About This Issue

Welcome, and "Thanks" for choosing to read NCP and 3745/46 Today. It's hard to imagine that we're halfway through the year 2001 when it seems like only yesterday that we were concerned about the year 2000. Time stops for no man and definitely for no business. Everyday there is a new and unique challenge to address in our ever-changing world of business. Or should I have said **e-business?** That's what it's all about today — e-business.

Of course, we've all heard about the "dot-com" successes and the "dot-bombs." Maybe the dust has settled and maybe not. Nonetheless, the task at hand is how to move our businesses into the e-business future. That term means a lot. It means how to move our business to the Internet. It means how to intranet inside our business or organization. It talks to how we do business with our customers and our suppliers. It drives us to greater efficiency in our infrastructure and our superstructure. And it's not just for private enterprise. It's for all types of organizations — for profit, or not for profit, or government — as well. One other thing that "e-business" means is the explosion of data and the need for storage strategies for the future.

SAN/NAS

The first section of this issue is about storage strategies of the now and the future. We're fortunate to have some opening words by Linda Sanford, Senior Vice President and Group Executive, IBM Storage Systems Group. Linda offers a unique insight into the storage demands of today and the future, and how storage networking is going to meet those demands. Following Linda Sanford's introduction, we have an article on what storage networking is all about. Then the first section is rounded off with a some "hot off the press" articles about brand-new storage networking products from IBM. This is must-read information if you're in business and want to compete in business today.

Network Technology

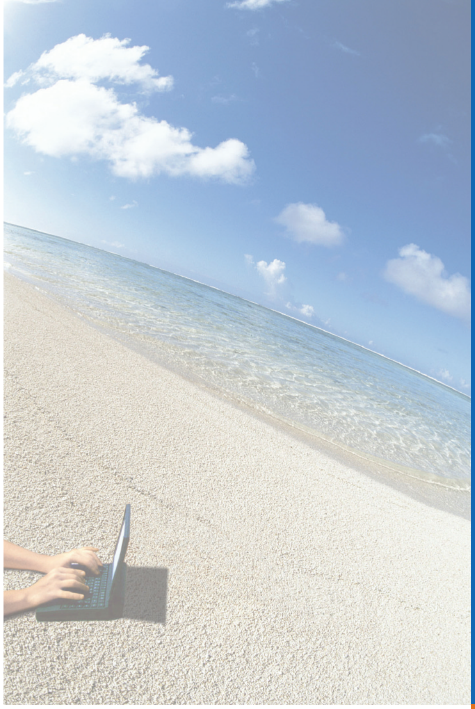
Of course, we realize that there are other topics that might pique your interest, so we have an entire second section dedicated to general networking technology subjects. These subjects cover the hard-to-answer questions of when and where to start the e-business transformation, how to merge SNA and TCP/IP environments, and how to balance those merged environments. They also touch on class/quality of service and security in these new environs. Again, this section is real "reach out and touch you" information for your business today. I know you'll enjoy it.

NCP and 3745/46

The third section is about the latest and greatest from us in the NCP and 3745/46 world. We've worked hard to respond to a number of new requirements that came directly from you, our customers. Then, we added a few new functions from ideas of our own. We hope this new support will make an already highly predictable and highly reliable product set more valuable to you and easier to use. These new functions span the product set from the 3745/46 to NCP, SSP, and NTuneMON™ and the service for each.

In our last issue, we included an NTuneMON demonstration CD inside the back cover and we received much favorable feedback — so much good feedback that we decided to do it again. The CD cover has all the instructions you need to download it, and detailed instructions for using it are in the downloaded material.

We would also like to remind you that you are important to us and we want to say "Thank you" for being our customers and for using our products and services.



Again, thank you for spending some of your valuable time reading our magazine. We sincerely hope that you find it stimulating, enjoyable and rewarding — both personally and professionally. We strive to make *NCP and 3745/46 Today* the best magazine in its class. The Summer 2000 Edition won an Excellence Award in the Society for Technical Communication (STC) 2001 Technical Publication Competition. Additionally, the enclosed NTuneMON CD won a Merit Award in the STC 2001 Online Information Competition.

If there's something we have missed or something you'd like to see in a future issue, please use the feedback form at the end of the magazine to contact us by mail, or reach us at **ibm.com**/networking/ncp

Look out for those "data explosions," and whatever you do, enjoy your year 2001.

Table of Contents

SAN/NAS

Network Technology

NCP and 3745/46

SAN/NAS

As IBM continues its drive to industry leadership in storage, you can count on our full support and commitment to open standards, to leadership in storage networking and to delivering world-class products, on-time.

IBM Continues Drive to Storage Leadership

by Linda Sanford, Senior Vice President and Group Executive, IBM Storage Systems Group

Two-and-a-half billion Web sites. Seven million more coming online every day. More than 500 billion documents online.

These estimates from the NEC Research Institute illustrate the enormous information explosion that has become a critical issue for IT executives around the world. The relentless growth in data is driving huge increases in storage requirements for businesses of all sizes. The average company's storage needs are more than doubling every year.

As companies consider ways to manage their data, secure it and make it accessible when and where it's needed, storage area networks (SANs) have become a popular alternative to traditional direct-attached storage. Three out of five companies have implemented SANs or are planning to initiate SAN projects in the near term, according to a recent IDC survey of large enterprises in North America. With the spending growth for SANs and network-attached storage (NAS) already outpacing the growth of direct-attached spending, SANs will almost certainly become the dominant model for storage environments in the future.

As storage moves into the network, IBM is committed to helping customers make this migration in a way that is evolutionary, not revolutionary. We want to protect your existing storage investments and help you integrate your storage resources to make your storage environment more manageable and your data more valuable.

Three out of five companies have implemented SANs or are planning to initiate SAN projects in the near term, according to a recent IDC survey of large enterprises in North America.

To respond to your accelerating storage requirements, IBM and our Business Partners have invested several billion dollars over the past 18 months to recapture industry leadership in storage. The IBM TotalStorage™ portfolio offers the industry's most complete array of disk and tape hardware, storage networking software and the services capabilities that customers need to put it all together. IBM has also been a leading voice in the industry behind an important cause: the development of open standards that allow for true and complete interoperability. In June, IBM joined five other storage vendors in announcing an unprecedented interoperability initiative that will make possible the first cross-vendor, jointly supported open storage networking solutions.

Open standards and interoperability are critical if we are to deliver on the true promise of storage area networks: universal access to information, where data can be stored and accessed from anywhere, anytime, from any applications running on any operating system on any hardware platform.

In this issue of *NCP and 3756/46 Today* magazine, Diane Pozefsky looks in-depth at the evolution of storage networking and where we see the technology heading. You'll also find stories on Tivoli® storage network management software, updates on key networking technology and information on our new NAS gateway products.

As IBM continues its drive to industry leadership in storage, you can count on our full support and commitment to open standards, to leadership in storage networking and to delivering world-class products, on-time.

You can also count on us to keep you informed of all the latest developments — through publications like this one. Please fill out the feedback form in the back of this issue and let us know what topics you'd like to see addressed in future issues.

Thanks for your business and for your confidence in us as a strategic partner you can count on to help you meet the storage challenge. ■

About the Author

Linda Sanford is Senior Vice President and Group Executive, IBM Storage Systems Group, the organization that develops and markets IBM's Enterprise Storage Server ("Shark^{TM}") and other open storage networking-related hardware and software that allow customers to use their data as the currency of e-business. Prior to assuming her current position, Ms. Sanford headed Global Industries, the IBM organization that manages relationships with IBM's largest customers worldwide and is responsible for generating about 70 percent of IBM's revenue.

Before that, Ms. Sanford was General Manager of IBM's S/390 Division, which develops, manufactures and markets large-enterprise systems. During the early 1990s, she guided the S/390 Division through one of the most comprehensive product transformations the computer industry has ever seen, reinventing S/390 as an open, enterprise-level server for today's e-business applications.

Ms. Sanford joined IBM in 1975 as an engineer in the Typewriter Division. She has held a number of executive positions at IBM, including executive assistant to the Chairman of the Board and Director of IBM Networking Systems.

One of the highest-ranking women at IBM, Ms. Sanford is a member of the Women in Technology International Hall of Fame and the National Association of Engineers. She has been named one of the 50 Most Influential Women in Business by *Fortune* magazine, one of the Top Ten Innovators in the Technology Industry by Information Week magazine, and one of the Ten Most Influential Women in Technology by *Working Woman* magazine.

Ms. Sanford serves on the Board of Directors of ITT Industries, St. John's University and Rensselaer Polytechnic Institute. She is a graduate of St. John's University and earned an MS in Operations Research from Rensselaer Polytechnic Institute.
lsanford@us.ibm.com

Storage Networking: More than an SNA Anagram

by Diane Pozefsky

This article introduces the concepts of storage networking and two key technologies used in today's storage solutions: the storage area network (SAN) and network-attached storage (NAS). It provides an overview of storage networking, and explains how the storage industry evolved to that model. It also covers the basics of SAN and NAS, and looks at where we see the industry and technology heading.

Separating Storage from Servers

It has always been true that a corporation's key IT asset is its data. When storage was tightly coupled to the server on which the data would be processed, the data and storage were simply subsumed under concerns for the server. Network design focused on availability of server access; scalability and manageability were defined by the server characteristics. Today, we are looking at significant changes to the storage environment. These changes involve separating storage from servers to create an environment that is generally referred to as *storage networking*.

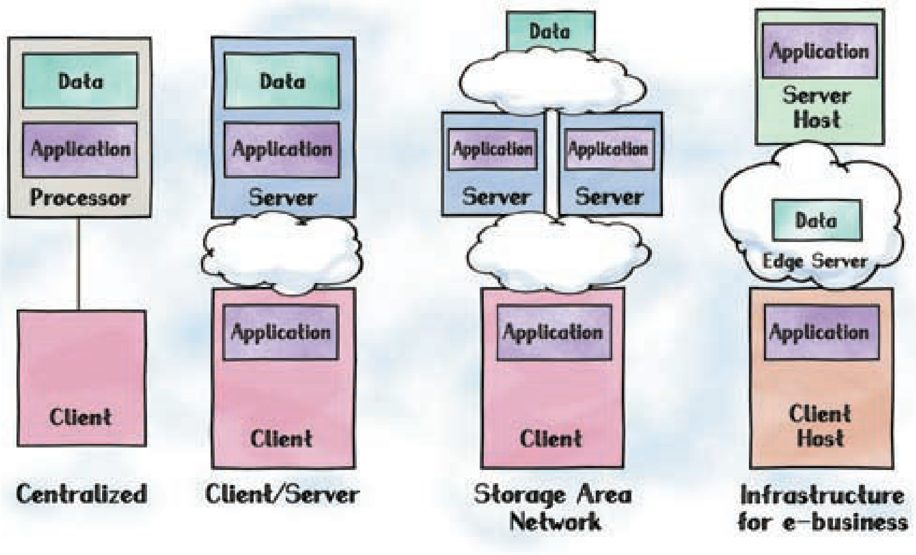
The move from server-based storage to network-based storage is as significant and fundamental a change as was the move from the centralized mainframe model to the client-server model or from client-server to Internet. But the separation of the storage is only the first step. We also want to treat the storage in the network as a single pool, leading to the virtualization of storage.

Once the storage-server axis is broken, we can begin to treat the information and the copy of the information as separate, and start looking at different ways to manage and access the information. This leads to the work in content distribution. Two of the key technologies of storage networking are the infrastructure model, known as *storage area networking (SAN),* and special-purpose file servers, known as *network-attached storage (NAS).*

Evolution of Information Access

In the centralized computing model, the complete application and all the needed data are stored and used in a single location — the mainframe. This is a very simple and efficient model. Access to information is local and there is no need to determine how to access data — it is all local. All that was necessary was to decide where the server should be located and to validate that network access allows users to reach the server with appropriate response times.

The advent of the client-server model did not affect the use of storage, but moved applications from a centralized model to a decentralized model. With the application running on distributed clients, access to data began to take on the first concerns of networking — how much data needed to be transferred, and was the network bandwidth and reliability sufficient for the transfer of data? The location of servers took on new levels of complexity, and network designers started to design the network with concern for data transfer.



The move to the World Wide Web exploded the amount of data being stored. Large corporations started seeing huge amounts of data, in the form of Web pages, added to the content that they needed to manage. This data increase was a significant motivator of the storage consolidation efforts seen, for example, in the growth of SANs.

The growth of e-business also introduced the notion of doing more for the end user. For example, once a customer enters any information into the system, well-designed Web sites continue to use this data in multiple ways and through multiple phases of processing. In some cases, the data is transformed and copied, but more often, the data is shared among different applications. As data sharing grows, the problem of keeping data with the application grows. How many applications must reside on the same server? Data sharing further enhanced the consolidation model and added the need for specific file sharing. While file sharing is not especially new, we saw a new surge of interest in the area with the growth of NAS.

The Changing Requirements of e-business

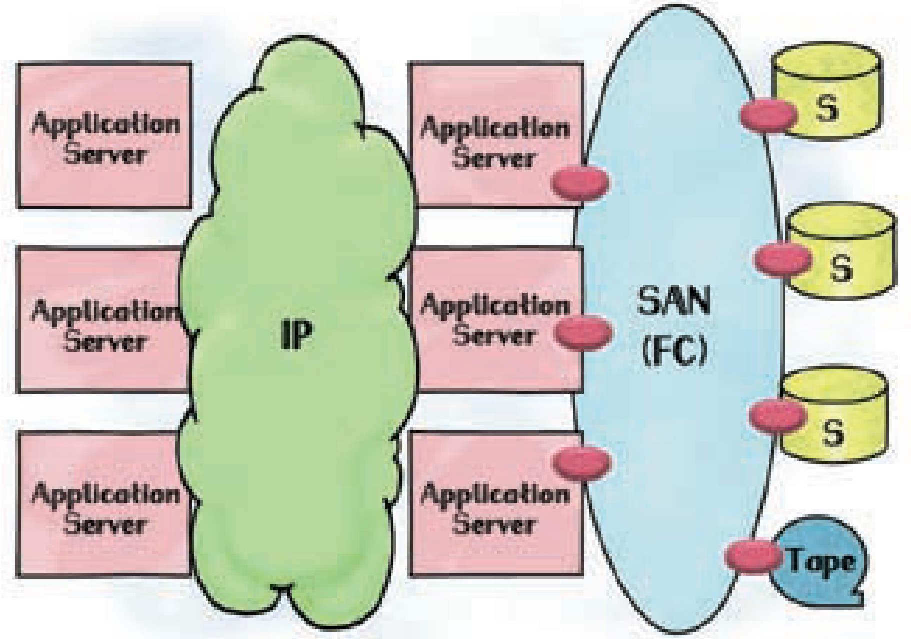
As the world moves to a rapidly growing dependence on e-business, the infrastructure for e-business is adapting to meet these needs. What are the changing requirements? The most significant is that information must be available reliably and efficiently wherever and whenever it is needed. The winners in the Web world are those who can keep customers returning to their sites. One of the keys to strong Web sites is that they are responsive. Users will come back to sites if they can do what they need to do efficiently. If it takes too long to check the weather, there are other solutions to find the information. If there are half-a-dozen sites that sell the same item at approximately the same price, users go to those that are the quickest.

What do these requirements mean for the world of storage? While companies still need and want very tight control of their data and information, they cannot afford to work with a single-copy model. There must be multiple copies to provide reliability, availability and appropriate responsiveness. Reliability and availability are nothing new to the corporate world, but managing response times is a significantly different problem from the one it used to be.

When people accessing the information can be anywhere around the world rather than in a few well-defined locations, problems can no longer be addressed by network design alone. And it is not just consumers who create this stress: business partners, suppliers, and businesses that are customers all require this same type of access. Even employees are no longer in well-defined locations. With business travel at an all-time high and the growth of the “teleworking” community, even that landscape has changed dramatically.

The need for responsiveness implies that there must be additional copies of data, and the need for control requires that there be synchronization rules for the data. There are multiple models, all the way from all data being copied everywhere to the simplest caching models. Models for distributing this data will continue to evolve and emerge, and the optimal selection of the distribution model will become a competitive advantage.

As the complexity of storage has increased, the skills available to manage it have shrunk. The old adage therefore applies that we need to learn to work smarter, not harder. The first step toward addressing these problems was the consolidation of storage into SANs and NAS devices.

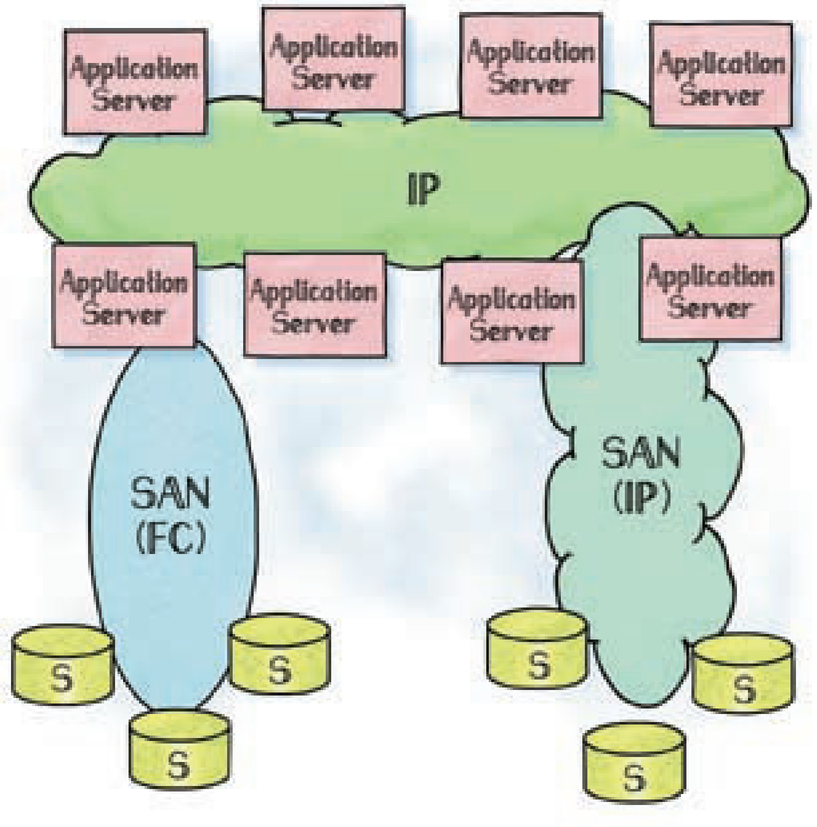


Storage Area Networks

When we refer to consolidation, the physical consolidation of the storage is not sufficient. Not only must the storage be brought physically together to be able to better manage it, but the storage must be accessible as a single pool of storage that can be used by whatever servers and applications need it. Also, the user must be assured that the storage is safe from inadvertent access or damage.

The first step is to create the physical pool of storage with a SAN. This is not a specific product, but an infrastructure for shared storage. In today's environment, SANs are implemented on a Fibre Channel base, but the fundamental principle of SANs is not tied to the media used. The fundamental characteristic is that the storage is on a network rather than directly attached to a server. The implications of this structure are that multiple servers can be physically on the SAN and have access to the same storage device, and that such a server has access to multiple devices. This structure provides the scalability and manageability that businesses need as their storage requirements grow. SANs also enable such key functions as backup and copying that do not interfere with either the servers that use the data or the network on which the corporation's other traffic is flowing.

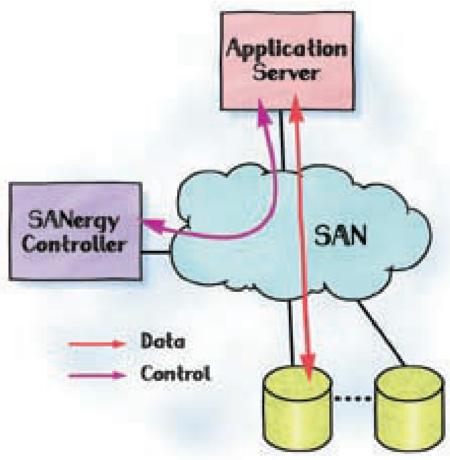
As already noted, consolidation requires more than physical consolidation, but before moving on to these additional requirements, let's spend a moment on the medium that is used. Today, SANs are built on Fibre Channel, and this restricts the placement of the servers that can access those storage devices. Although Fibre Channel extenders do exist, there are many environments in which they are not sufficient. For example, a company with a large number of remote offices will not be able to have servers at the remote offices directly accessing fibre-attached SAN storage. This is where IP storage comes into play. IP storage provides all of the benefits of SANs — the consolidation with its scalability and manageability — while extending the access to this storage across the wide area network (WAN).



Needed Enhancements

Returning to the requirements beyond physical sharing, a series of enhancements are needed to make the consolidated pool easy to use. In the industry, these enhancements are often referred to as virtualization. A simple definition of virtualization is the conversion of discrete physical entities into a single logical whole. Virtualization can occur at a number of different levels and we begin here by talking about disk, or logical unit (LUN), virtualization.

LUN virtualization covers those features that are needed to make this pool of storage behave as a single virtual disk. When a SAN has these capabilities, servers now connect to a virtual disk that can span multiple units. When new storage is added to this virtual disk, no changes are required at the server; the virtual disk simply gets larger. LUN virtualization also includes LUN masking, which can ensure that only one server has access to any specific storage.



Without LUN masking, every operating system attached to the SAN assumes that the entire SAN belongs to it. Clearly, this is not desirable and becomes one of the management complexities for a SAN. Virtualization removes this burden from administrators. LUN virtualization will be a continually enhanced function, not a one-time capability that is "done." The virtual disk and LUN masking are clearly the most important two features and therefore, the first ones to be provided, but as technology and business requirements advance, we can expect to see additional capabilities added at this layer.

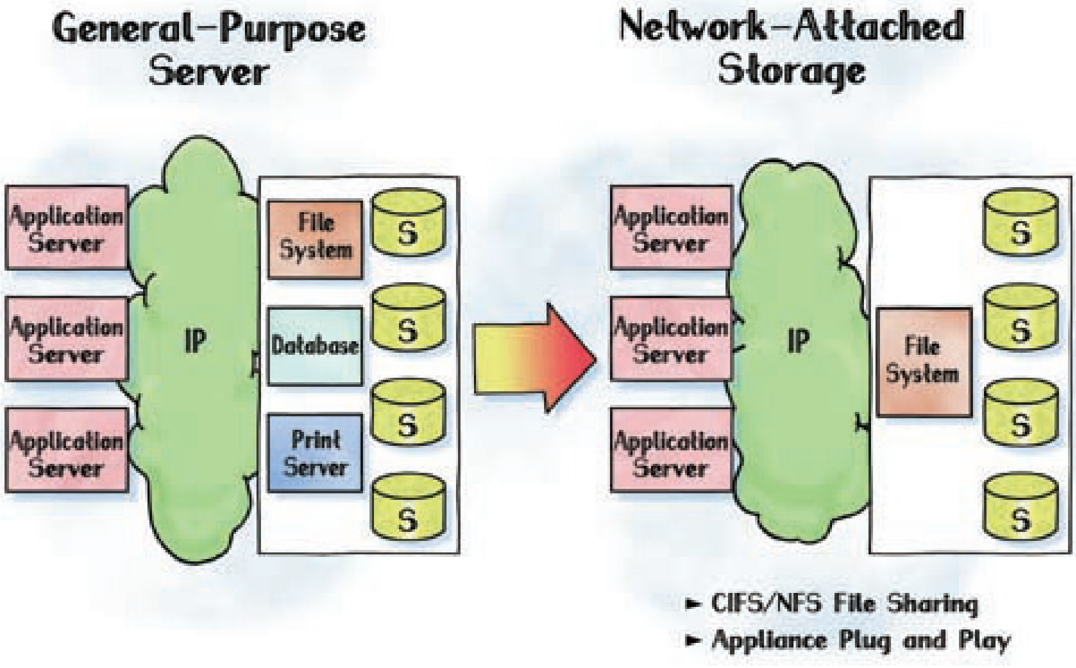
As virtualization is added to a SAN, there must be physical control of this logical entity, and that requires a single logical control point that is always used when requesting access to data. Clearly, such a single point of control can cause performance problems, but that is not necessarily the case. With products and technology such as the Tivoli® SANergy™ software, that single control point needs to be accessed only when the file is opened. By returning appropriate information and authority to the requesting server, additional access to the file can be directly from the server to the data.

