a leased line. Separate terminals are linked to a common control unit, which contains most of the digital logic and storage.

The control unit is attached to the modem by using a standard terminal-modem interface. In some cases, the modem is built into the packaging of the terminal itself, and terminal and modem can be tailored to each other's needs.

The termination of the telephone line is at a "patch panel," which with the aid of jack-plugs enables terminals to be switched to different lines.

In many cases, the terminal will use a public telephone dial line; then the modem must be connected to a telephone, or a data set with a telephone dial must be used.□