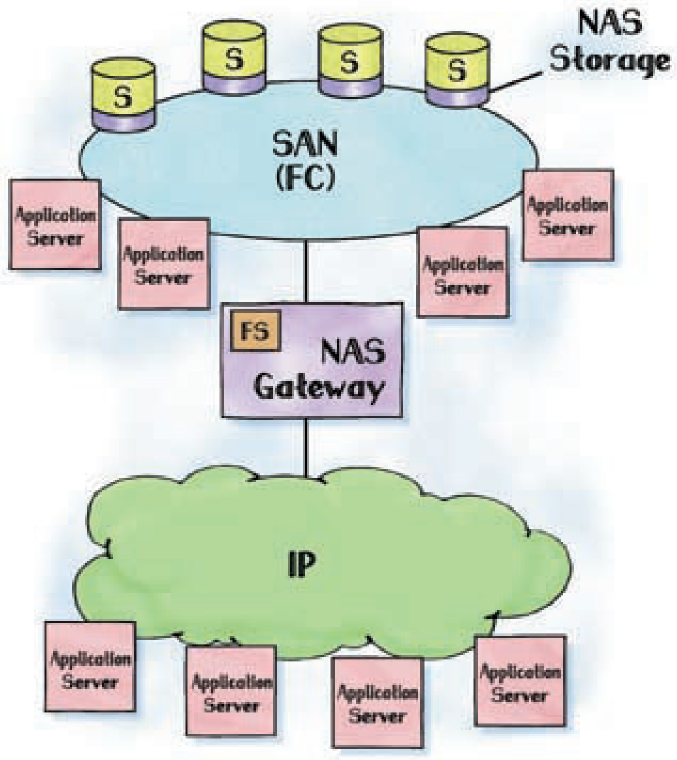
Network-Attached Storage

Like the SANergy software, NAS devices are built on the file paradigm and provide the important function of file sharing. IBM has delivered distributed file systems, Andrew File System (AFS®) and Distributed File System (DFS™), for a number of years. These systems allow multiple users to access the same file, and allow the user to access a file without knowing where it is located and whether the file being accessed is a primary copy or a secondary copy. Once the environment is set up, all copies of the files are automatically kept up-to-date. The key advantage to users of the system is that a widely distributed audience can all get local-access performance to the files.

NAS devices are simply file servers. File systems and file servers have been around for a long time. The primary reason why businesses began using file systems was for sharing of information. When many people or applications require access to the same information, it is more efficient to have a single copy that everyone knows is the most recent. File servers were originally just another component that ran on general-purpose servers. As the use and importance of file serving grew, it was recognized that much of the processing power of the server was being dedicated to file serving, and that the overhead and complexity of the general-purpose server was not only unnecessary, but indeed added to the administrative burden. Given those two observations, special-purpose file serving devices — NAS appliances — were born.

There is clearly a need for these appliances. IBM's current focus is to bring the benefits of the NAS environment to those customers who also need the consolidation and scalability of SANs. We refer to this configuration as a *NAS gateway* because the device does not come with its own integrated storage, but simply provides the access to existing SAN storage. Our goal is to integrate these two technologies that customers have found so valuable: the scalability and manageability of large amounts of storage on a SAN with the simplicity and sharing of file systems provided by NAS. When a NAS gateway is placed on a SAN, it is typically not the only server on the SAN. The existing application servers can continue to use the SAN as usual. The NAS gateway is simply another server that can share the pool of storage. As the need for more NAS storage grows, the NAS gateway simply expands its share of the SAN-attached devices.





A NAS device, whether containing integrated storage or a gateway to a SAN, provides a single access point to a large quantity of storage and the files on it. Our NAS product line features a shared file system and a cluster model that allows multiple servers to serve them. With the load distribution capability of accessing any of these servers through a single address, the server needs to know only that the file requested is available through this NAS server. The actual server that will be used, the volume that it is stored on, and whether it is on direct-attached storage or a SAN are all transparent to the user. Because of this, the NAS cluster itself is able to move files when new storage becomes available, and provide backup and mirroring capabilities transparently to any user.

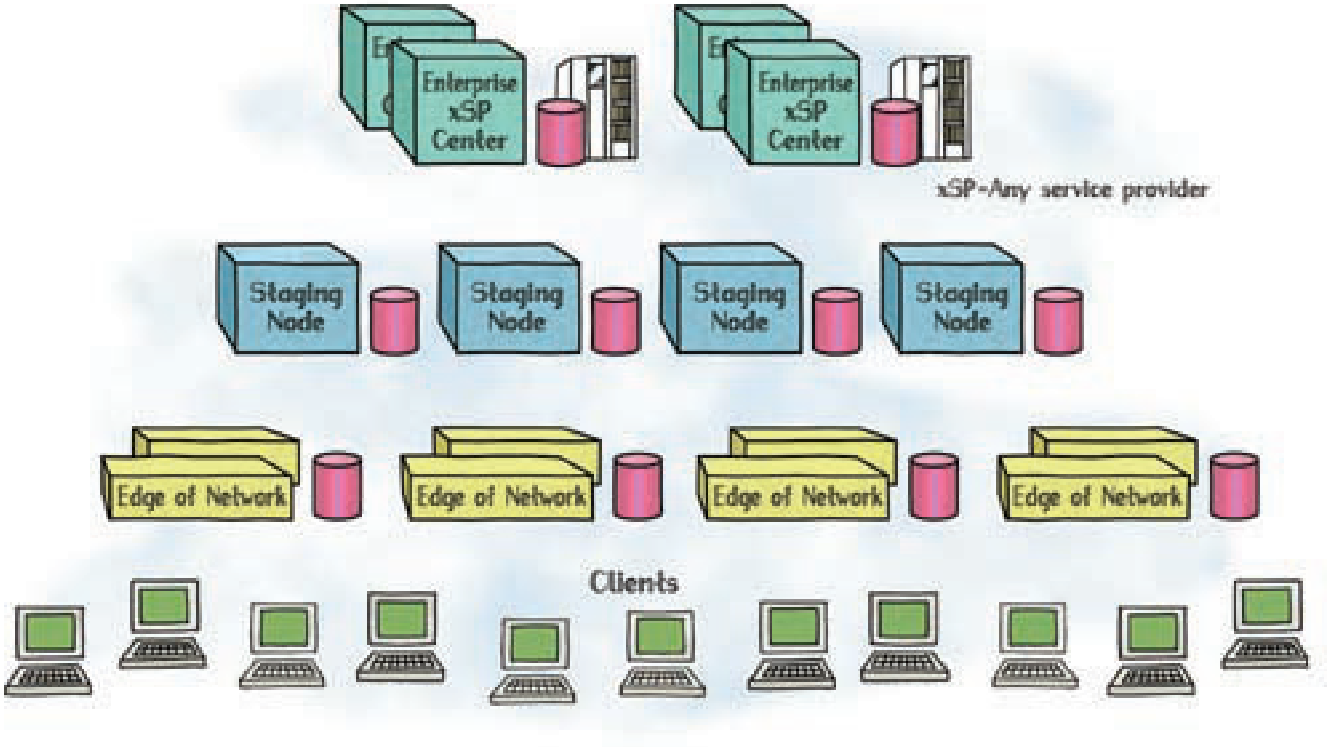
A word of warning about shared file systems: There is no magic in either distributed file systems or shared file systems that allows multiple parties to simultaneously update a file. At the most basic, all of these systems ensure consistency by prohibiting this from happening. This same distributed file model is needed and expected in our NAS environment. Although each NAS cluster will have a shared file system, there will often be multiple geographically separated clusters. In this environment, you will want to be able to think of the multiplicity of NAS environments as a single file system, and you will often need the performance advantage of the closer file access. We refer to this as a *federation of NAS clusters.*

Content Distribution

Content distribution is in many ways simply the next step beyond virtualization. Not only should you be able to think of all the storage in the network as a single pool, but the network should be able to manage the placement of the information that is stored in a manner that meets the criteria that you have established for that information. The criteria that you need to establish include:

- The level of availability of the data
- The backup requirements (for example, the window of lost data that can be tolerated)
- The performance requirements for accessing this data (for example, latency)
- The constraints on assuring that all accesses to the data get the most current information

Some examples of the different types of requirements will help explain this. For a database such as current product information that is updated every evening and accessed constantly throughout the day on a worldwide basis, the availability and latency are critical, but a backup copy of the master data is required only once a day. The implication is most likely that you would want to replicate multiple read-only copies of the information immediately after update. On the other hand, data that tracks activities happening during the day, and is therefore updated throughout the day through a batch process from around the world, but needs to be always available and always correct, would be more apt to exist as a single copy that is constantly mirrored. The goals for content management are that the operational activities to provide these capabilities be automated, based on policy statements, and that this type of management can be focused on the policy definitions.



Key components of content distribution include the management of the data and the timely and efficient distribution of the data where needed. There are a number of options that can provide the replication function. In some cases, you will "push" the copies out to edge locations in advance of any requests for them and always keep them up-to-date as changes occur. This would be the traditional replication function. Variants on this model include not replicating until there is a request for data but then keeping it up-to-date as changes occur. This variant is useful if only a small portion of a large collection is likely to be used, but once it is used, it is likely to be accessed multiple times. This capability might be called *intelligent replication.*

On the "pull" side of getting the information out to the users, the basic function is caching. The assumption with caching is that if there is one use of information, others are to follow. Here, too, we can add intelligence to the function. The caching policy can be adapted, based on history or on the types of data that have been requested. It is also possible to "prefetch" based on the same characteristics or information stored with the data. For example, when a Web home page is loaded, it may be worth prefetching all of the pages at the next level. Once one of the next-level pages is loaded, it may be worth prefetching more depth than just the next level.

For most companies, the most important data stays in their data centers. These use tightly managed SANs with data backed up, archived, and protected.

The replicated or cached information is stored in a device that we refer to as an edge server. Besides this function, the edge server will support a number of capabilities needed for pervasive devices. Two of these key functions are personalization and transcoding. Personalization is the specification of what information is relevant to the recipient.Transcoding refers to the functions that take a standard Web page and convert the presentation to match the device to which it is to be sent (for example, screen size, graphic capabilities, and color scale).

When considering personalization, one recognizes that there is a hierarchy of information. You can imagine, for example, that there is a locality of news that is likely to occur. Most requests for U.S. news are going to come from within the U.S., for state news from within the state, and for local news from within the local area. It is therefore important that the storage and data hierarchy be multilevel to get the most benefit from caching.

To pull together the technical description of storage devices needed to support content distribution, there are consequently four tiers of devices to consider:

- The master copy of storage stored or created at a single central location
- The cached or replicated copy at the edge of the network, which provides the needed response time and supports streaming multimedia
- Potential intermediary caching and staging points to take advantage of commonality
- The actual clients

As one considers the environment described for content distribution, it is unlikely that most corporations will actually deploy this type of network on their own. It is impractical to expect most companies to have a network that reaches the breadth that this access implies. Therefore, it becomes a significant service provider strategy to create this part of the storage networking infrastructure.

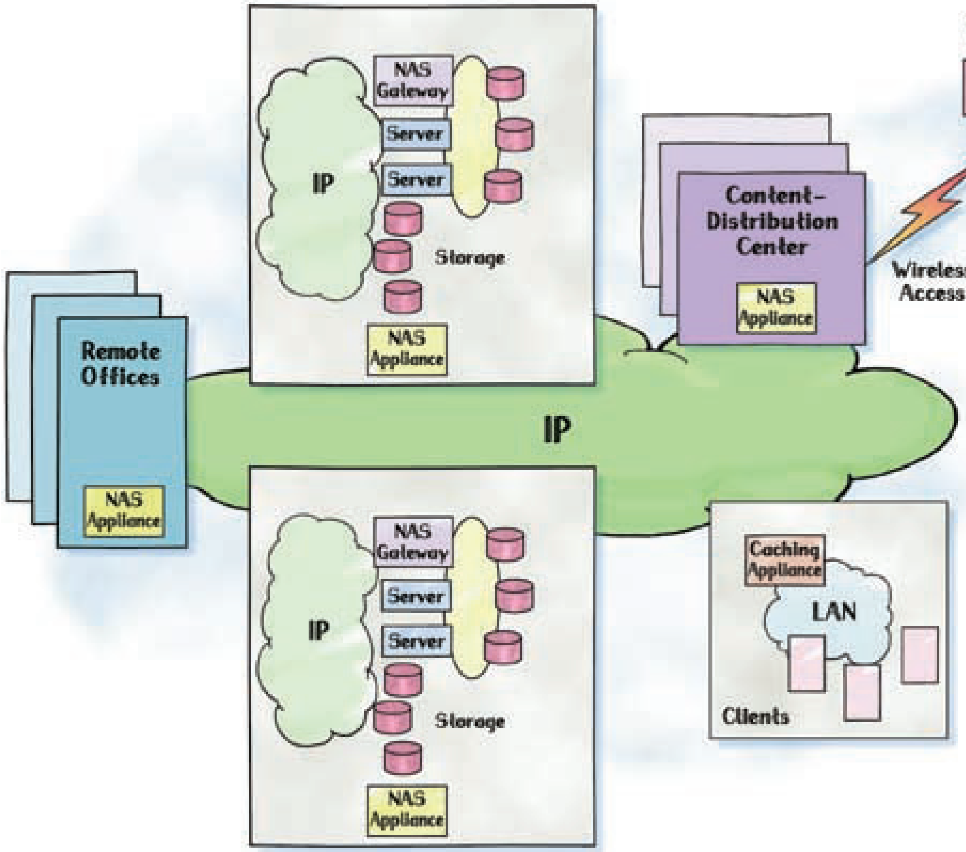
Policy-Based Management

With all of this technology to place storage where it's wanted and provide access to the data as it's needed, the administrative cost and complexity of this environment can still be overwhelming. You need to deal with the management and security of data at the business level:

- What is the value of the data to the corporation?
- How critical is it that there be no lost data?
- What is the cost to the business if the data is not accessible for some amount of time?
- Where do I need to keep archives and how many versions are needed?

These are the questions that the business wants to focus on — independently of the type or location of the data. To have this capability, you need policy-based management.

Policy-based management exists today for areas outside of storage, and for storage, it exists in specific environments. For example, Tivoli SecureWay® Policy Director provides these types of capabilities for security access, and storage products such as Data Facility Storage Management Subsystem (DFSMS) for the storage that they manage. Where we need to be is a single-policy definition of storage that supports all types of data (Web pages, databases, electronic mail, and all other files as well) independently of where they are stored. The policy management should drive key products, such as Tivoli Storage Manager and Tivoli Storage Network Manager, and support all major vendors' storage. For policy-based management to be most valuable to users, it must be based on open standards with proven interoperability so that it is able to deal with all of the business data in a holistic manner.



As already stated, policy must address all data, no matter where it is, and a corporation's data is spreading far and wide. It is no longer all sitting in the data center. Let's take a look at what is happening to corporate data. For most companies, the most important data stays in their data centers. These use tightly managed SANs with data backed up, archived, and protected. Most of the content is managed here. But leaving the data just at the data center no longer works. Remote offices that need fast access to data often can't handle the network delays that are inherent in remote access. Corporations therefore want copies of that data at the remote sites, but they can't afford to manually ensure that the data is up-to-date. The policy-based management needs to take on the responsibility of ensuring consistency.

In other cases, those response-time requirements do not exist, and the remote sites want direct access to all the data — both data that is accessed through file systems and data that is accessed directly on a SAN. In this case, IP storage can be used. For other organizations or for other data, the storage might not be at the enterprise at all, but at a storage service provider. In this case, the requirements still remain the same — the service-level requirements become the responsibility of the SSP, but the business still wants to manage the policy in a single place with a single set of decisions.

As access to information becomes ubiquitous across the Web, many of these same requirements surface for the information going out to the company's customers and other audiences (stockholders, press, business partners, and so on). Here, there is not the option of the corporation building the breadth of network to meet these needs, so they will require the services of content-distribution providers who can take advantage of their global presence to provide the performance required.

No matter the size of the organization or the business, or whether the storage is maintained in-house or at a service provider, we see the volume of data and the needed access to it growing at ever-increasing speeds, and these requirements for easier management of data and storage becoming critical to the growth of business. Our goal is to provide the products and services to tame the information beast. ■

About the Author

Diane Pozefsky is an IBM Fellow and is currently Director of Storage Networking Architecture in IBM Storage Systems Group. Diane joined IBM in 1979 and spent most of her career in networking. She was one of the lead architects for APPN and team leader and architect of AnyNet®, which permits customers to run SNA applications on IP networks, and IP applications on SAN networks. She also worked on the network for the Nagano Olympics and on networking issues for the internal IBM network prior to taking her current position.
dpoz@us.ibm.com

iSCSI: Furthering IBM Storage Networking Leadership

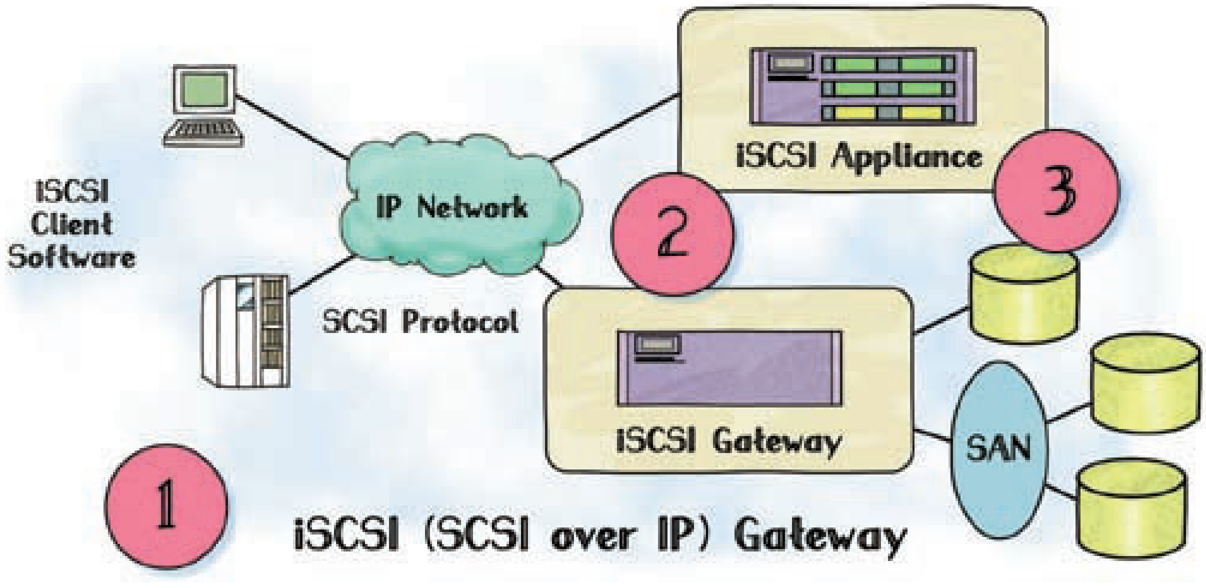
By Jim Tuckwell

Today's incredibly fast-paced, ruthlessly competitive, global business environment has led to the ever-increasing strategic importance of leading-edge e-business applications and the transformation of data into meaningful information with which to fuel competitive advantage.

This need for information, and the exponential growth of stored data to provide it, have led to a transformation in the IT industry. Not long ago, servers were viewed as the central component of the network, and storage was often seen as an afterthought. Today, an organization's data is the heart and lifeblood of its network. Providing ease of access to exploit this asset, in a secure environment, is crucial to success.

While pursuing the promise of the competitive advantage that e-business applications can provide, many customers, if not the vast majority of them, have seen significant growth in the numbers of servers and their associated databases in their networks. This phenomenon has led to the growing complexity of systems and storage management, increasingly inefficient storage utilization and increased cost of ownership. These issues are compounded when, as is the case with many customers, support staffs often have very limited resources. The deployment of storage area networks (SANs) is designed to address these requirements, and increasingly, customers are seeing improved availability and service levels, improved administration, increased flexibility and lower cost of ownership as a result of implementing SANs. However, these same requirements exist throughout many areas within organizations that are not served by SANs, as well as in organizations where Fibre Channel SAN networking technology isn't justifiable.

Remote locations, departments, and small- to medium-sized businesses are increasingly voicing the need for directly addressable, pooled storage solutions, optimized for database-intensive applications, which can be deployed over familiar IP "fabric." In effect, they require a solution often referred to as *SAN over IP.* What these customers require is a lower-cost, more-easily managed and more-affordable approach than currently available SAN solutions. In addition, these customers typically don't require the enterprise-level scalability, reliability and performance that are the hallmarks of SAN technology. These customers also demand that pooled storage solutions be based on open standards so that they can maintain the maximum in flexibility and control.



What is iSCSI?

Developed by IBM Research in Almaden, CA and Haifa, Israel, "iSCSI" is a new technology that provides the encapsulation of SCSI commands with TCP/IP for transport over IP network "fabric." iSCSI product development is currently taking two different paths. One approach is based on developing an iSCSI-based storage "appliance." These devices include imbedded storage, and utilize iSCSI technology to attach directly to the IP network. The second approach features a "gateway" that utilizes iSCSI technology and provides an "extension" for Fibre Channel SANs, to link users of the Fibre Channel SAN environment and users in the IP network environment. These gateways do not contain internal storage.

Each of these approaches provides three separate components:

Initiators (1): These are the "device drivers" that reside on the client. They intercept SCSI commands, encapsulate them with TCP/IP, route them over the IP network and provide access to the "target" device.

Target Software (2): This software receives iSCSI commands from the IP network, and can also provide configuration support, storage-management support and other capabilities.

Target Hardware (3): This can be a storage appliance containing internal storage, or a gateway that has no internal storage of its own, but extends the reach of the SAN to users on IP networks.

There are some in our industry who are eager to predict that IP-based storage will replace SAN technology. IBM doesn't share this view, but sees iSCSI as an exciting new technology capable of providing solutions that are an extension of, and complementary to, the benefits that SANs offer. Just as SAN features and functions continue to be enriched, IBM expects the capabilities provided by iSCSI technology to grow, as a number of different factors evolve over time, such as:

- The introduction of 10-Gigabit Ethernet (expected late in 2002 or 2003)
- The introduction of Host Bus Adapters (HBAs) to offload TCP/IP overhead from storage appliances and clients, thus improving performance/ throughput (this is beginning to happen now)
- Improved network and storage management capabilities

IBM is committed to being a storage networking leader. Nowhere is this commitment more self-evident than in the announcement of the IBM TotalStorage™ IP Storage 200i in February, 2001.

IBM Storage Networking Leadership

IBM is committed to being a storage networking leader. Nowhere is this commitment more self-evident than in the announcement of the IBM TotalStorage™ IP Storage 200i in February, 2001. One of the first iSCSI solutions to be introduced into the market, this new storage "appliance" clearly extends IBM's leadership in IP-based storage networking research, development and open standards to deployable customer solutions. The IBM TotalStorage IP Storage 200i positions IBM as a technology and open-standards leader, and is IBM's first step in delivering solutions for the emerging market for pooled storage over IP networks.

Designed for environments that don't require and cannot justify the bandwidth, enterprise-level reliability, or scalability provided by Fibre Channel SANs, the TotalStorage IP Storage 200i begins to bring the advantages of SAN technology to users on Ethernet and Gigabit Ethernet "fabric," and is specifically engineered for:

- Departments and workgroups
- Remote branch/plant sites
- Service providers
- Small- to medium-sized businesses

The IP Storage 200i can provide an excellent pooled-storage solution if you have multiple servers and your storage-management requirements have grown in complexity to the point where your IP staffs lack the resources or skills to effectively manage them. Its use of the familiar IP fabric makes this approach more manageable and affordable in such a situation.

IBM TotalStorage IP Storage 200i

There are three models of the IP Storage 200i, the Model 100, 200 and EXP (expansion unit). All feature RAID reliability and a preloaded operating environment (microcode) based on Linux, and are designed for block I/O storage. Initiators are downloadable from the Web at no charge and support Windows NT®, Windows® 2000 and Linux clients.

IBM is committed to iSCSI as the open standard for the transport of block I/O over IP and, as a result, actively participates with other industry leaders in the Internet Engineering Task Force (IETF) standards efforts. The standards definition is not yet finalized, although a final proposed standard is expected in October of this year. IBM's iSCSI strategy includes the adoption of this proposed standard.

IBM TotalStorage IP Storage 200i Model 100

The IBM IP Storage Model 100 is a tower configuration that includes 36.4-GB hard disk drives (three to six, providing from 108 to 216 GB). The Model 100 features an 800-MHz Pentium® III processor, takes advantage of standard IBM components to minimize cost, and is bundled with microcode that provides a browser interface for straightforward configuration and management. With a suggested retail price starting at USD19 995, the Model 100 is an excellent fit if you want to begin taking advantage of the benefits offered by an IP-based, pooled-storage solution, and position yourself for the future.

IBM TotalStorage IP Storage 200i Model 200

The IBM IP Storage 200i Model 200 comes as a tower configuration and is designed for environments requiring increased performance. The Model 200 offers up to two 800-MHz Pentium III processors, and imbedded storage ranging from 216 GB to 1.74 TB (utilizing the same 36.4-GB hard disk drives as the Model 100). Suggested retail prices range from USD40 000 to USD100 000.

IBM TotalStorage IP Storage 200i Model EXP (Expansion Unit)

The IBM IP Storage 200i Model EXP provides additional storage capacity for the Model 200. It holds 14 slim-high 36.4-GB hard disk drives (ships with 3), providing up to 509 GB storage capacity per unit. Up to three Model EXPs can be attached to a Model 200, providing 1.7 TB of storage. Each Model EXP has dual 250 W, hot-swap, redundant power supplies.

IBM Storage Networking Alternatives

IBM offers multiple approaches to extend the advantages offered by storage area networking technology to users throughout the entire organization. Each of these approaches is designed and optimized for specific requirements. When file serving and file sharing is required, NAS solutions are the recommended approach. The xSeries 150 NAS family, and the groundbreaking new IBM TotalStorage Network Attached Storage 300G file-serving gateway offer outstanding solutions.

If you require solutions optimized for database performance, block I/O approaches are recommended. In addition to SAN solutions, iSCSI solutions are also block I/O in design and are recommended. The focus of this article is the IBM TotalStorage IP Storage 200i. However, Cisco Systems is also a leader in iSCSI technology. Their first iSCSI solution, the Cisco 5420 Storage Router, was announced in April. The 5420 is a gateway that can offer a bridge between users of Fibre Channel and IP networks, and is sold by IBM as well as Cisco Systems.

IBM IP Storage 200i Model 100

IBM IP Storage 200i Model 200

There have never been more options available to you to tailor your storage environment to maximize service levels and obtain lower cost of ownership. IBM is in a unique position to offer solutions covering the entire storage networking spectrum. These solutions offer direct access to data, be it connected through a SAN or on IP "fabric," and we also provide options optimized for file sharing as well as database-intensive environments.

Not only is IBM's storage networking strategy designed to cover the depth and breadth of an enterprise's requirements, it also calls for the continued introduction of exciting new capabilities.

Positioning — Which Approach for Which Requirement?

Where do each of these solutions best fit? To address this question, and better understand which pooled storage solution is optimal for a particular requirement, it is best to begin with the business problem that we're working to solve. Every situation is different. There will be exceptions, but in general, the following guidelines are good to keep in mind when evaluating requirements and identifying where to begin to find the best approach:

- If the requirement being addressed is focused on end users, and needs to address the problems inherent in the day-to-day management and support of PC clients (which tend to be file-oriented), then typically the optimal solution will be a NAS solution, which is optimized for file serving and file sharing.
- If the requirement evolves around servers, and addresses the problems inherent in the day-to-day management and support of storage accessed by database-intensive applications, then typically the optimal solution will be a SAN or iSCSI solution, which is optimized for block I/O, database-intensive environments.
- If enterprise scalability, reliability and throughput are required, then clearly, a SAN solution offering the optimum in scalability, bandwidth, service levels, and network management software should offer the best approach. In this environment, extending the SAN through the installation of a gateway might prove to be an ideal "marriage," offering SAN benefits to users not directly attached to the Fibre Channel network.
- If installation of a Fibre Channel SAN isn't a realistic alternative, due to initial cost, complexity, support skill investments, and so on, then the IP Storage 200i or IBM's NAS family should be the recommended solution.

The following table summarizes IBM TotalStorage IP Storage 200i Model 100 and 200 features.

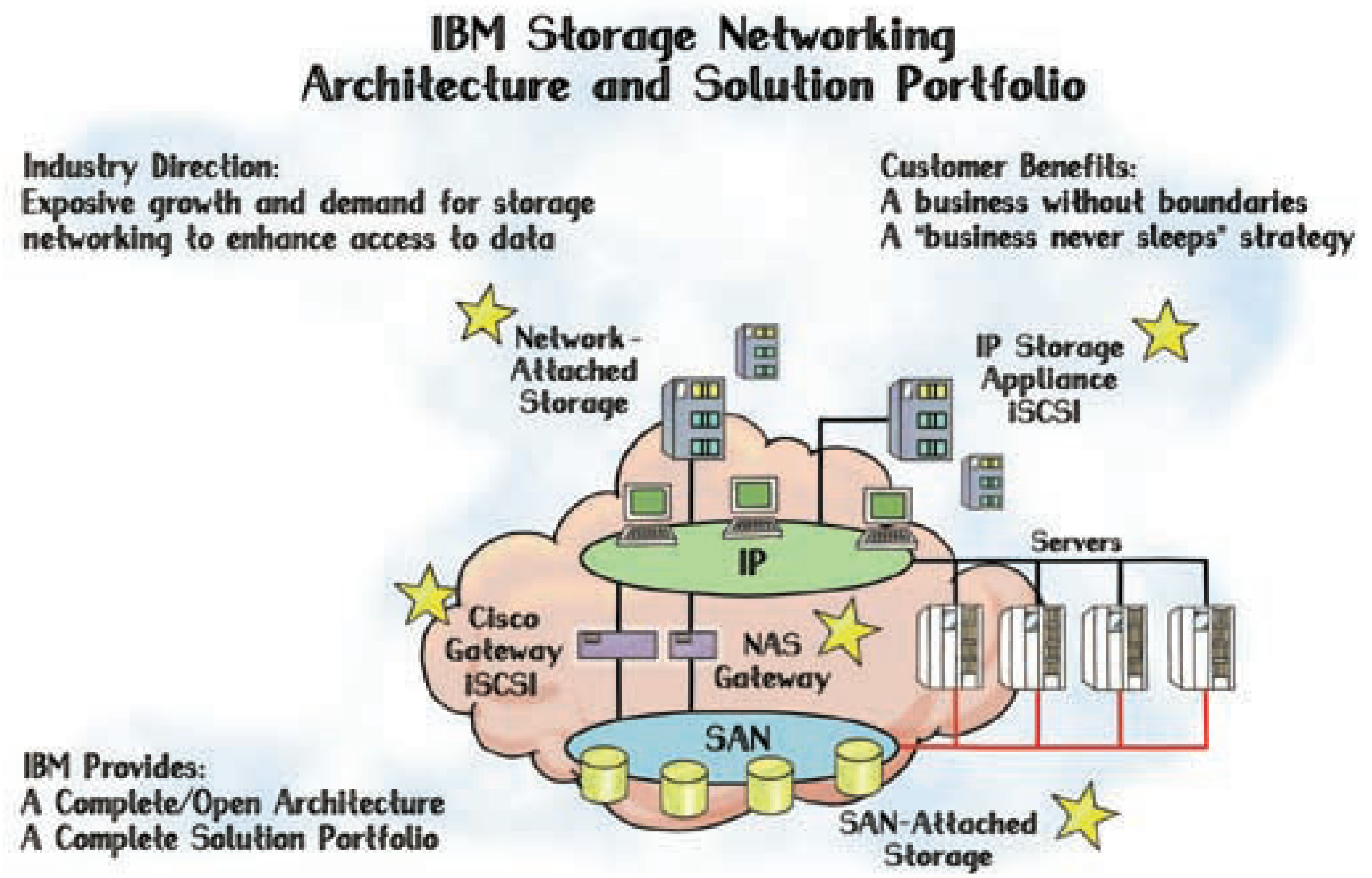
Feature	**Model 100 (4125-100)**	**Model 200 (4125-200)**
Scalability	108 GB - 216 GB	216 GB - 1.74 TB
Drives	3 - 6	6 - 48
Engines/nodes	1	1
Processors	1	2
Protocol attachment to hard disk drive (HDD)	SCSI	SCSI
Redundancy	Hot-swap HDD Hot spare Hot-swap fans Redundant power supply	Hot-swap HDD Hot spare Hot-swap fans Redundant power supply
Backup	Normal system procedure	Normal system procedure

Why IBM?

IBM is committed to being a leader in open standards-based storage networking solutions. IBM has a depth and breadth of products, and complete solution offerings that we believe you will find unparalleled anywhere in the industry. Finally, IBM offers the best depth and breadth of support and service, and complete consulting and integration services designed to help you realize the best ROI.

Storage networking is an amazingly fast-growing area of our industry. It offers an exciting and fast changing landscape. The Web sites on the following page offer additional information.

What customers require is a lower-cost, more-easily managed and more-affordable approach than currently available SAN solutions.


IBM Storage Networking
Architecture and Solution Portfolio
Industry Direction:
Exposive growth and demand for storage
networking to enhance access to data
Customer Benefits:
A business without boundaries
A "business never sleeps" strategy
Network-
Attached
Storage
IP Storage
Appliance
iSCSI
Servers
IP
Cisco
Gateway
ISCSI
NAS
Gateway
SAN
IBM Provides:
A Complete/Open Architecture
A Complete Solution Portfolio
SAN-Attached
Storage

IBM Corporation
ibm.com

IBM Storage Systems Group
ibm.com/storage

IBM Storage Networking
ibm.com/storage/snetwork/index.html

IBM Tivoli Storage
tivoli.com/search/query.html?qt=storage&col=products&submit=Search

Storage Networking Industry Association
snia.org

Internet Engineering Task Force
ietf.org

Cisco
cisco.com/go/ibm

Cisco storage
cisco.com/warp/public/756/partnership/ibm/storagenetworking.shtm

Joint IBM & Cisco Web site
cisco.com/ibm

Cisco AVVID Partner Program
cisco.com/go/avvidpartners

Fibre Channel Industry Association
fibrechannel.org

ANSI
ansi.org

SCSI Trade Association
scsita.org ■

About the Author

Jim Tuckwell joined IBM's General Systems Division as a marketing representative in 1975 after earning BBA and MBA degrees, majoring in marketing, at the University of Wisconsin. Jim has held a number of sales and sales management positions covering new business through Fortune 100 customers. In 1992, Jim joined the AS/400 Division, where he helped create the AS/400® Partners in Development team. There, Jim created the TLC program, designed to better support ISVs and directly link IBM developers with their counterparts in the industry who were developing applications. Jim joined the Storage Networking Division in November, 2000 and has worldwide marketing responsibility for IBM's initiatives involving iSCSI storage networking technology.
jwtuckw@us.ibm.com

IBM TotalStorage Network Attached Storage 300G Answers Small and Medium Business Needs

By Daniel Powell

IBM's Storage Networking Division recently unveiled two new products designed to help you deal with burgeoning growth in storage.

Two models of the IBM TotalStorage™ Network Attached Storage 300G are now available to enable you to connect an IP or Ethernet network with storage area networks (SANs). The IBM NAS 300G will act as a versatile link, allowing IP-attached clients running applications based on Windows®, UNIX® and NetWare to share data stored on a Fibre Channel-based SAN.

The growth of SANs is primarily due to the need to control the rapid expansion that most companies are experiencing in storage requirements. Today, many customers are faced with an infrastructure consisting of isolated pockets of servers and their direct-attached storage. SANs offer the combined benefits of centralized management and efficient assignment and use of storage. This is also a prime reason for the recent explosion in demand for network-attached storage (NAS) devices that combine the access features of the IBM NAS 300G with integrated storage (DASD) in a single device.

The IBM NAS 300G connected to a SAN provides NAS functions with huge scalability and, in many cases, the redeployment of existing storage devices. Although many analysts are taking a SAN-versus-NAS line, a better view is to see them as complementary architectures. Because IBM offers both SAN and NAS devices, we are in a position to create storage solution offerings that best suit your specific environment.

The single-node IBM NAS 300G Model G00 is shipped according to your configuration in a single rack-ready frame. You can attach to the built-in 10/100-Mbps Ethernet adapter or order any combination of additional 10/100 or Gigabit Ethernet PCI adapters. (The IBM NAS 300G has four available PCI slots.) The device also has one Fibre Channel port, and can be enhanced with one additional PCI Fibre Channel adapter to provide direct connection of a Fibre Channel device like a tape backup device. The Model G00 is powered by dual 933-MHz processors and a base of 1 GB of memory. This can be expanded to a total of 2 GB for maximizing NFS performance.

The Model G00 comes preloaded with numerous software elements specifically selected to focus this storage server on its single-use function — rapid file delivery. The operating system is Windows Powered OS, a derivative of Microsoft Windows 2000 Advanced Server. The IBM NAS 300G will ship with a finely tuned version of this OS that has been tweaked to maximize its rate of delivering NFS, CIFS, Novell, HTTP and FTP files. Complementing this are management and backup clients, a server box utility and a gateway configuration utility.

A key software application preloaded in the IBM NAS 300G is the Columbia Data Products Persistent Storage Manager, which allows users to maintain multiple backup images of files and to easily recover an earlier version of any file. The IBM NAS 300G is also preloaded and ready to license with Tivoli® SANergy™ software, which can provide performance enhancements for high-bandwidth requirements. For individual computers requiring higher-bandwidth access beyond the capability of any LAN connection, the SANergy software can be licensed and used to provide increased I/O performance to large data files located on the SAN.

The dual-node Model G25 is essentially two Model G00s that are linked together using clustering and failover-protection code. This design eliminates any single point of failure from SAN connection through to the LAN connection, and provides a fault-tolerant design, allowing continuous access to data. Should a failure occur, the automatic failover function will switch the failing node's data request to the second node, which will then provide data access to both nodes' files. The dual-engine model includes an active/active design which provides fault-tolerant features so that access to data is maintained even in the event of an engine failure. At the same time, a high performance level is maintained by utilizing both nodes for data delivery. This ensures that users will have access to data through the alternate engine. Each engine also provides a hot-swappable redundant power supply and fan for increased availability.

Both models of the IBM NAS 300G have been tested and verified to work with a number of SAN and Fibre Channel products. IBM will continue to test and verify compatibility of the IBM NAS 300G products with other SAN and Fibre Channel products, and the following list of compatible hardware products is not meant to be definitive:

IBM Enterprise Storage Server™ 2105 Models F10 and F20

IBM 2106 Modular Storage Server

IBM 2109 Fibre Channel Switch

IBM 2032 Fibre Channel Director

IBM 2042 Fibre Channel Director

IBM 7133 Serial Disk System

IBM 7139 SLIC Router

IBM Netfinity® FAStT500

IBM Netfinity FAStT200

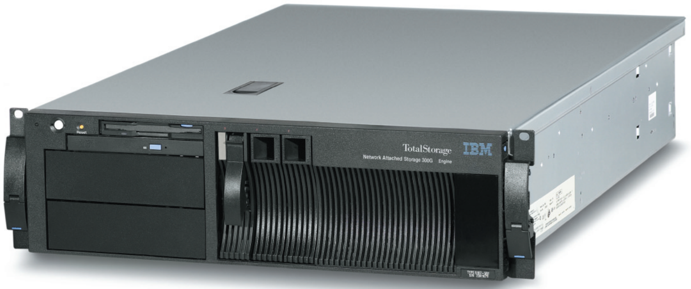
Compatible software includes Tivoli Storage Manager client, Veritas Backup Exec and Legato Networker.

You will benefit from the IBM NAS 300G technology in multiple ways. You may be able to consolidate or eliminate servers that are currently acting as file servers. The finely tuned IBM NAS 300G storage server has a significant performance advantage over a general-purpose server (in some cases up to 3 times better NFS file performance). You will also benefit from cost savings due to the elimination of Fibre Channel connections to multiple servers. The server consolidation requires a single Fibre Channel connection to the IBM NAS 300G, rather than multiple servers. Finally, the storage server appliance model eases product installation, configuration, support and management.

More and more customers are seeing the value of storage consolidation and are looking for appliances that ease the transition to the new world of open storage systems. The IBM NAS 300G answers the call by providing an easy-to-use, competitive, high-availability (Model G25) offering that delivers great total-cost-of-ownership value. ■

About the Author

Dan Powell has 22 years of development experience in IBM. He started his career as a chip and board designer and has held numerous positions in development, management, planning, quality and project management. He is currently the Product Manager for Storage System Group's recently announced storage gateway products. dcpowell@us.ibm.com



IBM NAS 300G

IBM TotalStorage Network Attached Storage 200 and 300 – Efficient Solutions for Storage Networks

By Eric Dunlap

The IBM NAS 200 and IBM NAS 300 bring affordable, IP-attached storage to your business for use in a variety of applications. These additions to the product line provide you with one more building block to increase the flexibility, efficiency and effectiveness of your storage network. The IBM NAS 200 and IBM NAS 300 provide exceptional performance at affordable prices. They combine to provide a wide range of scalability, offering data storage ranging from 108 GB to 3.24 TB. Redundant subsystems enhance their value, providing a high-availability solution to ensure the continuation of your business operations even in the event of a component failure.

Why NAS?

Data storage began as direct-attached devices, such as the hard disk drive in a PC, but evolved with innovations in storage networking such as storage area networks (SANs) and network-attached storage (NAS). Although direct-attached storage provides good performance and is relatively inexpensive, it is limited in the amount of data it can store and constrains an organization's ability to share data. SANs have overcome these obstacles, providing a means of not only sharing data, but doing so with many clients, while simultaneously providing the highest level of performance. SANs also offer excellent data protection through their use of completely redundant systems. NAS combines the best attributes of direct-attached and SAN storage. NAS solutions, in particular the IBM NAS 200 and 300, create a solution that enables many clients to access the same data simultaneously, with exceptional performance, while also providing component-level redundancies to protect the data.

IBM TotalStorage™ Network Attached Storage 200 and 300 are the newest additions to the IBM Storage Networking portfolio of hardware, software and services.

The IBM Network Attached Storage 200 Series

The IBM Network Attached Storage 200 series is an affordable solution if you are undergoing server consolidations, or otherwise considering the use of NAS. Both versions of the IBM NAS 200 provide exceptional performance in Windows® environments and excellent mixed Windows/UNIX® performance. Their multiprotocol support eliminates the need to have separate servers for each supported protocol, further enhancing their ability to consolidate storage from multiple general-purpose servers. Redundant, hot-swap power supplies and hot-spare, hot-swap hard disk drives provide system availability even in the event of a subsystem failure.

These additions to the product line provide you with one more building block to increase the flexibility, efficiency and effectiveness of your storage network.

The included Persistent Storage Manager software creates True Image data views, which allow client file restorations in the event of an accidental file deletion. Persistent Storage Manager also includes an open file manager that allows online backup without the need to stop user access to the system. Installation of the IBM NAS 200 series has been simplified compared with general-purpose servers through a tightly integrated suite of preloaded software, allowing you to get up and running as soon as possible.

The single-engine IBM NAS 200 series offers two versions — a tower version with 108 to 216 GB of storage for workgroup-type applications, and a rack version with 216 GB to 1.74 TB for departmental and small enterprise environments. The IBM NAS 200 tower version uses an 800-MHz Pentium® III processor and a ServeRAID™ 4L controller with high-throughput SCSI hard disk drives. It can be upgraded with an additional 800-MHz processor and up to 2 GB of total system memory. The IBM NAS 200 rack version utilizes two 800-MHz Pentium III processors, 1 GB of system memory, a four-channel ServeRAID 4H controller and high-throughput SCSI hard disk drives. It can be upgraded to 2 GB of total system memory as well.

Environments and Applications

The IBM NAS 200 tower version is well suited for use in workgroups, including corporate headquarters locations as well as remote settings such as the branch offices of a bank, insurance company or retail facility. The low-cost-per-megabyte advantage of the IBM NAS 200 tower version also makes it an excellent choice for use in Internet service provider (ISP), application service provider (ASP) and storage service provider (SSP) environments where engine failover capability is typically not required.

The IBM NAS 200 rack version is best suited to departmental environments, including corporate headquarters locations and regional offices. Within these settings, the IBM NAS 200 rack version is an excellent solution for storing e-mail attachments, temporary storage of daily transactions prior to downloading to a corporate location, or as a means of running sales force support applications.

IBM NAS 200 Tower

IBM Network Attached Storage 300

The IBM Network Attached Storage 300 is an affordable, scalable and high-performance solution if you require failover capability to support mission-critical applications. Its dual engines and overall system design create increased levels of performance compared with the IBM NAS 200 in either Windows or mixed Windows/UNIX environments. By scaling from 360 GB to 3.24 TB, the IBM NAS 300 also provides the storage capacity necessary to support mission-critical applications. Like the IBM NAS 200 series, its support of Windows (CIFS), UNIX (NFS), HTTP, FTP and Novell (NetWare) lets clients share files using different protocols, and eliminates the need to have separate servers for each supported protocol.

In addition to the failover protection afforded by its dual-engine design, the IBM NAS 300 also provides redundant, hot-swap power supplies, hot-spare, hot-swap hard disk drives and dual RAID controllers for increased system availability. In addition to those Persistent Storage Manager features provided on the IBM NAS 200 series, the IBM NAS 300 has the additional capability to instantly restore volumes of data in a minimum of steps. This quick-restoral ability eliminates the need to point, click, drag and drop potentially thousands of files, saving valuable time for scarce IT resources. The same tightly integrated suite of preloaded software that is found on the IBM NAS 200 is standard on the IBM NAS 300 as well.

IBM NAS 300

The two engines of the IBM NAS 300 each use dual 933-MHz Pentium III processors and 1 GB of system memory to bolster I/O throughput. The use of Fibre Channel technology in its hubs and hard disk drives further increases performance. The IBM NAS 300 is upgradable to a total engine memory size of 2 GB.

Environments and Applications

The IBM NAS 300 is well suited for use in larger departments or for enterprise-wide support. Because of its dual-engine failover capability, the IBM NAS 300 is ideally suited to mission-critical applications such as:

- Accounts receivable processing and storage
- Payroll support
- Customer service support

For further information regarding the IBM Network Attached Storage 200 and 300, as well as other IBM storage networking products, go to **ibm.com**/storage/nas ■

About the Author

Eric Dunlap is Product Marketing Manager for Integrated NAS products and is based in Research Triangle Park, North Carolina. Eric has a diverse IT background, bringing several years of experience to the Storage Networking Division.
sedunlap@us.ibm.com

Taking SAN Management to the Next Level with Tivoli Storage Network Manager

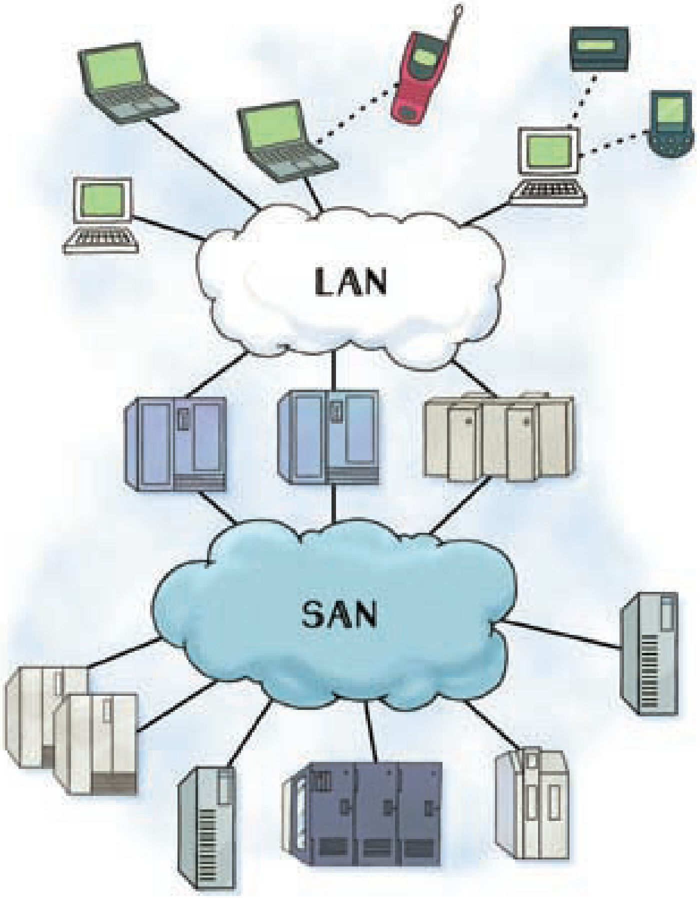
By Sarah Jeong and Steve Luko

Storage area networks (SANs) have emerged as the hottest technology to hit the storage industry this decade. Why all the hype? Because SANs offer real promise.

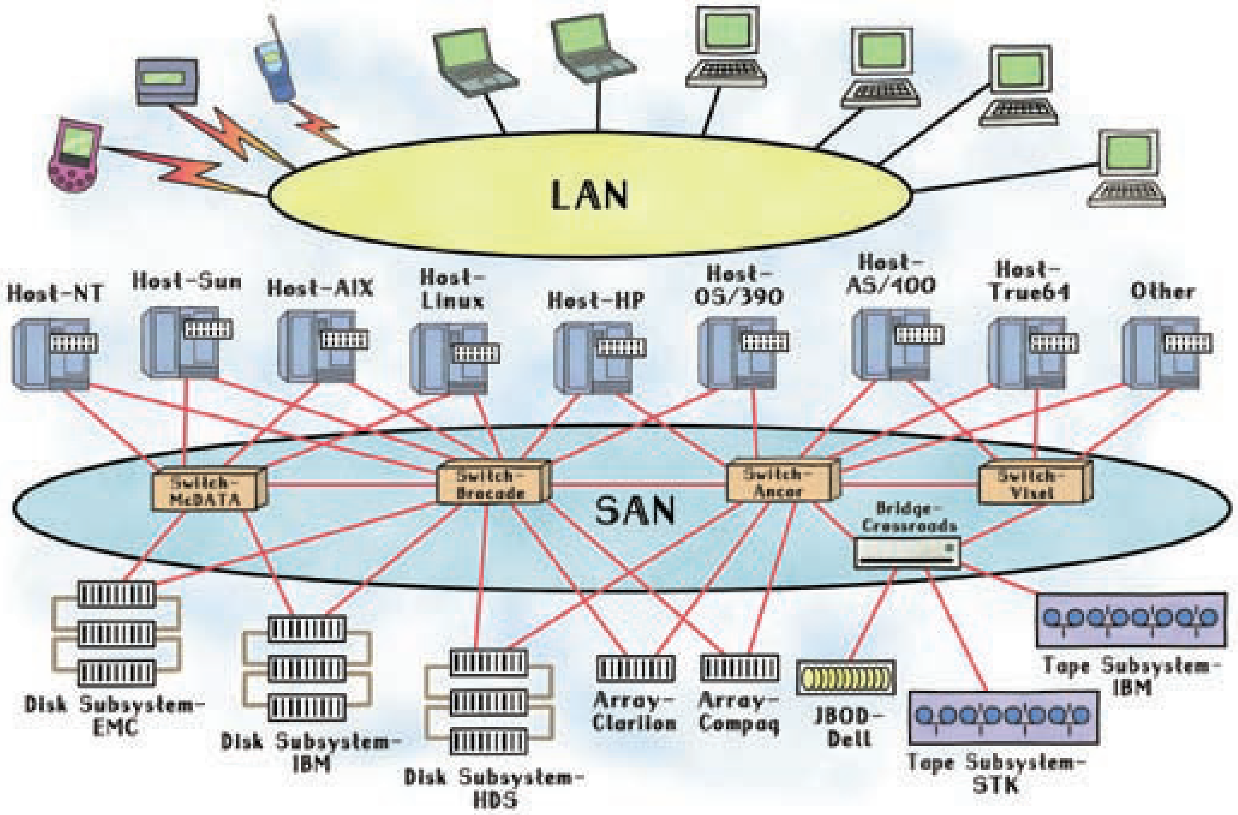
SANs offer the promise of reduced costs, improved speeds, centralized management, consolidation, flexibility, scalability, resource sharing and simplified administration, all the necessary ingredients to make organizations more competitive. Industry analysts estimate that most storage will be networked by 2005, and that SANs will be a key technology for the support of e-business.

Although SANs may be the future of enterprise storage management, they come with some very specific problems. The reality of SANs is that they can be extremely complex. When you look at an enterprise SAN, you will most likely see hosts from one vendor, storage devices from another, non-similar file systems and catalogs, different access methods, and so on. Such network complexity, coupled with massive amounts of data, make it a challenge to maintain your existing business processes and to make strategic decisions to enter new markets to expand your business. Complexity is especially apparent when it comes to managing this environment.

Tivoli® Storage Network Manager is a comprehensive solution for managing SAN infrastructures and their associated storage resources. This unique solution provides features to manage SAN topology, assign available disk resources to managed hosts, and automatically extend file systems using administrator-defined policies.



The costs of managing this type of environment are extensive. It is estimated that over the life of the asset the cost of managing storage is three to eight times the cost of the actual storage itself. In addition, there is a shortage of storage administrators, with no relief in sight. Industry analysts predict that by the year 2003 there will be 1.5 million open, unfilled IT positions. With this said, SAN management has become a very popular topic. In September 2000, one of the industry's leading analysts said that SAN management is more important than the network itself.



Getting past the intimidation of so many components and connections involved with networked storage, we realize that there are three key requirements for managing a storage network:

1. Integrate new SAN and storage technologies quickly and easily into enterprise infrastructures to alleviate the storage constraints.
2. Simplify the management of existing and new technologies, and reduce the overall cost of managing the environment.
3. Provide quick, reliable and secure access to data for application processing.

Simplifying the Complexities of SAN Management

Tivoli Systems has a solution for simplifying the complexities of managing heterogeneous enterprise SANs. Tivoli Storage Network Manager provides a complete solution by moving beyond the traditional industry definition of SAN management, which is: attempting to manage the SAN through the proprietary management software of Fibre Channel switch vendors. Although switch management is necessary, Tivoli Storage Network Manager enables administrators to take SAN management to the next level by taking advantage of all the management intelligence and interoperability built into the ANSI standards-based Fibre Channel protocol.

Tivoli Storage Network Manager is a comprehensive solution for managing SAN infrastructures and their associated storage resources. This unique solution provides features to manage SAN topology, assign available disk resources to managed hosts, and automatically extend file systems using administrator-defined policies. This solution is built upon an architecture that can easily scale to handle very large and complex configurations. Tivoli Storage Network Manager provides:

- SAN discovery and monitoring for problem identification and resolution. This assists in the maintainability of the SAN infrastructure to ensure continuous application availability. The console used for SAN discovery can integrate directly with the Tivoli NetView® program or operate in standalone mode to provide vendor-independent graphical SAN management, allowing administrators to monitor and quickly resolve problems within the SAN infrastructure.

- Identification and assignment of heterogeneous disk storage resources by logical unit number (LUN) to provide a simple, secure, and efficient method of assigning resources to host systems. This software-based LUN masking/disk management function allows you to securely manage disk resources attached to the SAN. These resources can be assigned to hosts in the SAN using a graphical display to show the assignments in effect and the disk storage resources available. Policy-based control provides flexibility and productivity by allowing granular control at the individual file-system level or host level or by grouping hosts together and configuring them as a set. Policy-based administration reduces the effort required to configure the SAN by providing for inheritance or the use of default policies.
- Monitoring of file systems to automatically assign and extend additional disk resources to maintain continuous application processing while reducing administrative workload and costs. This file system automation function uses administrator-defined policies to automatically monitor file systems, assign additional storage, and extend file systems in an out-of-space condition. Policies can be explicitly set for a group of hosts, an individual host, or a file system on a particular host.

Complete SAN Topology Discovery using Inband and Outband Methods

In order to manage a SAN, it is first necessary to identify the network topology — what is physically connected to what. Tivoli Storage Network Manager performs SAN topology discovery and display of the components and storage resources across the SAN based on the ANSI FC-MI, the industry standard for Fibre Channel management. This discovery uses a combination of discovery mechanisms: inband, through the SAN network itself, and outband, through TCP/IP using Simple Network Management Protocol (SNMP) capabilities for the most accurate results.

Tivoli Storage Network Manager uses multiple, industry-standard discovery techniques. These include, but are not limited to, the SNMP-based Fibre Alliance Fibre Channel (FC) MIB, "legacy" SCSI commands and the newest management service of native inband Fibre Channel commands. These commands include extended link services such as request node identification (RNID) and request topology information (RTIN), name server queries, and management server queries, as well as selected vendor-specific interfaces. The use of multiple discovery techniques and a correlation engine allows Tivoli Storage Network Manager to identify devices discovered by two or more discovery mechanisms to provide an accurate topology.

Although both inband and outband methods have a valid role in SAN management and are complementary, the inband method provides superior function. The current functional advantage of inband over outband is expected to increase due to the greater ongoing investment in inband technology.

The outband Fibre Channel MIB was developed originally as a de facto standard by the Fibre Alliance vendor consortium. The rationale was to provide basic SAN manageability quickly and with broad device coverage using the well-established and easy-to-implement SNMP protocol. The Fibre Alliance FC MIB corrects the basic flaws in SNMP, including the lack of topology awareness and conflicting data definitions. The FC MIB has made SNMP a useful tool for basic SAN management during the early phase of SANs.

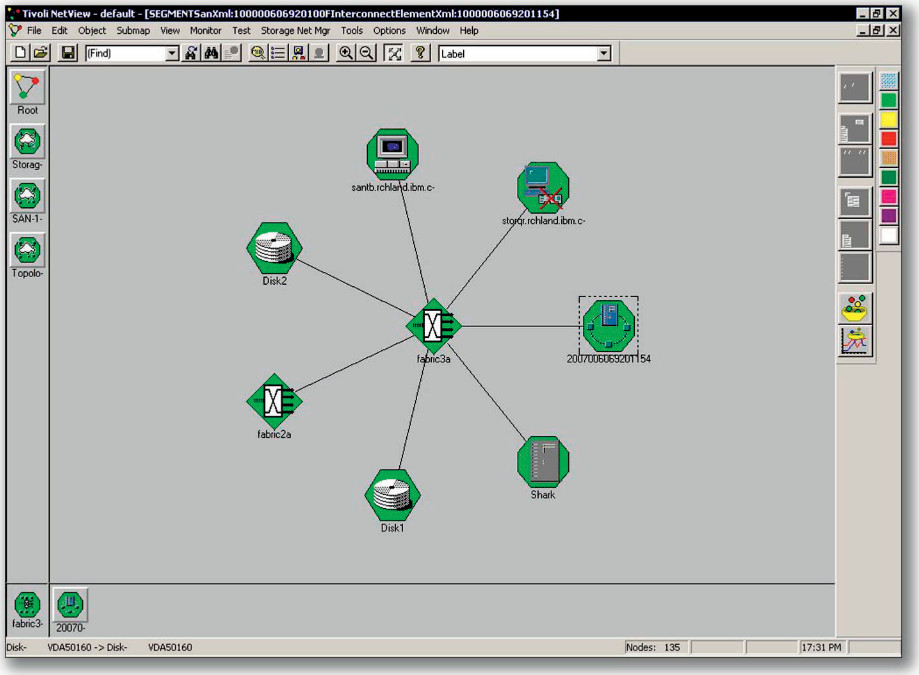
The requirement for an SNMP agent on a SAN-attached device has both an advantage and disadvantages. The advantage of the outband agent is that it provides an alternative method to detect and forward errors where the inband method fails due to the outage of the primary SAN data path. The disadvantages are that the agent must be installed, consumes memory and CPU resource, requires a LAN adapter and connection, and adds an additional point of failure. If your mission-critical servers are already resource-constrained, these can be weighty costs. Storage service providers are even more sensitive to these concerns because they often must ask their customer to bear these risks where the customer retains primary operational responsibility for the system. An additional disadvantage of the outband approach is that it lacks the visibility and insight into the status of adjoining SAN devices.

The SAN industry set out to build a network with new levels of performance, reliability, manageability and interoperability. The SAN architects benefited from industry experience gained in both wide and local area networking — a much more powerful and less costly technology — and the freedom to start from a blank slate. A current advantage of inband over outband is the ability of inband-compliant devices to discover and provide error reporting for adjoining devices.

For example, a switch can use inband facilities to discover and manage the physical and logical connections to a fibre-attached disk device. Similarly, inband management is used by a switch to discover and manage fibre-attached CPUs through contact with their Host Bus Adapters (HBAs). This eliminates the need for agents where only basic SAN fabric management is required, increases the number of systems manageable in a SAN management domain, and decreases the cost of deploying the management system.

A future version of the inband SNIA HBA standard is expected to include the ability of the HBA to query the host system and provide the host name, IP address, and OS type and level, and to identify multiple HBAs on a single host. Multiple HBAs on host systems are expected to be the most common implementations, and the ability to identify these will enhance error detection and fault isolation.

Another future capability expected under the forthcoming inband GS-4 standard for switches is a common zone control mechanism that would allow the setting of zones across multiple switch vendors. These capabilities will extend inband's ability to deliver lower management costs through improved administrator productivity and improved application service levels.



Monitoring the SAN Infrastructure

With Tivoli Storage Network Manager, you can now continuously monitor the components within the discovered SAN topology and capture data to be used for reporting. Tivoli Storage Network Manager collects and consolidates information about events as well as provides for multi-vendor, multi-component fault isolation and problem determination in the SAN. When the SAN exceeds tolerances or experiences a failure, alerts are generated and icon colors are changed on the Tivoli Storage Network Manager Console.

Using the icons on the console, you can drill down into individual elements to view configuration information as well as alert and failure information. If a device must be manipulated or reconfigured, you can launch specific "vendor-provided" management tools from within Tivoli Storage Network Manager to assist in resolution of the problem. In addition, Tivoli Storage Network Manager can integrate directly with the Tivoli NetView program, allowing you to monitor and control the SAN infrastructure from the same interface that you use to manage your LANs and WANs. These networks can now be viewed from the same console.

One feature of a SAN is the ability to support logical subsets or zones in the network. Said another way, just because there is a physical connection between the host, SAN fabric and storage, does not mean that there is a logical connection. With Tivoli Storage Network Manager, once a physical map is produced, additional information allows logical views to be overlaid showing zones, host-to-storage and storage-to-host mappings. Thus, the SAN topology can be displayed in both physical and logical views.

Assigning Storage Resources to Host Systems

In a SAN, where you have a lot of hosts attached to a lot of disks, something has to govern which disks are accessible by which hosts. This is what will ensure that a Windows NT®, AIX®, or Sun machine has its own set of disks and that they don't step on each other. Windows NT and Windows® 2000, as part of their Plug and Play architecture, are incompatible with other operating systems on a SAN in that they will attempt to write signatures on all the disks they can see, an action that makes it impossible for other systems to read from those disks. Because you are potentially in an any-to-any networked configuration, without such control mechanisms there would be chaos in the SAN. Tivoli Storage Network Manager provides this control with its software-based LUN masking capabilities.

It is important to note that LUN masking is also available from the high-end disk subsystem vendors like EMC and HDS, in addition to IBM. If you are an all-EMC shop, you can do it using the EMC Symmetrix. If you are an all-IBM shop, it can be done inside the IBM Enterprise Storage Server™ (ESS). Rather than depending on the software LUN masking to be installed on the managed system, intelligent disk subsystems have their control point at the disk controller. Because intelligent disk subsystems perform LUN masking at the controller level, there is no exposure to a "rogue" host. A rogue host is an unauthorized or improperly configured system on the SAN. Although Tivoli Storage Network Manager can alert the operator when a rogue host joins the SAN, it cannot prevent the rogue host from improperly accessing data.

Although intelligent disk subsystem controller-based LUN masking offers protection from rogue hosts, it has two potential drawbacks. The first is cost. There is a large investment in "Just A Bunch of Disks" (JBOD) that can be economically and safely adapted to the SAN environment using proper precautions. Software-based LUN masking used in conjunction with best-practice data-center physical security, change control and data protection, and supplemented with port-based zoning, can create a safe habitat for JBODs on the SAN.

The second drawback to the use of intelligent subsystem LUN masking exists where subsystems are from multiple vendors and thus have multiple different configuration interfaces adding to the complexity of this area of SAN management. Tivoli Storage Network Manager reduces the chance of configuration error in the setup and maintenance of LUN masking by presenting the storage administrator with one interface for diverse systems and storage devices. You can easily and securely assign discovered disk resources (at the LUN level) from the heterogeneous storage subsystems attached to a SAN to specific computers connect to the SAN, ensuring that the right host is looking at the right disk. Also, it is possible to allow multiple systems to have access to the same LUN. Tivoli Storage Network Manager effectively allows multiple computers to securely share the same SAN resources and the same storage subsystems.

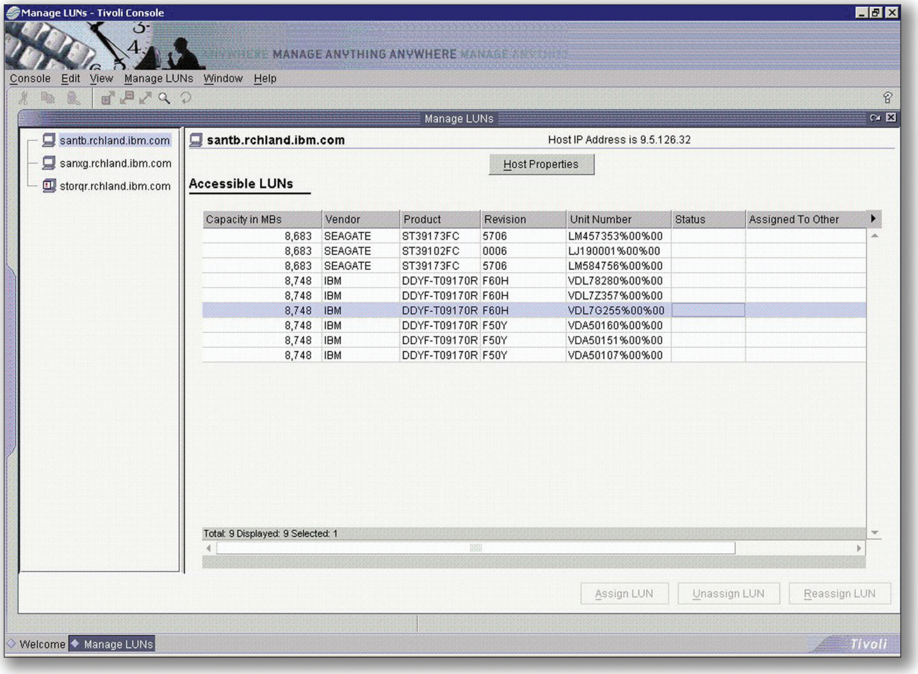
You can easily view all the assigned LUNs for a particular host using the Tivoli Storage Network Manager GUI. You can also view available LUNs from a heterogeneous list compiled by Tivoli Storage Network Manager, and assign them as resources are needed. To assign a LUN, you simply select it and apply the change. The additional resources are immediately available to the host. You can just as easily unassign a LUN from one host and reallocate it to another.

Another way to ensure that the right systems have access to the right disk storage is port-level zoning from the switch. Port-level zoning uses zones that allow devices plugged into specific ports to pass data between them. The disadvantage of using only port-level zoning to govern system-to-disk access is that it does not allow for the sharing of disk resources at the granular level of the LUN. Such granularity allows for more efficient use of storage assets and is achievable only by using LUN masking from intelligent disk subsystems or software-based LUN masking, such as the one available from Tivoli Storage Network Manager.

Tivoli Storage Network Manager's policy-based file system automation ensures continuous access to mission-critical applications.

Automating File System Extension

Tivoli Storage Network Manager's policy-based file system automation ensures continuous access to mission-critical applications. For example, on a Web site that is up 24x7, you cannot have an outage due to an out-of-space condition. Tivoli Storage Network Manager enables clients to automatically assign more space when a predetermined threshold is reached. This enables a client to stipulate, "I don't want the location where content is stored to be over 80% full. If it does, take action to fix it so that I don't run out of space in my Web environment."



santb.rchland.ibm.com Host IP Address is 9.5.126.32

Accessible LUNs

Capacity in MBs	Vendor	Product	Revision	Unit Number	Status	Assigned To Other
8,683	SEAGATE	ST39173FC	5706	LM457353%00%00		
8,683	SEAGATE	ST39102FC	0006	LJ190001%00%00		
8,683	SEAGATE	ST39173FC	5706	LM584756%00%00		
8,748	IBM	DDYF-T09170R	F60H	VDL78280%00%00		
8,748	IBM	DDYF-T09170R	F60H	VDL7Z357%00%00		
8,748	IBM	DDYF-T09170R	F60H	VDL7G255%00%00		
8,748	IBM	DDYF-T09170R	F50Y	VDA50160%00%00		
8,748	IBM	DDYF-T09170R	F50Y	VDA50151%00%00		
8,748	IBM	DDYF-T09170R	F50Y	VDA50107%00%00		

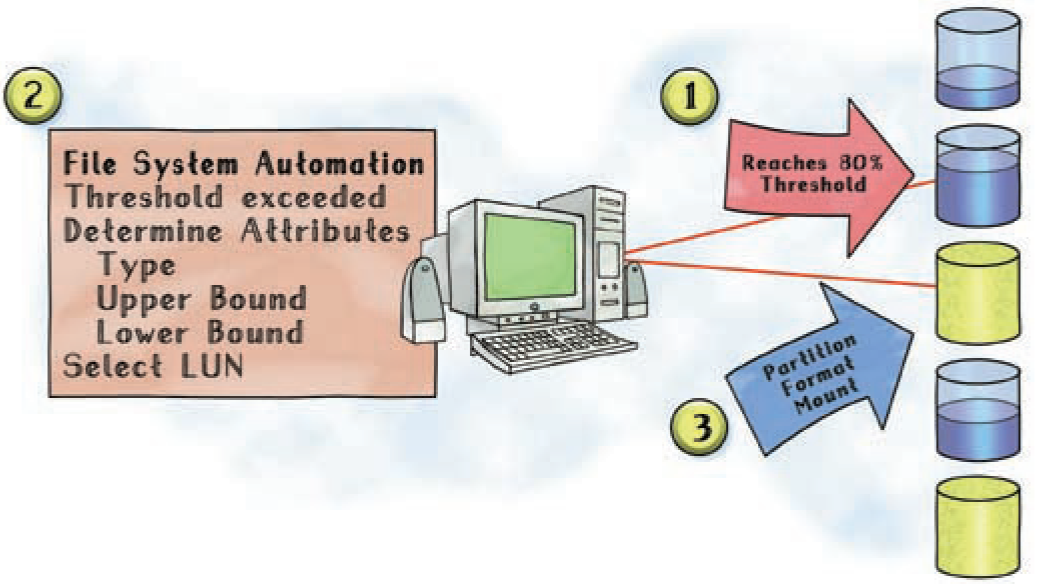
Total: 9 Displayed: 9 Selected: 1

Tivoli Storage Network Manager continuously monitors these resources as they approach a policy-defined threshold or capacity level. When the threshold is exceeded, Tivoli Storage Network Manager automatically extends those file systems. This task varies depending on the operating system. For example for Sun Solaris, Tivoli Storage Network Manager:

1. Writes a label to the disk so that it can be used by the Veritas Volume Manager
2. Configures the disk for use with the Veritas Volume Manager
3. Adds the disk to the group disk where the file system's logical volume resides
4. Increases the size of the file system and underlying system's logical volume by adding all the available space from the assigned LUNs

Tivoli Storage Network Manager intelligently selects the most appropriate available LUN based on the size required, or other characteristics. During initial configuration, the upper and lower bounds of the size of an acceptable LUN are specified. Later, when an additional LUN is required, the largest available LUN that meets the acceptable size criteria will always be chosen. Also, if the system is using software-based mirroring, Tivoli Storage Manager will automatically select and assign multiple LUNs as required to preserve the mirroring.

This unique automation capability can greatly reduce administrative workload and ensure continuous application availability. There may be cases where automatic extension is not required, such as when users don't want to pay for additional storage resources and would prefer to be notified so that they can manually delete files that are not longer needed. Tivoli Storage Network Manager can meet this requirement by sending an alert indicating that a warning threshold has been reached, and the user can be notified. Whether a system is either mon-



itored or automatically extended, Tivoli Storage Network Manager sends SNMP and Tivoli Enterprise Console events to report these activities to the designated administrator's or management console. This unique monitoring and extending feature is available to both internal and external disks.

Integration with Complementary Tivoli Offerings

Tivoli Storage Network Manager is a key component of the overall Tivoli Storage Management Solutions strategy for comprehensive SAN and storage resource management. An application that gives an overall storage management view of the enterprise for the operational integrity of the business, Tivoli Storage Network Manager can operate and integrate with:

- Tivoli NetView
- Tivoli Enterprise Console
- Tivoli Decision Support for SAN resource and management analysis

Supported Platforms

Console Platform
Microsoft Windows 2000 Professional Edition, Server Edition or Advanced Server Edition

Managing Server Platform
Windows 2000, Advanced Server Edition

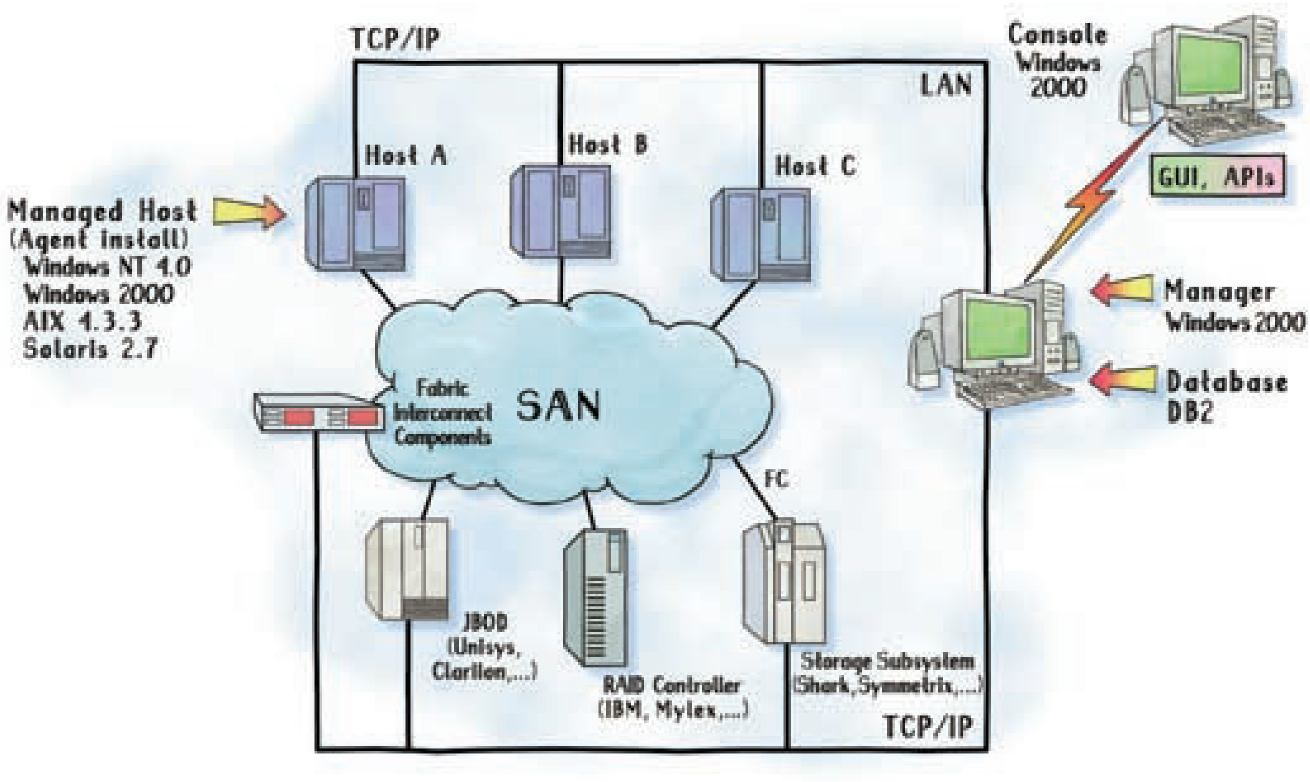
Managed Host Platform
Microsoft Windows NT Version 4.0 Server; Windows 2000, Professional Edition, Server Edition, or Advanced Server Edition for all functions including file system monitoring, except automatic file system extension; IBM AIX, Version 4.3.3; Sun Solaris, Version 2.7 for all functions.

For more information about current platform support, device compatibility, application support, and the latest information on features and developments, go to tivoli.com/tsnm ■

About the Authors

Sarah Jeong is currently the Product Marketing Manager for Tivoli Storage Network Manager. Sarah has an extensive background in the open-systems arena and storage. and has worked in nearly every facet of the product development life cycle, including development, technical support, systems analysis, training, implementation, product management and product marketing. She holds a BS in Computer Science from California Polytechnic State University, San Luis Obispo.
sjeong@tivoli.com

Steve Luko is currently the Product Manager for Tivoli Storage Network Manager. Prior assignments in his 19 years with IBM included architect with IBM Business Recovery Services designing advanced recovery solutions for open systems, storage software product marketing, network consulting, and pre-sales technical sales support for networks and systems management. He holds an MBA from University of California, Irvine.
sluko@tivoli.com



How Do I Learn More About SAN and NAS?

by Mary Lovelace

IBM Learning Services offers courses on a wide variety of storage and Storage Networking topics to help you increase your skills and stay competitively sharp.

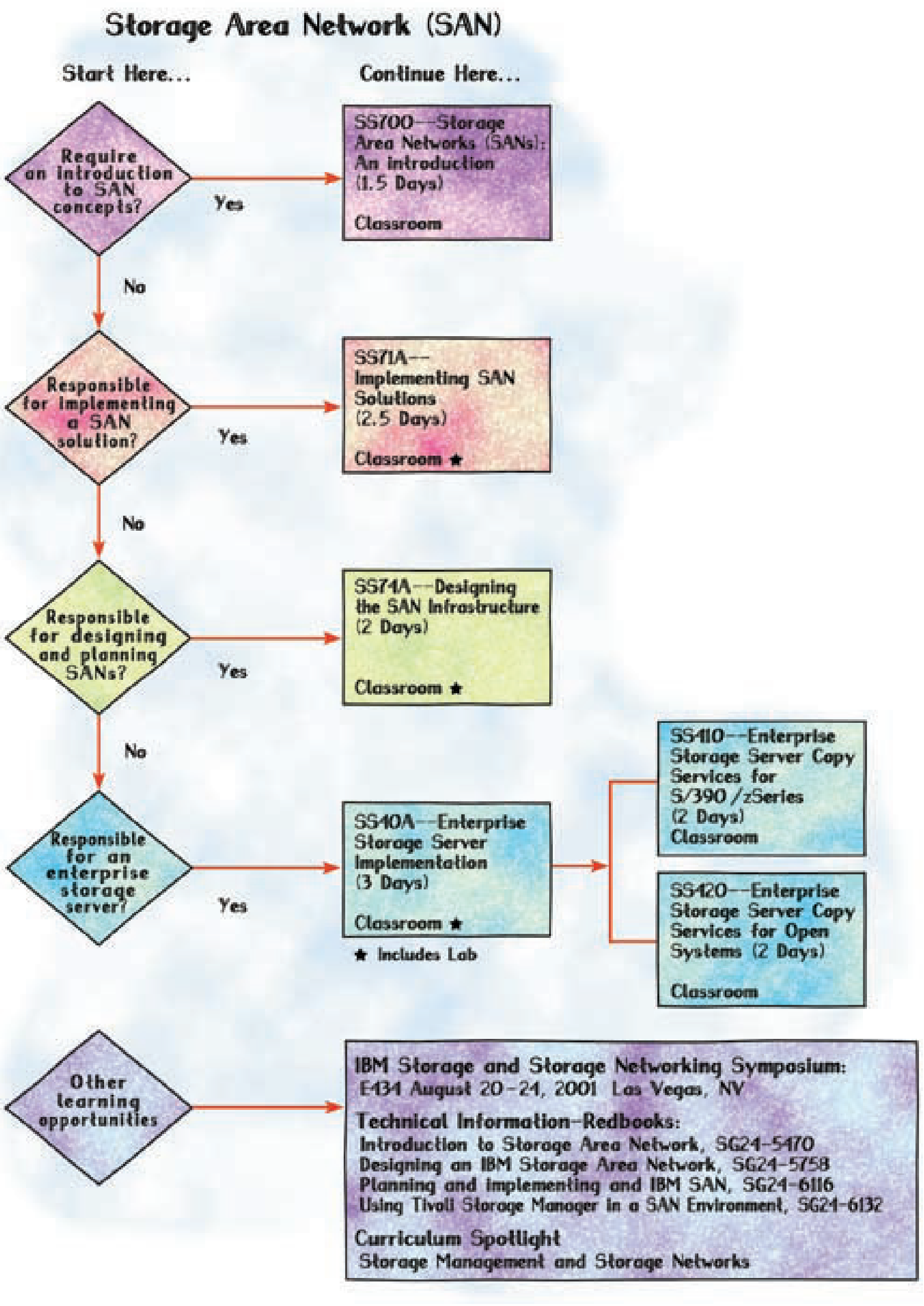
Storage Area Network Courses

Storage area networks (SANs) are the foundation of a storage network. They harness the breadth of IBM technical expertise, products, software, and services. They manage the explosive growth of data, offer high application availability, and make it easier to manage resources and growth.

(SS700) Storage Area Networks (SANs): An Introduction (1.5 days, USD595)

Examine products and strategies associated with managing the explosive growth of business data across the enterprise in today's networking economy. Learn the basic concepts and terminology associated with SANs, and map the promise of SANs to the complications of managing islands of information among heterogeneous environments with disparate operating systems, data formats, user interfaces, and limited integration of products from assorted vendors. Learn how to enable your enterprise to take advantage of this relatively new approach to information management.

To enroll, call 1 800 IBM-TEACH (1 800 426-8322) or visit ibm.com/training/spotlight/storage for more information and a complete list of storage and Storage Networking classes.


Storage Area Network (SAN)
Start Here...
Continue Here...
Require an introduction to SAN concepts?
Yes
SS700—Storage Area Networks (SANs): An introduction (1.5 Days)
Classroom
No
Responsible for implementing a SAN solution?
Yes
SS71A—Implementing SAN Solutions (2.5 Days)
Classroom ★
No
Responsible for designing and planning SANs?
Yes
SS74A—Designing the SAN Infrastructure (2 Days)
Classroom ★
No
Responsible for an enterprise storage server?
Yes
SS40A—Enterprise Storage Server Implementation (3 Days)
Classroom ★
★ Includes Lab
SS410—Enterprise Storage Server Copy Services for S/390 /zSeries (2 Days)
Classroom
SS420—Enterprise Storage Server Copy Services for Open Systems (2 Days)
Classroom
Other learning opportunities
IBM Storage and Storage Networking Symposium:
E434 August 20–24, 2001 Las Vegas, NV
Technical Information–Redbooks:
Introduction to Storage Area Network, SG24-5470
Designing an IBM Storage Area Network, SG24-5758
Planning and Implementing and IBM SAN, SG24-6116
Using Tivoli Storage Manager in a SAN Environment, SG24-6132
Curriculum Spotlight
Storage Management and Storage Networks

(SS71A) Implementing SAN Solutions (2.5 days, USD2495)

This intermediate-level, hands-on course provides the in-depth working knowledge and skills needed to implement Fibre Channel SAN solutions. This course builds on your knowledge of Fibre Channel SAN basics, adding technical discussions of protocols, SAN connectivity and security considerations, and specific SAN product training. Lecture and lab activities focus on how to plan, design, install, and configure SANs in Windows NT® and UNIX® environments. The skills you develop can also be applied to other environments. Lab activities include more than 8 hours of hands-on configurations and demonstrations with a variety of Fibre Channel products and configurations.

(SS74A) Designing the SAN Infrastructure (2 days, USD2295)

This intermediate-level course provides the knowledge and skills that senior consultants require in order to plan and design SANs for customer requirements. The course emphasizes the critical considerations required for creating effective Fibre Channel SAN logical and physical designs. You will practice design steps for four solution-centric projects from start to finish. Based on extensive SAN consulting engagements, this course offers practical advice and tested options that you can apply immediately to your next SAN project.

Enterprise Storage Server (ESS) Courses

The IBM Enterprise Storage Server™ (ESS) is the ultimate SAN utility, providing the information “fuel” that runs the e-business “engine.” IBM Learning Services has courses to help you use the ESS to address any or all of your strategic and tactical business initiatives and give your organization the business advantage needed to survive and thrive in the e-world.

(SS40A) Enterprise Storage Server Implementation (3 Days, USD1930)

This intermediate-level, hands-on course is intended for storage administrators, system programmers, and support-group members responsible for storage subsystems in open-systems and mainframe environments. It provides information about the IBM 2105 Enterprise Storage Server (ESS) and its supporting software products, including StorWatch™ ESS Specialist and StorWatch ESS Expert. Hands-on labs using ESS Specialist to configure the ESS subsystem provide training in each of the products. You will learn to identify the components and features of ESS, host attachments, and attachment considerations of ESS for open-systems platforms. You will also learn about the logical configuration of ESS, how to access logical volumes defined for ESS, the copy services function of Concurrent Copy, FlashCopy™, Peer-to-Peer Remote Copy (PPRC), and Extended Remote Copy (XRC), as well as the setup and functions of ESS Expert.

New Copy Services Courses

Two new courses on Copy Services, one for zSeries and S/390® and one for open systems, deliver the skills needed to provide the reliability, scalability and availability that have come to be expected in the mainframe world, along with the openness and interoperability of a solution that will work across a mixed computing environment. You can take these courses at one of the IBM Learning Services training locations or, if you like, they can be delivered at your location. IBM Learning Services will also work with you to understand your training challenges and to tailor a training solution that is a perfect match for your audience.

(SS410) Enterprise Storage Server Copy Services for S/390 / zSeries (2 Days, USD1465)

Get a detailed look at the ESS Copy Services functions for S/390 and zSeries environments, and the procedures required to implement these functions. Perform an Extended Remote Copy (XRC) volume copy, a FlashCopy, and a Concurrent Copy, and set up a Peer-to-Peer Remote Copy (PPRC) volume copy within a configured ESS. Study the methods for performing these functions with job control language (JCL) and Time Sharing Option (TSO) commands, and using the ESS Web Copy Services interface.

(SS420) Enterprise Storage Server Copy Services for Open Systems (2 Days, USD1465)

Get detailed information on the Enterprise Storage Server (ESS) Web Copy Services functions and the procedure required to implement these functions in an open systems environment. Perform a FlashCopy and set up a Peer-to-Peer Remote Copy (PPRC) volume copy within a configured ESS. Study the methods for performing these functions using the ESS Web Copy Services and command line interfaces (CLIs). ■

About the Author

Mary Lovelace is the SAN Education Focal Point for IBM Learning Services. Before joining ILS she worked with the International Technical Support Organization, San Jose Center, authoring Enterprise Storage Redbooks. Mary has more than 20 years of experience with IBM in large-systems and storage-product education, system engineering and consultancy, marketing support, and system programming.
mhl@us.ibm.com

User Group Meetings/Events

July	2001	
Int'l Workshop on Intelligent Data Acquisition & Advanced Computing Systems	July 1 - 4	Foros, Ukraine
Int'l Conf. on Dependable Systems & Networks	July 1 - 4	Göteborg, Sweden
6th IEEE Symp. on Computers & Comm	July 3 - 5	Hammamet, Tunisia
Int'l Conf. on Intelligent Agents, Web Technologies, & Internet Commerce	July 4 - 6	Las Vegas, NV
Computing Systems	July 5 - 7	Kuala Lumpur
International Conference on Enterprise Information Systems	July 7 - 10	Setúbal, Portugal
Share/Guide Association Conference	July 9 - 10	Singapore
COMMON Malaysia Conference	July 9 - 10	Kuala Lumpur
Internet World Chicago	July 9 - 12	Chicago, IL
Int'l Conf. on Networking	July 9 - 13	Colmar, France
Interaction 2001 Conference	July 15 - 17	Sydney, Australia
Wireless Developers Seminar	July 17	Atlanta, GA
e-business Analytics Seminar	July 17	Atlanta, GA
Perl Conference 5	July 23 - 27	San Diego, CA
SHARE Conference	July 22 - 27	Minneapolis, MN
Wireless Developers Seminar	July 31	Boston, MA
e-business Analytics Seminar	July 25	New York, NY
Migrating Enterprises to Linux Seminar	July 24	Reston, VA.
Migrating Enterprises to Linux Seminar	July 25	New York, NY
7th World Conference on Computers in Education	July 29 - Aug. 3	Copenhagen, Denmark

August

10th Int'l Symp. on High-Performance Distributed Computing	Aug. 7 - 10	San Francisco, CA
14th Int'l Conf. on Parallel & Distributed Computing Systems	Aug. 8 - 10	Dallas, TX
IBM Storage and Storage Networking Symposium	Aug. 20 - 24	Las Vegas, NV
10th USENIX Security Symposium	Aug. 13 - 17	Washington, DC
Solutions, The IBM Technical Developer Conference	Aug. 13 - 16	San Francisco, CA
2nd Int'l Conf. on Software Eng., AI, Networking, & Parallel/Distributed Computing	Aug. 20 - 22	Nagoya
13th Int'l Conf. on Parallel & Distributed Computing & Systems	Aug. 21 - 24	Anaheim, CA
IEEE Int'l Symp. on Network Computing & Applications	Aug. 22 - 24	Boston, MA
SecureWorld: IBM's End-To-End Security Conference	Aug. 27 - 31	Washington, DC
European Conf. on Parallel Computing	Aug. 28 - 31	Manchester, UK

September

6th Int'l Conf. on Parallel Computing Technologies	Sept. 3 - 7	Novosibirsk, Russia
5th Int'l Enterprise Distributed Object Computing Conference	Sept. 4 - 7	Seattle, WA
Wireless Developers Seminar	Sept. 5	Santa Clara, CA
Migrating Enterprises to Linux Seminar	Sept. 6	Santa Clara, CA
IBM @server iSeries (AS/400) Technical Conference	Sept. 6	Orlando, FL
Networking Solutions Technical Conference	Sep. 10 - 14	Las Vegas, NV
COMMON Latinoamerica Peru Conference	Sep. 12 - 14	Lima, Peru
Storage World Conference 2001	Sept. 14	San Jose, CA
Internet World	Sept. 26 - 27	Glasgow

October

ACM Symp. on Mobile Ad Hoc Networking & Computing	Oct. 4 - 5	Long Beach, CA
IEEE Int'l Symp. on Network Computing & Applications	Oct. 8 - 10	Cambridge, MA
IEEE 3rd Int'l Conf. on Cluster Computing	Oct. 8 - 11	Newport Beach, CA
e-Power Technical Conference Windows 2000 Enterprise Solutions	Oct. 8 - 12	Washington, D.C.
IBM @server pSeries (RS/6000) Linux and NUMA-Q Technical University	Oct. 8 - 12	Atlanta, GA
z/OS and OS/390 Expo	Oct. 8 - 12	Orlando, FL
ISPCON	Oct. 9 - 11	Las Vegas, NV
Gartner Symposium and ITxpo	Oct. 9 - 11	Lake Buena Vista, FL
9th Int'l Conf. on Networks	Oct. 10 - 12	Bangkok
30th IEEE Applied Imagery Pattern Recognition Workshop	Oct. 10 - 12	Washington, D.C.
iSuc Conference	Oct.17 - 19	Osaka, Japan
COMMON Conference and Expo	Oct. 21 - 25	Minneapolis, MN
Storage Networking World	Oct. 22 - 24	Orlando, FL
IMS Technical Conference	Oct. 22 - 25	Miami Beach, FL
Networking Solutions Technical Conference,	Oct. 22 - 26	Malta
ISPCON Europe	Oct. 23 - 25	London, UK
Storage Tank Management and Technology Conference	Oct. 23 - 26	Phoenix, AZ
Wireless e-business Conference and EXPO	Oct. 28 - 31	Orlando, FL

November

20th IEEE Symposium on Reliable Distributed Systems	Nov. 4 - 7	New Orleans, LA
5th Annual Linux Showcase and Conference	Nov. 6 - 10	Oakland, CA
XFree86 Technical Conference	Nov. 7 - 8	Oakland, CA
9th Int'l Conf. on Network Protocols	Nov. 11 - 14	Riverside, CA
COMDEX 2001	Nov. 13 - 17	Las Vegas, NV
Internet World	Nov. 14 - 15	Manchester, UK
GLOBECOM 2001	Nov. 25 - 29	San Antonio, TX
1st Int'l IFIP TC11/WG11.4 Conf. on Network Security	Nov. 26 - 27	Leuven, Belgium
ISPCON Germany	Nov. 26 - 28	Frankfurt, Germany
ISPCON France	Nov. 27 - 29	Paris, France
IEEE Int'l Conf. on Data Mining	Nov. 29 - Dec. 2	San Jose, CA

December

15th Systems Administration Conference (LISA 2001)	Dec. 2 - 7	San Diego, CA
Int'l Conf. on High-Performance Computing	Dec. 17 - 20	Hyderabad, India
Int'l Symp. on Artificial Intelligence	Dec. 18 - 20	Kolhapur, India

Events in 2002

Conference on File and Storage Technologies (FAST)	Jan. 28 - 29	Monterey, CA
BSDCon	Feb. 11 - 14	San Francisco, CA
Share/Guide Association Conference	Mar. 3 - 8	Nashville, TN
The Systems and Network Administration Conference	Mar. 17 - 21	Dallas, TX

Network Technology

A Net390 project helps you to make the appropriate strategy, infrastructure, product, design, system management, and implementation decisions required to leverage S/390 and @server z900 enterprise servers into the e-business infrastructure.

Leveraging SNA while Transforming to e-business

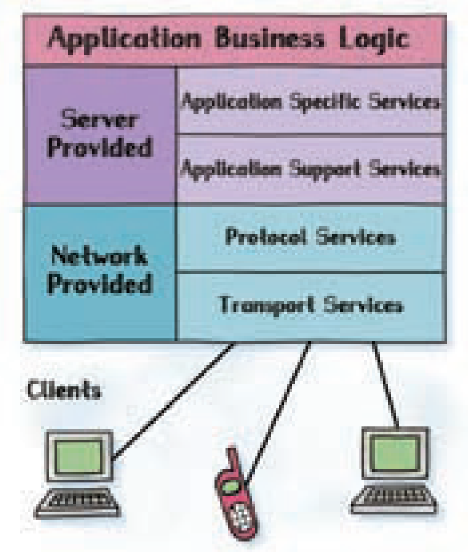
By Robert Brinkman, Jim Goethals, and Bob Louden

SNA powers a majority of the financial transactions that traverse the Web, and most of the airline and hotel reservations booked over the Web rely on SNA. This article explores ways to preserve and even enhance SNA investments while enabling for e-business. At the IBM Global Services (IGS) Center for Infrastructure Solutions in Research Triangle Park, NC, we have worked with many forward-thinking organizations that recognize the value in their SNA investments and want to continue leveraging this value. These organizations are transitioning their SNA environment to preserve their significant investment in SNA applications, endpoints, and processes while positioning themselves for TCP/IP in the future.

Through our work with many clients, we have identified four infrastructure architectures for S/390® and z900 environments. These architectures can serve as marker posts to where an organization is today and where it may want to be tomorrow. They help to guide investments while minimizing significant and costly changes, avoiding potential traps while leveraging past and future investments, and enabling new capabilities while continuing to operate today. Before discussing these four architectures, we will first explore a service delivery model, which is the strategy that many organizations adopt to remain competitive. And lastly, we will discuss the Net390 architecture; Net390 is the data-center structure that will optimally support the four architectural models. Both Net390 and service delivery are best-practices approaches to networking enterprise systems and, hence, to leveraging the experiences of many organizations.

It seems that SNA has been pushed to the backwaters of networking and systems as application developers and end users become increasingly focused on TCP/IP and the Web. Unknown to many, there is still significant business value that can be leveraged from their SNA investments.

Change is inevitable; leveraging the past can ease this change. Knowing where you want to go even further eases the impact of change. Strategically adopting a service delivery model leverages the past while enabling the new.


Application Business Logic
Server Provided
Application Specific Services
Application Support Services
Network Provided
Protocol Services
Transport Services
Clients

Service Delivery

Working with many companies, especially with the influence from the emerging service provider business segment, we have found that looking at infrastructure from a perspective of what services it provides to the devices that attach to it can simplify the way we view today's complex IT environment. By looking at the services that are provided by the infrastructure, we can determine which ones are critical to the business and which ones have little or no usage. We can identify where duplicate services exist or where a service may be missing. By adopting a service delivery model, an organization can deploy new applications faster because the required underlying infrastructure is in place. They can reduce their overall cost by eliminating duplicate services, and they can improve the availability of their IT environment because the service delivery model provides an effective way to look at IT from an end-to-end perspective.

This approach works for centralized organizations that are becoming distributed, and for distributed organizations that are centralizing. Adopting a service delivery approach focuses the organization on the delivery of services to the end-user client, as opposed to the technologies and products that seem to change on a daily basis. The end-user client device can be a PC, a browser, a personal digital assistant (PDA), a phone, and so on. To remain competitive, IT organizations are transitioning to become service providers for their companies. In doing so, the biggest challenges they face are provisioning the required services and meeting the service-level agreements and objectives that have been set between them and the end users. Over the past few years, the typical focus has been on the delivery of network-provided services (transport and protocol) in the lower two layers of the service delivery model shown on the left page.

Progressive organizations are raising their sights into the higher layers of this service delivery model to better enable their business by:

- Enabling new capabilities that in turn enable new business opportunities and associated revenue growth
- Optimizing their infrastructure to reduce costs
- Improving their infrastructure to achieve continuous availability and better response times

These higher layers can be divided into two types of services: application-support services and application-specific services.

As shown in the figure, application support services are built on top and have affinity with the networking protocols. As the name suggests, applications rely on these support services, and hence, when a network support service fails, a broad range of applications will be affected. Examples of these services include host access, directory service, dynamic host configuration protocol, messaging, file transfer, print and Web serving. Application support services impact the largest number of applications and should be architected for the highest level of availability. Additionally, the number of technologies used to provide the same service (capability) to the end user should be minimized, and reduced to one if possible, to simplify support while reducing costs. Reducing the number of these options to one simplifies the client and enables the organization to focus on higher availability of this one service.

Application-unique services are built on top of the application support services. They provide a set of capabilities that are specific to the application, and therefore impact only that set of applications. Examples of these services on the S/390 server are:

- Transaction Processing: CICS® and IMS™
- Enterprise JavaBeans™
- Application-unique API (LU6.2, LU0 and Sockets)
- Lotus Notes®
- PeopleSoft
- SAP R/3

When moving into these higher layers of the networking model, the integration of the network and the server becomes paramount, hence the need for a systems or enterprise view of computing. For enterprise servers, we have found through our extensive work with our clients that there are four infrastructure architectures, given the types of services that the S/390 server delivers.

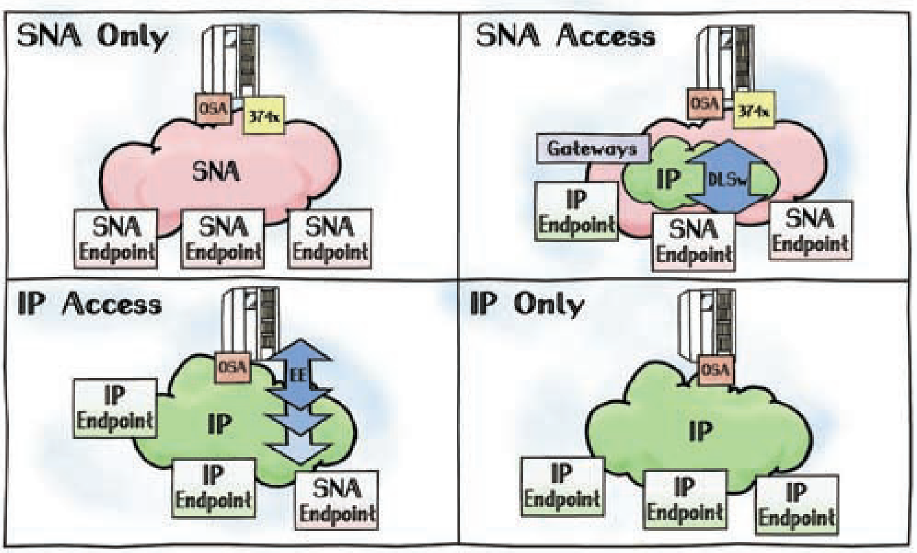
The Four Infrastructure Architectures for Enterprise Computing

When building a service delivery model with the S/390 and z900 server, it is important to understand what application support and application-unique services need to be delivered, what protocol will deliver those services (SNA, IP, or both), and what transport will be used to carry those protocols. The four infrastructure architectures for enterprise computing are:

- SNA only
- SNA access
- IP access
- IP only

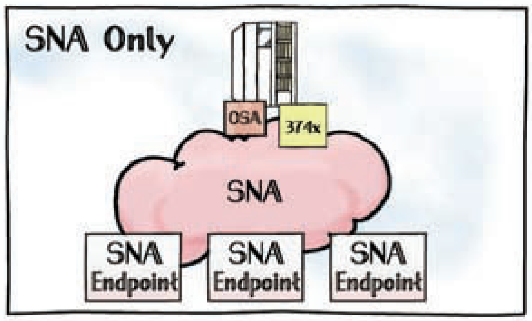
From a systems or enterprise view, the driving factors that cause these models to differ are:

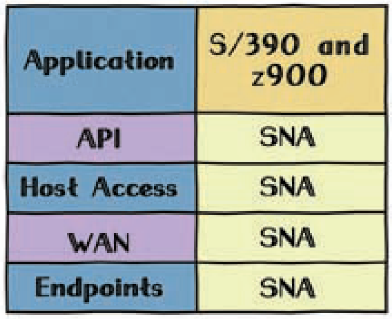
- The application programming interface (API), either SNA or IP, that is used to access the application on the enterprise server
- The host-access method, either SNA or IP, used to integrate the enterprise server with the network
- The WAN protocol, either SNA or IP, used to communicate to the endpoint
- The endpoint, either SNA or IP, used to deliver the application



SNA Only

Most applications, networks and endpoints were SNA Only a decade or more ago. Companies made significant investments in mainframe host-based SNA applications and associated networking (VTAM®, NCP, 3745, 3746, SDLC, the NetView® program, and so on).





Application	S/390 and z900
API	SNA
Host Access	SNA
WAN	SNA
Endpoints	SNA

The Advanced Peer-to-Peer Networking® (APPN®) extension to SNA was an enhancement that supported communications with and between midrange systems, and enhanced the ability of SNA to accommodate change dynamically.

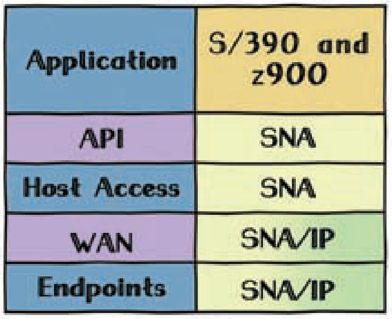
Due to the high cost of host cycles at that time, mainframe systems were usually front-ended by communication controllers such as the IBM 3745 Communication Controller running NCP, which provided, among other things, line attachments and offloaded link-level protocol support and device management. Wide area network (WAN) links carried traffic from SNA-based endpoints such as 3270 terminals or personal computers with 3270-emulation software. Fewer SNA-only networks remain today.

IBM introduced SNA networking almost 30 years ago, at a time when WAN bandwidth was limited and expensive. Consequently, SNA products were originally designed for very efficient support of less-reliable, low-bandwidth media environments. Although the cost of bandwidth and the availability of higher-bandwidth links have improved substantially over the years, there remain parts of the world where the efficiency of SNA still plays a critical role in support of business operations. Also, given the maturity of the SNA networking products, SNA-only networks enjoy a level of stability that would be very difficult to match with newer networking technologies at this time.

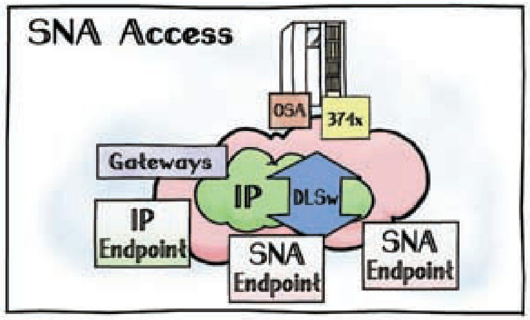
Applications and endpoints using networking protocols other than SNA (such as NetBIOS and TCP/IP) grew in popularity in the 80s and 90s. Consequently, networks had to evolve to carry those protocols in addition to SNA. Some networking solutions, such as multiservice switches or frame relay, separated protocols into different logical networks over shared physical links. Other solutions, such as Data Link Switching (DLSw) transported other protocols (SNA, NetBIOS) over an IP-based network. Both approaches are widely used today.

SNA Access

Also, in order to leverage the investment in SNA applications and host-based data, organizations began to deploy gateway devices that allow non-SNA endpoints to access SNA-based applications and data without requiring changes to the applications or host-access devices (for example, 3745s). These gateways were often placed in the data center just outside of the existing host-access devices. TN3270 servers, which allow TCP/IP-based personal computers (running special 3270 emulation "Telnet client" code) to access SNA applications, are the most commonly found examples of such gateways today. As the need for SNA access grew, gateway use proliferated, and many companies found themselves struggling with the challenges of scaling and managing gateway environments. Gateways also negatively impacted SNA application performance and availability.

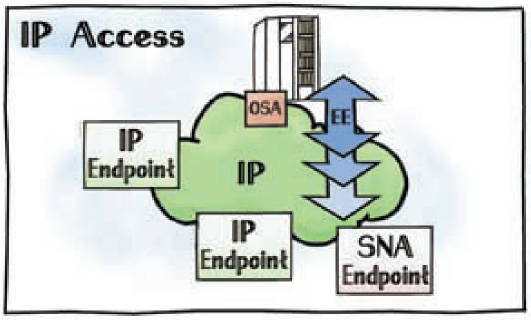


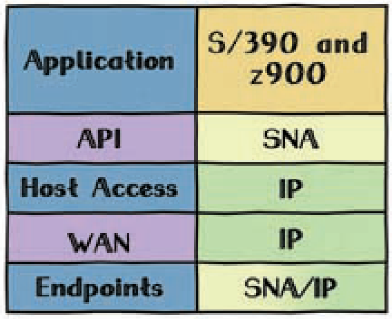
Application	S/390 and z900
API	SNA
Host Access	SNA
WAN	SNA/IP
Endpoints	SNA/IP



IP Access

Recognizing the strategic importance of TCP/IP, many organizations have focused on IP enablement of their host environments using the Open Systems Adapter-Express (OSA-Express) as a low-cost, high-performance, IP Access into their host environment. They have also used later releases (V2R8 and later) of the OS/390® software for its TCP/IP functionality. This positions the organizations for the deployment of IP-based applications into the highly scalable and very highly available S/390 or z900 environment. In recent years, the cost of mainframe host cycles has fallen to the point where it is usually less expensive to move gateway function (such as TN3270 servers) into the host environment while, at the same time, reducing complexity and improving performance and availability to SNA applications. Because of its scalability and availability, the mainframe host environment may also prove to be the ideal place to run certain Web services — particularly those that need access to SNA applications or data.



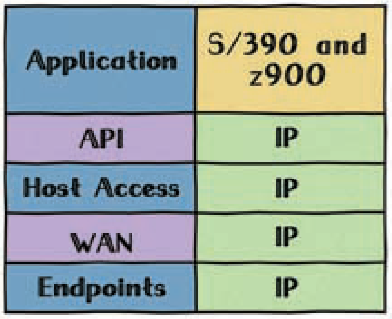


Application	S/390 and z900
API	SNA
Host Access	IP
WAN	IP
Endpoints	SNA/IP

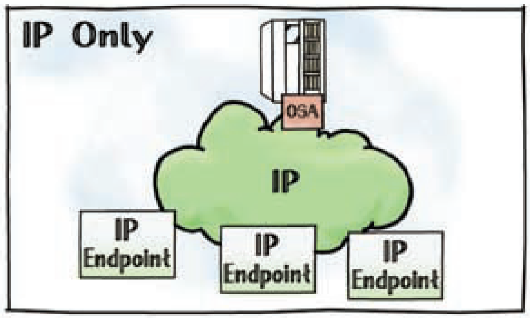
SNA endpoints, "legacy devices," and specialized logical unit types (for example, LU0, LU6.2) can be given access to host applications using Enterprise Extender (EE), which very efficiently transports SNA traffic over an IP WAN and host-access infrastructure. EE is supported at the remote site in some routers and in most servers (for example, with IBM Communications Server for Windows NT® or IBM Communications Server for AIX®). EE is also integrated into OS/390 V2R6 and later.

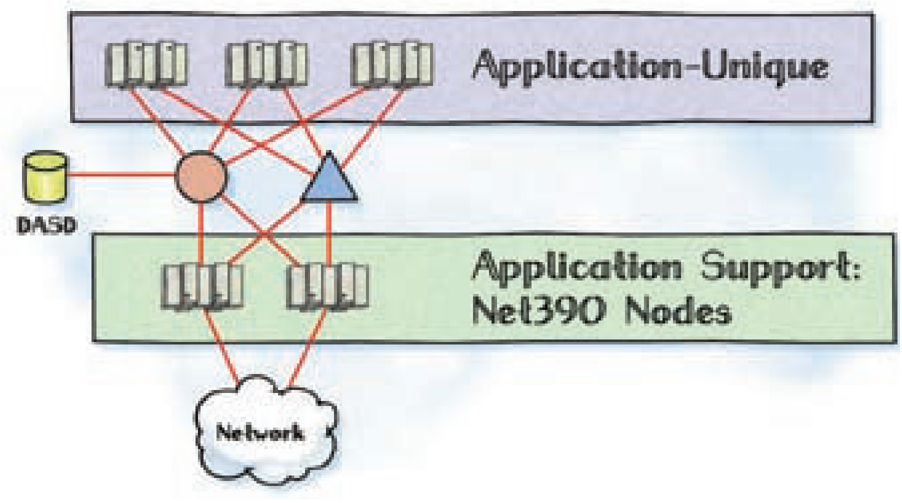
IP Only

As organizations replace SNA endpoints with IP endpoints or Web browser-based endpoints, it will eventually no longer be necessary to support SNA endpoints. Likewise, SNA applications may be changed to support IP APIs or be replaced with IP applications as business processes are reengineered. Such transitions will likely take a very long time for most organizations; however, such an IP-only environment may be the ideal infrastructure "platform" to position an organization to more quickly support unanticipated business requirements as they arise, while minimizing complexity and operational costs.



Application	S/390 and z900
API	IP
Host Access	IP
WAN	IP
Endpoints	IP





Data Center Architecture for Enterprise Computing: Net390

The Net390 structure, which enables the S/390 and the z900 in the Parallel Sysplex® configuration to be the premiere service-delivery platforms in the industry, using separate logical partitions or servers in the Parallel Sysplex configuration for application-specific and application-support services, as shown in the figure above. The value of this approach is that the application-support nodes, which should be continuously available, can be isolated from the applications that run on the application-unique logical partitions. This protects the application-support services from any potential availability impact that the applications might present, while it enables an organization to deploy new IP services on the S/390 and z900 without impacting the existing application. This reduces the risk to the business as new S/390 applications are added or scaled.

Additionally, hardware and software upgrades to the application-support nodes, called Net390 nodes, can be performed independently of the application. In the SNA-only and SNA-access infrastructure architectures, which have been in place for the past 10 - 20 years, these logical partitions were known as *Communications Management Configuration (CMC)* logical partitions with well-understood and proven benefits. In effect, Net390 extends the benefits of the CMC design to IP-access and IP-only enterprise models.

As shown in the figure above, the Net390 nodes provide access and integration to the network. The back-end application-unique nodes need not directly attach to the network, though in most cases they will be attached. New application-unique nodes can be added without incurring changes in the network. This architecture also reduces the complexity of the application-unique nodes because they rely on the Net390 nodes for all of their underlying services.

S/390 and the z900 are well known for providing high-volume online processing services and for their ability to intermix and manage multiple application workloads and support large batch workloads. For online processing, the predominant transaction managers are CICS and IMS, and access to these subsystems has historically been provided by SNA. Today's customers are also leveraging the S/390 and z900 into Web serving and Linux environments. To fully leverage S/390 and z900 into the e-business infrastructure, three major changes need to be accommodated. These are:

- The implementation of Parallel Sysplex architecture to enable continuous availability and unrestrained scalability of the host complex
- The enablement of IP access for improved reach and heightened performance
- The accommodation of SNA to leverage and protect the existing investment

Net390 is an advanced network computing system that enables all three of these changes. It is built atop a S/390 Parallel Sysplex configuration and is extended into the network by combining the best attributes of SNA with the flexibility and reach of TCP/IP and the Web. The benefits of a Net390 are:

Continuous Availability

Net390 enables continuous availability to the end user by integrating dynamic network routing protocols (for example, OSPF and High-Performance Routing [HPR]) into the continuous operation characteristics of the Parallel Sysplex server.

Scalability

By building on OS/390 features such as generic resources, Workload Manager (WLM), and Service Policy Agent, and by extending these features into the network with Enterprise Extender, TN3270 Server, and Sysplex Distributor, Net390 provides scalability and enhances performance.

Security

Net390 provides secure access to mainframe applications and data by leveraging the cryptographic features of the S/390 server and the network.

We have found that looking at infrastructure from a perspective of what services it provides to the devices that attach to it can simplify the way we view today's complex IT environment.

Investment Protection

Net390 protects current investments in existing applications, routed networks and SNA equipment (including the 3745 and 3746 Communication Controllers) by enabling OS/390 access from the intranet, Internet and extranets. Connectors between WebSphere® software on the S/390 server and backend systems (CICS, IMS and DB2® software) make existing data and application logic available to Web users.

Predictable Response Times

Net390 enables the preservation of SNA service levels, and leverages new networking technologies that enable IP service levels through the use of bandwidth management, application content-specific prioritization and workload distribution to achieve system-level response-time management.

Flexibility

Net390 is a flexible structure that is based on many existing and emerging technologies. As a best-practices approach, only those Net390 components that provide the greatest benefit, given current business requirements and timeframes, need to be implemented.

Next Steps

It is difficult to present in a single paper the exact configuration that will support your business, given your current investments and business strategy. It is for this reason that IGS has developed the Net390 project, which is based on the Center for Infrastructure Solutions work with numerous clients. A Net390 project helps you to make the appropriate strategy, infrastructure, product, design, system management, and implementation decisions required to leverage S/390 and z900 in the e-business infrastructure. The intent of the Net390 project is to:

- Review current and proposed applications that involve the S/390 and z900 to develop a Net390 service-delivery architecture capable of supporting your e-business model
- Recommend appropriate refinements to the IBM-developed Net390 architecture to support your business model, your investment in the S/390 server other server types, and the network, while exploiting Parallel Sysplex scalability and availability
- Architect the data center network structure, including intra-data center, data center-to-data center, and disaster recovery as appropriate to optimally support IP and SNA access to the S/390 server
- Architect and design, based on your business, the Internet access network that provides the integration between the S/390 server and the Internet/extranets
- Provide a detailed plan that identifies the steps to enable you to build a Net390 environment

The value of this session is that it allows you to maximize your existing data, SNA applications, endpoints and infrastructure by bringing together appropriate business units to address your overall company's e-business infrastructure needs. The solution focuses on:

- Preserving your existing investment in hardware and application software
- Taking advantage of the opportunities created by the migration to e-business solutions
- Creating a highly reliable, scalable, cost-effective solution with the flexibility to react and adjust to future business requirements
- Identifying a clear set of steps to achieve your required e-infrastructure

For more information on Net390 workshops, visit our Web site at **ibm.com**/services/its/us/netcenter.html

You can also send a note to netctr@us.ibm.com or talk to your local IBM representative.

To learn more about z900 and S/390 networking visit **ibm.com**/servers/eserver/zseries/networking ■

About the Authors

Bob Louden is a Certified IT Specialist in networking at IBM, specializing in enterprise systems, wide area networking, local area network interconnection, SNA and TCP/IP. His 19 years of experience in networking have enabled him to consult with a wide variety of businesses and help them to design optimal solutions to meet their business needs. Bob has also taught IBM Education courses including broad discussions of data communications and the details of SNA and TCP/IP.
blouden@us.ibm.com

Robert Brinkman is an IBM consultant working in the Center for Infrastructure Solutions in RTP, NC. In this capacity, he uses his 20-year background in networking and systems to assist clients with the their infrastructure strategies, architectures and designs to support their changing business requirements. Robert has worked with over 300 U.S. and international clients to enable them to deploy infrastructure quickly and effectively.
His primary focus is networking, including transport technologies, addressing, routing protocols, DNS, DHCP, and Quality of Service (QoS) mechanisms. Robert has supported all types of industries including process, engineering, higher education, K-12, manufacturing, banking, insurance, telecommunications, and retail. He has in-depth knowledge of major applications such as SAP R/3, Lotus Notes®, Web, and TPF and the impact of these application on infrastructure.
rbrinkma@us.ibm.com

Jim Goethals has been associated with IBM networking products for 24 of his 32 years at IBM. Currently the z900 and S/390 Networking Offering Manager, Jim specializes in enterprise customer networking requirements and solutions. Jim is responsible for marketing of the OSA-Express adapter, for the IBM z900 and Cisco joint e-business solutions, for TCP/IP and SNA support in CS for OS/390, and for helping customers leverage TCP/IP and their current SNA investments while evolving to the world of e-business on z900 and S/390 servers.
jimgo@us.ibm.com

Analyzing Traffic Problems with the NTA TCP/IP Analysis Tool

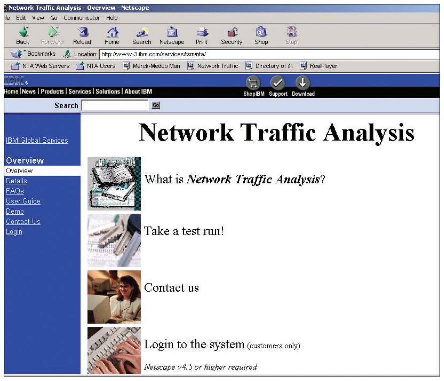
By Bob Springsteen

NTA has been available since 1989, with the Web-based TCP/IP analysis tool becoming available in mid-1999.

This article takes you through an analysis using NTA's Web-based TCP/IP tool. We will use a trace file with IP packets captured by a LAN protocol analyzer. An analysis can be done on many LAN-captured traces or an OS/390® packet CTRACE.

Because this tool is on the Web and has a demo function that allows actual analysis of a trace, you can follow along if you have Internet access and Netscape Version 4.5 or higher.

First, go to the demo at **ibm.com**/services/tsm/nta/. Click on the last picture on this page, with the caption "Login to the system."


Network Traffic Analysis - Overview - Netscape
ile Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location:
NTA Web Servers NTA Users Merck-Medco Man Network Traffic Directory of /n RealPlayer
IBM
Home | News | Products | Services | Solutions | About IBM
ShopIBM Support Download
Search
Go
IBM Global Services
Overview
Overview
Details
FAQs
User Guide
Demo
Contact Us
Login

Location: http://www-3.ibm.com/services/tsm/nta/

Network Traffic Analysis

What is *Network Traffic Analysis*?

Take a test run!

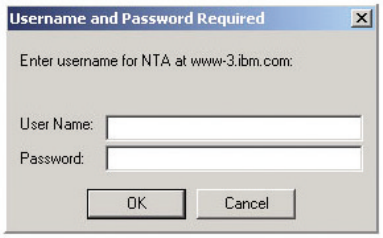
Contact us

Login to the system (customers only)

Netscape v4.5 or higher required

IBM Global Services (IGS) can now analyze TCP/IP performance problems very quickly using a TCP/IP analysis tool that is part of IBM's Network Traffic Analysis (NTA) service offering.

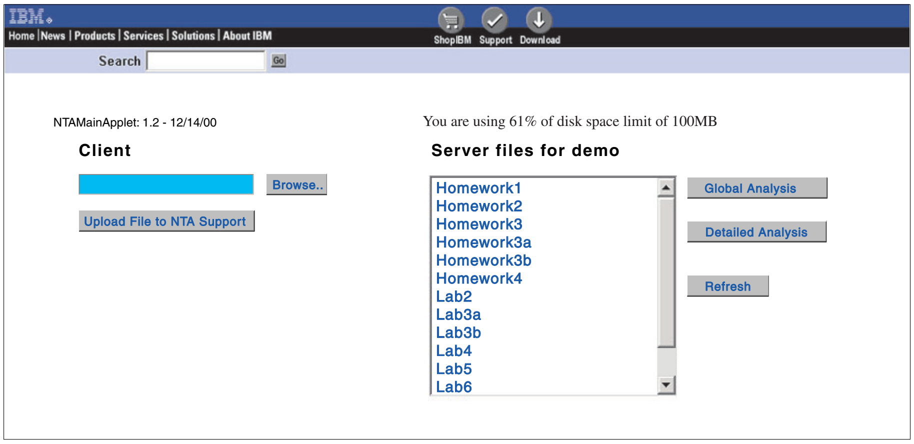
A pop-up window will ask for your user ID and password. Enter **demo** for the user ID, leave the password field blank, and click **OK**.



Enter username for NTA at www-3.ibm.com:

The next display shows a list of traces under "Server Files for demo."

Click **homework1.** This trace will then be highlighted. Now click **Detailed Analysis.**


IBM
Home | News | Products | Services | Solutions | About IBM
ShopIBM Support Download
Search
Go
Browse..
Upload File to NTA Support
Global Analysis
Detailed Analysis
Refresh

NTAMainApplet: 1.2 - 12/14/00

You are using 61% of disk space limit of 100MB

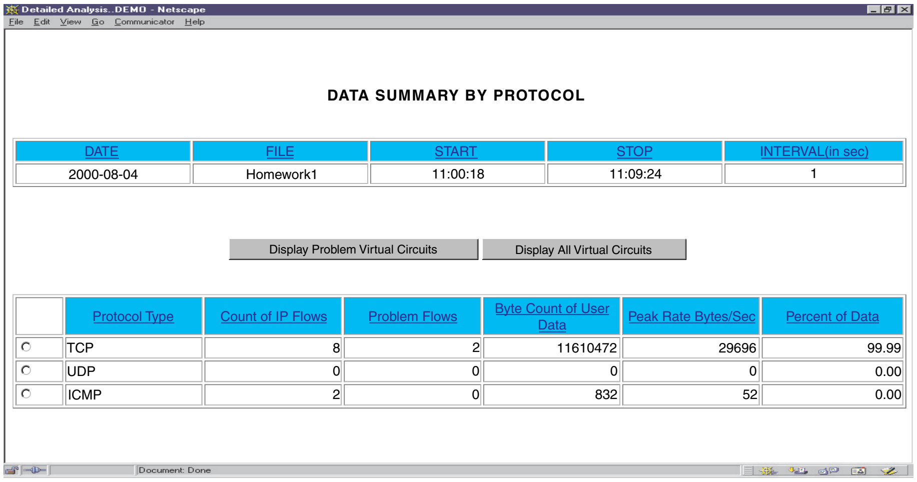
Client

Server files for demo

Homework1
Homework2
Homework3
Homework3a
Homework3b
Homework4
Lab2
Lab3a
Lab3b
Lab4
Lab5
Lab6

Saving time when analyzing network performance problems is what NTA is all about.

This displays a count of IP flows and subprotocols such as **TCP.** The next-to-last column shows the peak 1-second data rate for all of the connections. Check the TCP radio button and then click **Display Problem Virtual Circuits.**


Detailed Analysis..DEMO - Netscape
File Edit View Go Communicator Help
Document: Done

DATA SUMMARY BY PROTOCOL

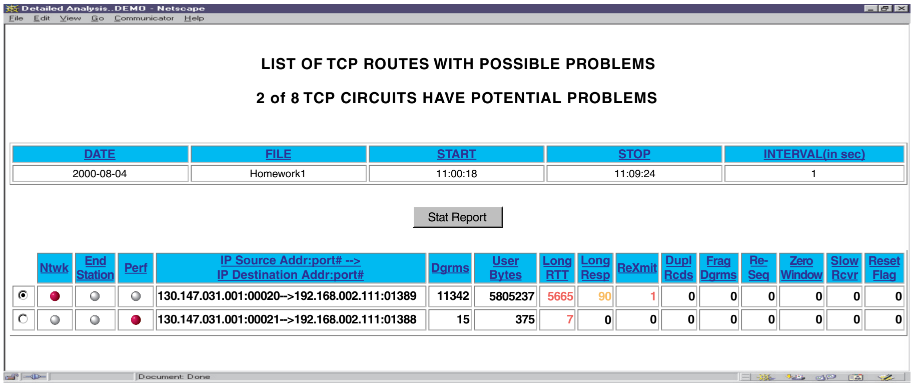
DATE	FILE	START	STOP	INTERVAL(in sec)
2000-08-04	Homework1	11:00:18	11:09:24	1

Display Problem Virtual Circuits | Display All Virtual Circuits

	Protocol Type	Count of IP Flows	Problem Flows	Byte Count of User Data	Peak Rate Bytes/Sec	Percent of Data
○	TCP	8	2	11610472	29696	99.99
○	UDP	0	0	0	0	0.00
○	ICMP	2	0	832	52	0.00

The next display shows the problem: 5665 packets with long round-trip times (RTTs). The RTT is the time that elapses between sending a packet and receiving its acknowledgment (ACK). The RTT count is incremented when a packet's RTT is 200 ms or more.

In this example, the tracing is at the server, which is sending the data packets. This means that you actually see the full network time for this socket pair. Keep in mind that the round-trip time (RTT) for data packets received at the trace point, inbound to the server, will simply be the IP stack processing time. In this case, you are able to see the problem clearly because the trace point is at the sender. If you wanted to determine to what extent the 300-ms delay was in the network rather than at the client, you would analyze a trace taken at the destination of the packet flow, which is the client. Although you can quickly see the cause of the poor performance, you still don't know how to improve it.


Detailed Analysis..DEMO - Netscape
File Edit View Go Communicator Help
Document: Done

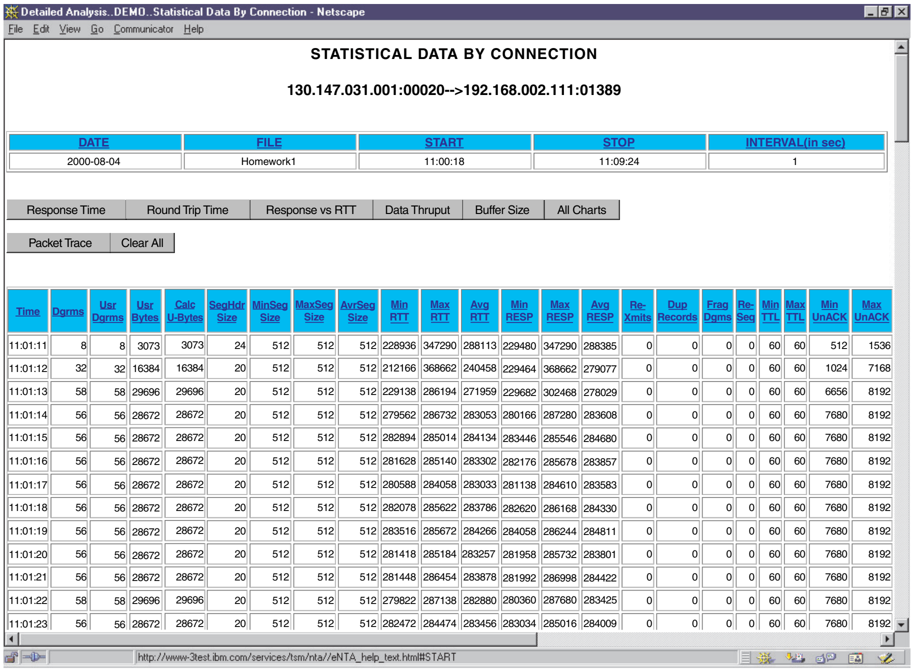
LIST OF TCP ROUTES WITH POSSIBLE PROBLEMS

2 of 8 TCP CIRCUITS HAVE POTENTIAL PROBLEMS

DATE	FILE	START	STOP	INTERVAL(in sec)
2000-08-04	Homework1	11:00:18	11:09:24	1

Stat Report

	Ntwk	End Station	Perf	IP Source Addr:port# --> IP Destination Addr:port#	Dgrms	User Bytes	Long RTT	Long Resp	ReXmit	Dupl Rcds	Frag Dgrms	Re-Seq	Zero Window	Slow Rcvr	Reset Flag
◉	🔴	⚪	⚪	130.147.031.001:00020-->192.168.002.111:01389	11342	5805237	5665	90	1	0	0	0	0	0	0
○	⚪	⚪	🔴	130.147.031.001:00021-->192.168.002.111:01388	15	375	7	0	0	0	0	0	0	0	0


Detailed Analysis..DEMO..Statistical Data By Connection - Netscape
File Edit View Go Communicator Help
Response Time
Round Trip Time
Response vs RTT
Data Thruput
Buffer Size
All Charts
Packet Trace
Clear All
http://www-3test.ibm.com/services/tsm/nta//eNTA_help_text.html#START

STATISTICAL DATA BY CONNECTION

130.147.031.001:00020-->192.168.002.111:01389

DATE	FILE	START	STOP	INTERVAL(in sec)
2000-08-04	Homework1	11:00:18	11:09:24	1

Time	Dgrms	Usr Dgrms	Usr Bytes	Calc U-Bytes	SegHdr Size	MinSeg Size	MaxSeg Size	AvrSeg Size	Min RTT	Max RTT	Avg RTT	Min RESP	Max RESP	Avg RESP	Re-Xmits	Dup Records	Frag Dgms	Re-Seq	Min TTL	Max TTL	Min UnACK	Max UnACK
11:01:11	8	8	3073	3073	24	512	512	512	228936	347290	288113	229480	347290	288385	0	0	0	0	60	60	512	1536
11:01:12	32	32	16384	16384	20	512	512	512	212166	368662	240458	229464	368662	279077	0	0	0	0	60	60	1024	7168
11:01:13	58	58	29696	29696	20	512	512	512	229138	286194	271959	229682	302468	278029	0	0	0	0	60	60	6656	8192
11:01:14	56	56	28672	28672	20	512	512	512	279562	286732	283053	280166	287280	283608	0	0	0	0	60	60	7680	8192
11:01:15	56	56	28672	28672	20	512	512	512	282894	285014	284134	283446	285546	284680	0	0	0	0	60	60	7680	8192
11:01:16	56	56	28672	28672	20	512	512	512	281628	285140	283302	282176	285678	283857	0	0	0	0	60	60	7680	8192
11:01:17	56	56	28672	28672	20	512	512	512	280588	284058	283033	281138	284610	283583	0	0	0	0	60	60	7680	8192
11:01:18	56	56	28672	28672	20	512	512	512	282078	285622	283786	282620	286168	284330	0	0	0	0	60	60	7680	8192
11:01:19	56	56	28672	28672	20	512	512	512	283516	285672	284266	284058	286244	284811	0	0	0	0	60	60	7680	8192
11:01:20	56	56	28672	28672	20	512	512	512	281418	285184	283257	281958	285732	283801	0	0	0	0	60	60	7680	8192
11:01:21	56	56	28672	28672	20	512	512	512	281448	286454	283878	281992	286998	284422	0	0	0	0	60	60	7680	8192
11:01:22	58	58	29696	29696	20	512	512	512	279822	287138	282880	280360	287680	283425	0	0	0	0	60	60	7680	8192
11:01:23	56	56	28672	28672	20	512	512	512	282472	284474	283456	283034	285016	284009	0	0	0	0	60	60	7680	8192

By clicking the radio button for the socket pair with high RTTs and then clicking **Stat Report,** you can display a statistical report that shows about 40 fields. Each row represents a 1-second collection of data for this socket pair

(130.147.031.001:00020-->192.168.002.111:01389).

Notice that the RTT columns show consistent values a little less than 300 ms. The data rate is mostly 28 672 bytes per second and the maximum segment size is 512 bytes. That's small and is probably a default value that should be changed. Most important, look at the field labeled **Max RcvWin.** You will need to be looking at the statistical report after login to NTA to see this field. There are too many fields in this report to display all of them easily in the figure, and therefore, the **Max RcvWin** field is not shown here. Its value is 8576. This means that the receiver of this socket pair, 192.168.002.111:01389, is sending back an advertised window of 8576 bytes.

Now look at the field labeled **Max Unack.** This is the count of the maximum number of bytes that are in transit. In other words, it is the number of bytes transmitted by the sender but not yet acknowledged by the receiver. This value can be as large as the advertised window, or the congested window size. Its value is 8192 bytes. The problem is that the advertised window of 8576 bytes is too small. The sender transmits a window of data and waits for an ACK. The ACK takes almost 300 ms to come back. Note that the number of bytes in transit, plus one more packet of 512 bytes, would mean that the number of bytes in transit would be 8704 (8192+512), which is larger than the receiver's advertised window allows.

The number of unacknowledged packets in transit is 16 (8192 ÷ 512). To get better throughput, the client will need to have a larger TCP receive buffer. The current size, 8576 bytes, is not large enough to allow for continuous transmission with a 300-ms round-trip delay. Setting the client's TCP receive buffer to at least 16K bytes should dramatically improve data throughput beyond the current 28 672 bytes per second.

The figure shows only a portion of the fields available in the statistical report that we use to determine the problem cause. There are many other reports available, including "Global" reports that provide information from all packets traced. You can display these reports while you are logged in to the demo.

In summary, this demo shows you how quickly a performance problem can be detected and fixed through a trace analysis using the NTA tools. Saving time when analyzing network performance problems is what NTA is all about. If you have any questions on the NTA service offering, please call the NTA team 1 800 876-8801 or 919 461-5131 from outside the U.S. and Canada, or e-mail us at ntateam@us.ibm.com. ■

About the Author

Bob Springsteen has worked in the IBM Network Traffic Analysis (NTA) Group. Bob was one of the original developers of NTA in 1987 and has retired as of April 2001. The NTA group helps clients resolve network performance problems in SNA and TCP/IP using NTA's expert system analysis tools.

Communications Server and the Open Systems Adapter: Optimized Integration for S/390 Networking

By Art Stagg

Rapid technological change has brought about a revolution in the manner in which business is conducted. The terms “Global Market” and “Global Economy” have moved from being trade-press buzzwords to being accepted aspects of the mainstream business world.

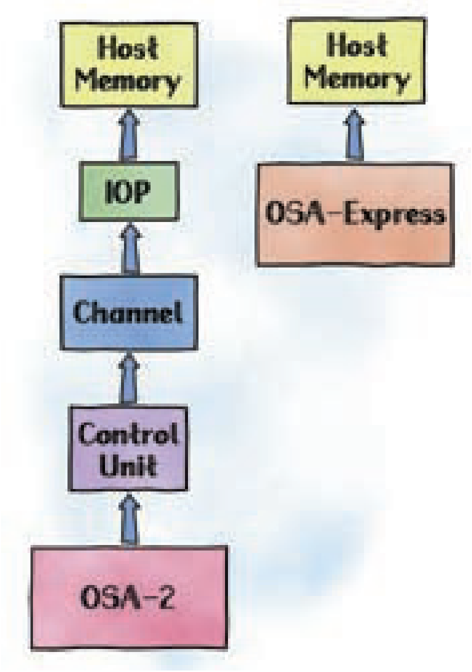
Today, it is possible for any vendor to reach potential customers anywhere in the world — an ability that was previously reserved for very large enterprises. This global revolution has had a profound effect on the business model, the effects of which are only now beginning to be understood.

One impact of the evolution of the global economy has been an increase in competition at all levels of the corporate community. We note the IBM commercial where the Japanese corporate conglomerate has a major supply problem — until the small business in Texas sends in a bid. Amusing it is, but it also underscores the growth of intense competition brought about by the exploitation of new and evolving technology by the business community.

This evolution has also introduced additional strain on the Internet. Businesses require the means to distinguish the shopper from the browser and to allocate network resources accordingly. An additional requirement is to provide the means to associate the allocation of system and network resources on a client basis, that is, preferred client optimization.

To assist you in meeting these challenges, IBM has introduced the OSA-Express Adapter for the IBM @server zSeries 900 and the S/390® Generation 5 and Generation 6 Parallel Enterprise Server™ platforms. This latest edition to the OSA family of adapters, coupled with substantial improvements in the Communications Server (CS) TCP/IP protocol, provides:

- World-class standard network attachment for Gigabit Ethernet, Fast Ethernet, and 155-Mbps asynchronous transfer mode (ATM). No longer is a special front-end processor required to connect your S/390 or z900 server to the network.
- Standards-based protocols to further eliminate the requirement for a front-end processor to translate between older channel based-interfaces and industry-standard network-attachment interfaces.
- Leading-edge performance, achieved by the use of a brand-new I/O attachment interface designed explicitly to provide higher bandwidth, lowerlatency and reduced CPU network costs.
- Dynamic configuration of the adapter through a direct interface with the TCP/IP stack, which simplifies network configuration integration, making it less prone to “user” definition errors.



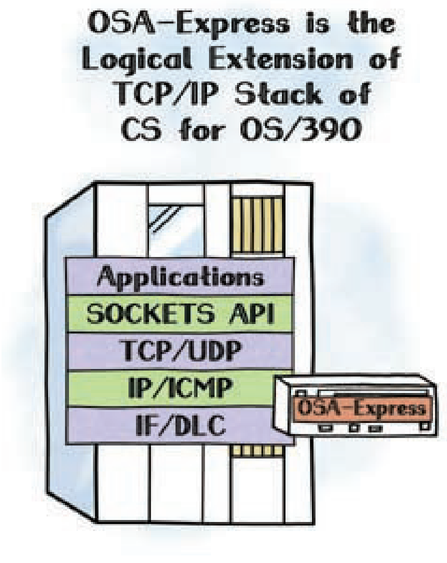
Introduced in OS/390® Release 7, the OSA-Express Gigabit Ethernet feature interface, utilizing new Queued Direct Input/Output (QDIO) interfaces, provides the means to economically exploit the new high-speed network media required to be competitive in today's dynamic market. OS/390 Release 8 provided QDIO support for the OSA-Express Fast Ethernet and ATM features.

Overview of OSA-Express Structure

OSA-Express is the third generation of the Open Systems Adapter (OSA) family developed by IBM for the S/390 and new z900 servers. The following figure provides an overview of the structure of the OSA-Express adapter and shows how the QDIO design reduces system resources required for network I/O.

The OSA-2 interface, like all traditional channel-based interfaces for controllers and routers, required the use of multi-stage hardware and software interfaces in order to transport data to and from the S/390 server. The OSA-Express QDIO interface replaces the OSA-2 interface with a single-step direct memory access from the OSA-Express adapter into S/390 and z900 memory. The removal of the intervening steps, and the related simplification of the I/O supervisor, results in reduced system cost for Network I/O.

Further, as shown in the figure, the new OSA-Express interface provides a logical extension of the CS for OS/390 TCP/IP network protocol. For example, IP functions are shared between the CS for OS/390 network stack and the OSA-Express adapter. The sharing of functions is determined by efficiency. Generally, the approach is to place compute-intensive functions on the OSA-Express adapter, thereby allowing more system cycles to be utilized by user applications.


OSA-Express is the Logical Extension of TCP/IP Stack of CS for OS/390
Applications
SOCKETS API
TCP/UDP
IP/ICMP
IF/DLC
OSA-Express

Specifically, OSA-Express provides the following functions and benefits:

- Automatic creation and updating of the Address Resolution Protocol (ARP) routing table for initial configuration as well as for dynamic updates
- No requirement for OSA/SF configuration
- Default routing, for example, the routing of "unknown" IP addresses on inbound data to a user-specified TCP/IP instance
- IP filtering to block non-IP datagrams from entering the system
- Dynamic access to media MIB data as well as ARP cache statistics

The figure below depicts the logical placement of OSA-Express within the CS for OS/390 TCP/IP protocol.

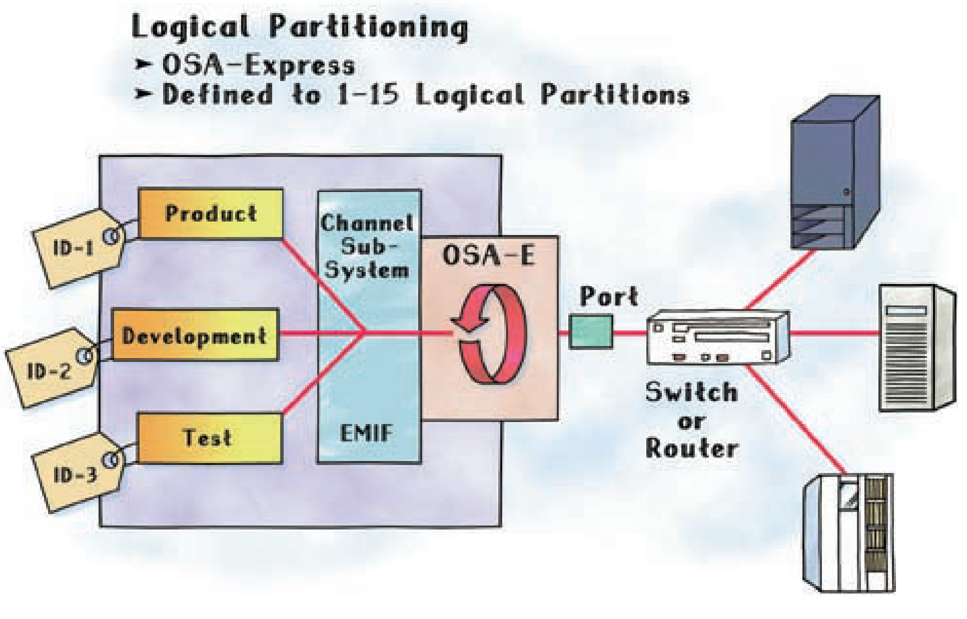
In addition, OSA-Express allows sharing of network attachment across up to 15 logical partitions. This reduces the total cost of network attachment by minimizing the need for multiple OSA-Express adapters, and optimizes the utilization of ports on expensive routers and switches.

The following figure demonstrates a possible configuration in a hypothetical S/390 data center.

In this example, the system has been configured to provide the following environments:

- Production
- Development
- Test

Each environment acts independently, with no cross-dependencies within the S/390 server. In this case, the network administrator controls the sharing of the OSA-Express, the allocation of network bandwidth, and access to downstream devices using administrative controls and traffic prioritization. Production would be set at the highest priority, with development and test prioritized based on internal rules.



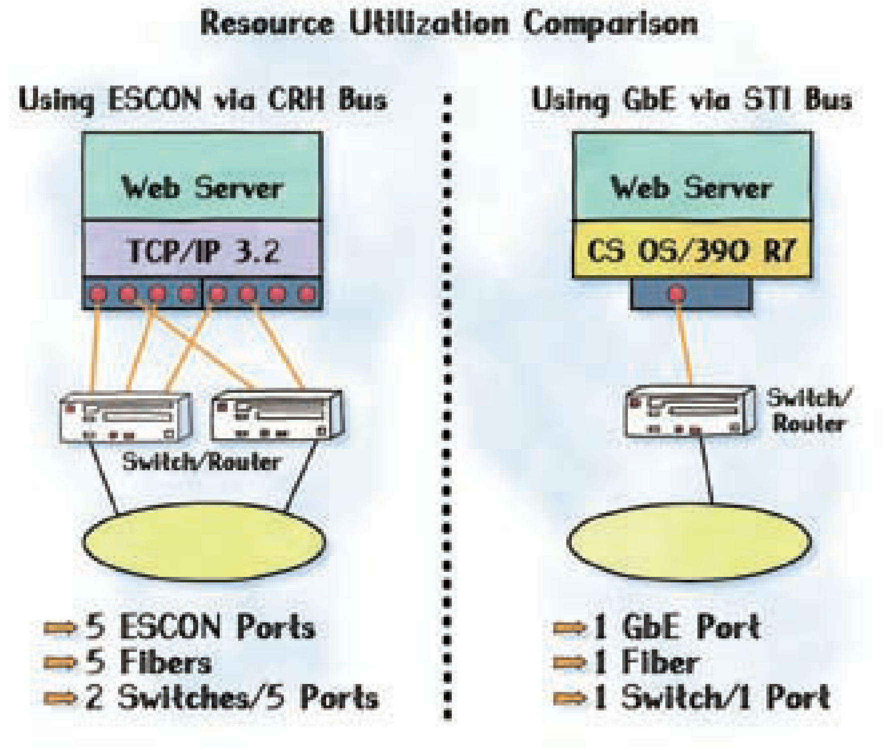
Cost Reduction and Resource Utilization

As previously mentioned, the new QDIO architecture provides a highly optimized interface for the transport of data to and from the S/390 or z900 server. The optimizations provided are tightly integrated into the CS TCP/IP protocol stack, and provide:

- Improved dispatching of network tasks
- More efficient storage management
- Optimized I/O supervisor interfaces

These improvements help to reduce the cycles required for a given unit of network-related work. The saved cycles can then be returned to the user application set.

Another aspect of cost savings relates to comparing the more costly front-end processors and routers with the cost of the OSA-Express adapter. The following figure demonstrates the resource reduction when comparing the channel environment with the OSA-Express environment.



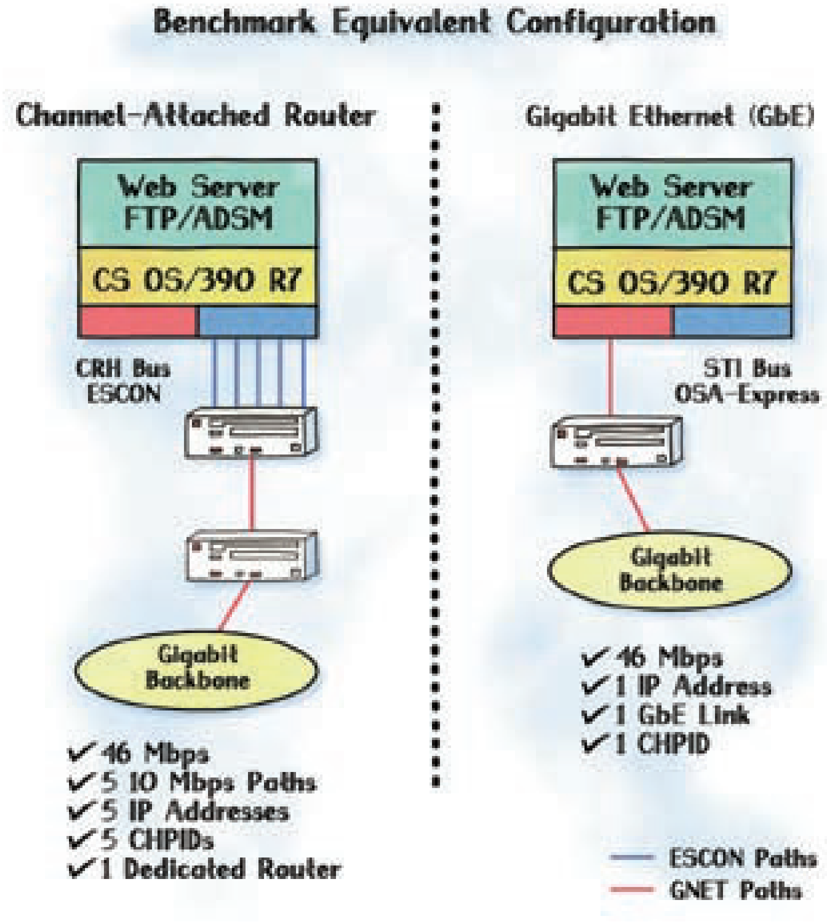
In this configuration, the use of OSA-Express has reduced cost in the following areas:

- The number of ESCON® channels required versus the use of a single OSA-Express, a five-to-one ratio. Note that not all users necessarily experience this kind of reduction. Actual reduction is a function of the aggregate bandwidth required.
- A reduction of the number of ESCON Director ports required for comparable configurations. These ports can now be made available for DASD attachment.
- A reduction of the number of ESCON ports from five to one.
- Elimination of the requirement to use expensive front-end routers.

The cost savings in moving from the expensive router to the OSA-Express configuration can be substantial. There is also another aspect of the savings that OSA-Express can provide, and that is the overall cost of system and network management. "People" cost is often the single most expensive aspect of the overall total cost of ownership (TCO) of a particular system. The OSA-Express helps reduce that cost by simplifying the work required to configure and manage the system and the network interfaces.

For example, if you consider a configuration in which three ESCON ports are replaced by a single OSA-Express adapter that is to be shared by two TCP/IP instances resident in two logical partitions, the following configuration tasks are reduced:

IBM has introduced the OSA-Express Adapter for the IBM @server zSeries 900 and the S/390® Generation 5 and Generation 6 Parallel Enterprise Server™ platforms.



TCP/IP address management: In the ESCON approach, a total of six IP addresses are required, one for each ESCON port per TCP/IP instance. OSA-Express requires just two, one per TCP/IP instance. This enables a 66% reduction in IP address management.

I/O configuration: Assuming that the network administrator has installed either MPC+ or CLAW channel interfaces, a total of 12 device address must be defined and mapped to the appropriate router. In the OSA-Express environment using QDIO, a total of only six device addresses must be defined.

These savings in people cost are yet another advantage of OSA-Express over the expensive traditional front-end or router configuration.

Performance

The new QDIO interface used by OSA-Express offers S/390 and z900 users a real performance boost over ESCON-based interfaces. The direct attachment of the adapter to the Self-Timed Interface (STI) bus, elimination of the use of the control unit image, the addition of direct memory access, and numerous optimizations within the CS TCP/IP protocol stack have all contributed to providing the S/390 and z900 user with leading-edge network connectivity.

Let's revisit the configuration used for the resource-utilization comparison, only this time to look at the real performance study on which it was based.

Note that the performance benchmarks for both configurations were identical (46 MBps sustained requirement). The resources required to meet that goal are greatly reduced in the OSA-Express environment.

The following table shows the throughput rate for the S/390 G5 server through a single OSA-Express Gigabit Ethernet adapter.

	Throughput (MBps)	
Connection	**OSA-Express 1500-Byte MTU**	**OSA-Express 9180-Byte MTU**
Single connection with RFC1323	33.5	46.0
Single connection without RFC1323	20.0	26.4
Multiple connection	48.0	75.0

The table shows actual test results using the standard 1500-byte MTU and the currently nonstandard (but widely adopted) 9180-byte MTU. The single connection test was run under two conditions: once using "Fat Pipes" (RFC1323[1]) and once not using Fat Pipes. The multiple connection test was only run with Fat Pipes. Returning to the previous figure, which compared the ESCON and OSA-Express benchmarks, it is worth noting that the OSA-Express numbers were obtained using the smaller MTU size. In short, one can state that a single OSA-Express is equivalent to at least five ESCON ports. The OSA-Express advantage can be expected to grow as the adapter evolves.

The performance advantage is also apparent for Fast Ethernet and 155 ATM. When attached using OSA-Express, both run full-duplex at media speed (100 Mbps and 155 Mbps respectively).

But what does all this extra throughput cost? Having all this bandwidth available is nice, but what effect does it have on my CPU? The following table shows the effect of using the OSA-Express QDIO interface on S/390 CPU utilization. The improvements (reductions) in S/390 CPU utilization for stream workload and a 1500-byte MTU were:

- 12.44% for the sender
- 12.47% for the receiver

	CPU Utilization	
Scenario	**Instructions per Byte for OSA-Express GbE**	**Instructions per Byte for MPC+ Channel**
Client (sender)	2.24	2.55
Server (receiver)	2.89	3.30

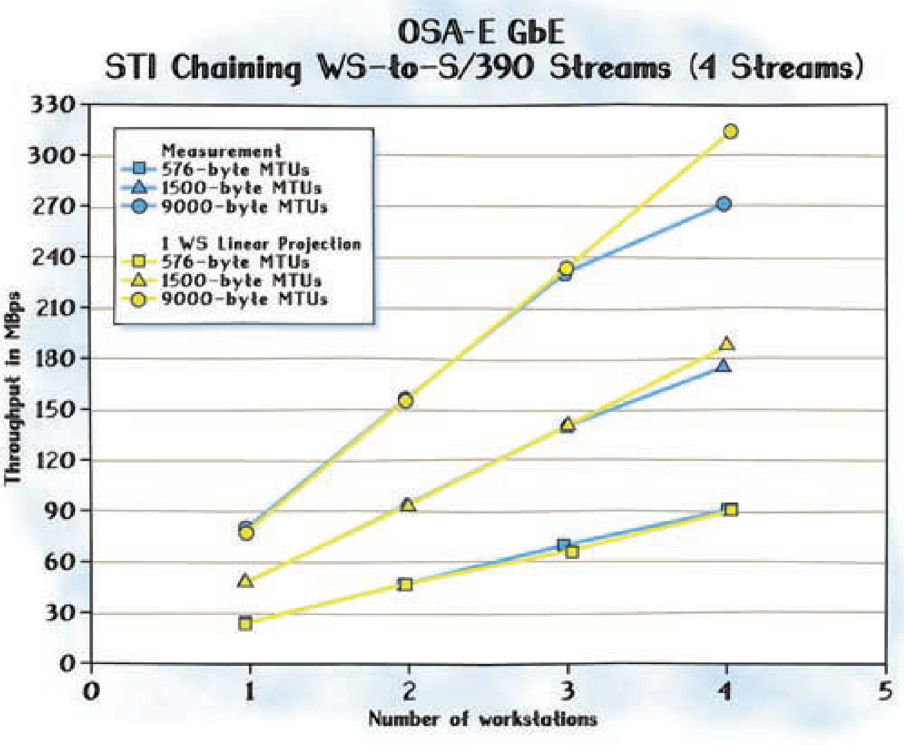
It is clear that OSA-Express delivers on its promise of higher bandwidth, increased network efficiency and reduced system cost. This is truly a case where one can "Have one's cake and eat it, too."

But what about scalablity? As your applications mature and your network grows, will you have to purchase dozens of OSA-Express adapters to keep up? What is the effective incremental benefit in the use of additional adapters?

The new QDIO interface used by OSA-Express offers S/390 and z900 users a real performance boost over ESCON-based interfaces.

[1] RFC1323, or "Fat Pipes," provides a mechanism by which the ends of a TCP connection can negotiate to grow the acknowledge window size from 64 KB up to 1 gigabyte.

To answer these questions, see the following figure showing the scalabilty of the OSA-Express adapter on a single Self-Timed Interface (STI) bus. The use of multiple adapters on a single STI bus is called chaining on a S/390 server and fanout on a z900.



In this S/390 configuration, multiple TCP/IP stream connections are used to concurrently drive up to four OSA-Express adapters. The X-axis represents the number of OSA-Express cards under test. The three sets of curves show the scalability of throughput for MTU sizes of 576 bytes, 1500 bytes and 9000 bytes as a function of the number of adapters. For each MTU size, a curve showing the theoretical scaled throughput for multiple cards, based on the throughput of a single card, is overlaid by the actual throughput realized with multiple cards. As the figure shows, the dropoff from the theoretical projections, which indicates limitations in scalability, is evident for higher MTU sizes (higher throughputs). Performance limits on S/390 hardware and software subsystem performance is indicated by the maximum throughputs realized for the 1500- and 9000-byte MTU scenarios. Therefore, the OSA-Express adapter provides real value in the S/390 configuration.

Additional Functions

IP Assist

IP Assist (IPA) was a new architecture introduced as part of the QDIO work effort in OS/390 V2R7. IPA architecture is designed to provide support for further distribution of function between CS and OSA-Express. The major design point for IPA is to offload and distribute compute-intensive functions from CS to the OSA-Express adapter. IP Assist functions include:

ARP Offload: This reduces CS cycles used in managing ARP cache. OSA-Express responds to ARPs received from the network as well as issuing ARP requests on behalf of the CS TCP/IP Stack.

MAC Handling: Previous S/390 LAN gateway interfaces required the CS stack to build the entire LAN header prior to transmission to the gateway. OSA-Express constructs the appropriate MAC header on behalf of the TCP/IP stack.

IP Filtering: Prior to inclusion of this support, broadcasts were always passed to the stack regardless of the network protocol that issued the broadcast, for example, IPX, DecNet, and so on. OSA-Express will now filter out all unsupported protocols.

IP Addressing: Dynamic assignment, for example, the addition or deletion of IP addresses, provides base support for the virtual IP address (VIPA) takeover during recovery of failed applications. The stack where the failure occurred deletes the address, and the recovery stack adds the address — all without human intervention. This enables the VIPA of a failed application to "follow" the application to the recovery stack.

The aggregate effect of these functions is to further reduce the use of S/390 cycles by relocating the function onto OSA-Express. The net result is a more intelligent use of system resources, the enablement of additional application recovery facilities and a further reduction of S/390 network system cost.

Priority Queuing

The QDIO architecture used by the interface between Communications Server and OSA-Express provides a priority-queuing mechanism in which four write queues are provided to order outbound traffic. This mechanism works in conjunction with the CS-provided Policy Agent, which enables you to specify different Quality of Service (QoS) class-based TCP/IP standards. These can be, for example, existing Type of Service (ToS) as well as the newly defined Differentiated Services (DiffServ). Properly used, this allows you to specify higher-priority service for mission-critical traffic and lower priorities for other less-critical traffic. The effect of using the prioritization of traffic is shown in the following figure.

The figure shows the results of a joint IBM and Cisco solution called "Delivering Predictable Host Integration Services." This solution demonstrates the impact of QoS policy on response time for critical traffic. The figure shows the response times of high-priority traffic under two conditions:

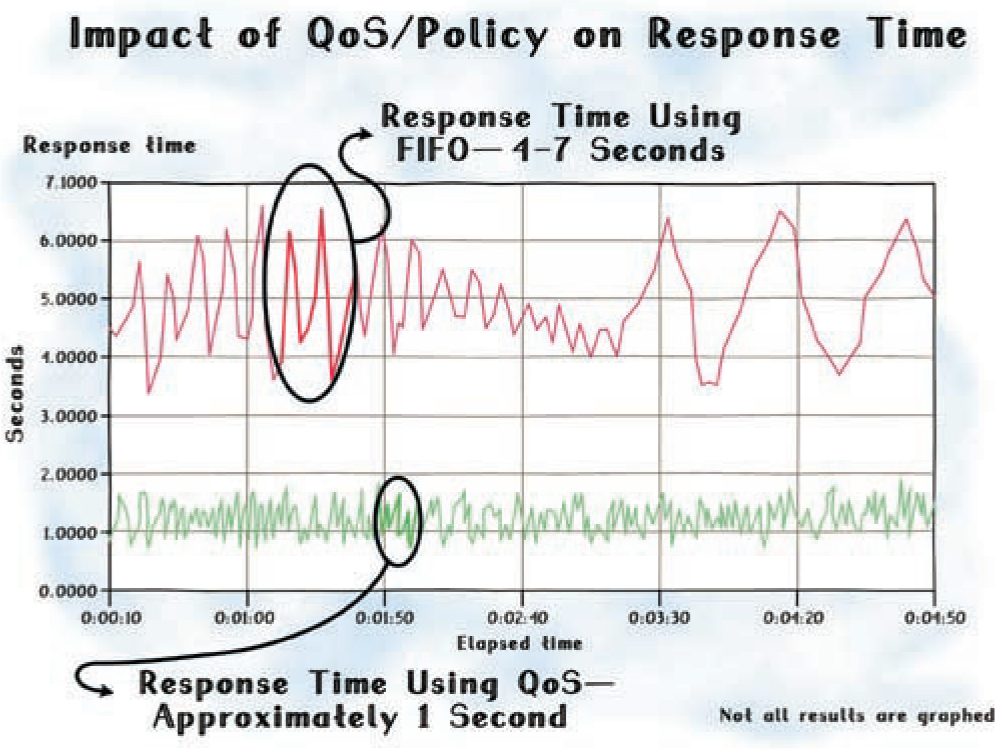
- Without exploiting QoS prioritization (upper graph)
- With QoS prioritization (lower graph)

In both cases, the network was first saturated with large FTP transfers and then high-priority traffic was inserted into the mix, in this case consisting of relatively small, interactive TN3270 transactions. The effect was dramatic. Response time went from 3 - 7 seconds to approximately 1 second. Not shown, but worth mentioning, is that there was no noticeable effect on the FTP traffic. Priority queuing in and of itself is a strong reason to move to the OSA-Express environment and is another reason why OSA-Express enables the best S/390 and z900 networking.

New for OS/390 R10

We continue to enhance this leading-edge interface, as amply demonstrated by the new additions in OS/390 R10:

- Extension of the QDIO interface to include Fast Ethernet[2] and ATM LANE (Ethernet mode)
- In conjunction with IBM middleware, the extension of QoS support to allow transactional QoS, for example, enabling priority to be set based on a particular item of work within a connection
- Enhanced management using a new interface that allows CS to extract ARP cache information



[2] Note that the Fast Ethernet and ATM LANE additions will be available on OS/390 R8 as a PTF.

New on zSeries 900

The OSA-Express adapter was enhanced again on the z900 with the addition of new packaging that provides two ports on each Gigabit Ethernet, Fast Ethernet, and 155 ATM feature. The Gigabit Ethernet adapter has been enhanced to attain line-speed performance. Protocol support, for example, QDIO (TCP/IP) and non-QDIO (TCP/IP and SNA/APPN®/HPR) support, remains the same as on S/390 G5/G6.

Bottom line is that, by every metric, OSA-Express, in conjunction with Communications Server, provides the best network connectivity for the S/390 and z900 servers.

For additional information on the OSA-Express adapters, the TCP/IP and SNA support provided by Communications Server for OS/390 and z/OS, joint IBM/Cisco partnership solutions, the most recent performance information, and latest news on z900 and S/390 networking, go to **ibm.com**/servers/eserver/zseries/networking. ■

About the Author

Arthur Stagg worked on various network-attachment efforts for the zSeries (S/390) for about 23 years. He worked in Research for 2 years on fully distributed, single-system image, server-clustering techniques. He has worked in the SNA, APPN, OSI and TCP/IP environments, dealing with the utilization of network media, including LANs (TR, ENET, GNET), WANs (ATM, X-25, ISDN), channel attachment (both copper and fiber), and the Open Systems Adapter.

Arthur was RTP inventor of the year in 1994 and named Master Inventor in 1997. He serves as co-chair for the Network Design Council sponsored by the Enterprise Sever Group (ESG) and as a member of the Network Project Development Team (PDT).
stagga@us.ibm.com

Enterprise Extender: A Key to SNA/IP Integration

By Sam Reynolds

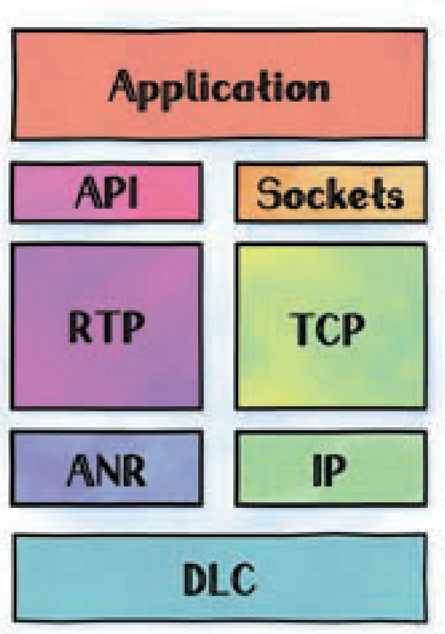
Motivation

With the latest releases of Communications Server for OS/390® and z/OS Communications Server software, the S/390 and the new IBM @server zSeries 900 servers are world-class platforms for native e-business (TCP/IP-based) applications. However, conversion of existing SNA applications to TCP/IP-enabled applications can be economically impractical. In many cases such conversions may even be technically impractical due to the lack of source code and adequate skills for the specific application. An additional complication is the wide variety of SNA-based endpoint devices, such as banking ATMs. So, how can we enable IP applications and preserve SNA-application and endpoint investment, while converging on a single network protocol?

For 3270-based applications, an inboard S/390 or z900-based TN3270 server can be a key component of the solution, allowing TN3270 clients to access SNA applications through an IP network, and limiting the SNA network path to the inside of a single Communications Server[1] image. This leaves the question of how to access non-3270-based applications without requiring a parallel SNA network path into the S/390 or z900 server.

Enterprise Extender (EE) enables e-business applications on the IBM S/390® and IBM @server zSeries 900 servers, while preserving the investment in SNA-based applications and endpoints and leveraging high-speed, industry-standard TCP/IP connectivity.

From Subarea SNA to Enterprise Extender



Systems Network Architecture (SNA) has evolved from the traditional subarea networks that have dominated the enterprise network landscape for years. The Advanced Peer-to-Peer Networking® (APPN®) extension was an enhancement to SNA that brought the ability to move logical units and change routing without coordinated system definition. High-Performance Routing (HPR) was an enhancement to APPN that enabled unparalleled availability by non-disruptively switching sessions around failures. Enterprise Extender is yet another evolution, providing a means for the efficient transport of SNA data across an IP network.

Enterprise Extender is an industry-standard solution defined by the APPN Implementer's Workshop (AIW) and the IETF (RFC 2353). With Enterprise Extender, the Rapid Transport Protocol (RTP) endpoint views its interface with the UDP layer of the stack as just another data link control, and treats the connection across the IP network the same as it would any SNA connection.

[1] In this article, *Communications Server* refers to Communications Server for OS/390 and z/OS Communications Server.

HPR's RTP component provides:

Error detection with selective retransmission of lost packets.

Non-disruptive reroute based on class of service requirements. HPR preserves the session without impact to the end user for planned and unplanned outages in the session path. If no alternative path is available, HPR can even be configured to preserve the session while the failing component is recovered.

Proactive congestion control. Enterprise Extender brings with it an enhanced version of HPR's Adaptive Rate-Based (ARB) congestion-control algorithm. The new version, Responsive-Mode ARB, is more aggressive in using available bandwidth and more tolerant of variations in network latency. Responsive-Mode ARB was introduced with Enterprise Extender to better allow HPR traffic to coexist with native IP traffic in the backbone network.

Prioritization. The SNA priority field is mapped to the IP Type of Service (ToS) byte, which is used by routing algorithms such as the Cisco Weight-Fair-Queueing algorithm. A set of standard UDP ports are also reserved based on priority, with packets mapped to them according to the SNA priority field. Furthermore, the Service Policy Agent (available on Communications Server for OS/390 V2R7 and later) enables priority schemes to be further refined, allowing for options such as setting the TOS priority based on the time of day or the specific client IP address.

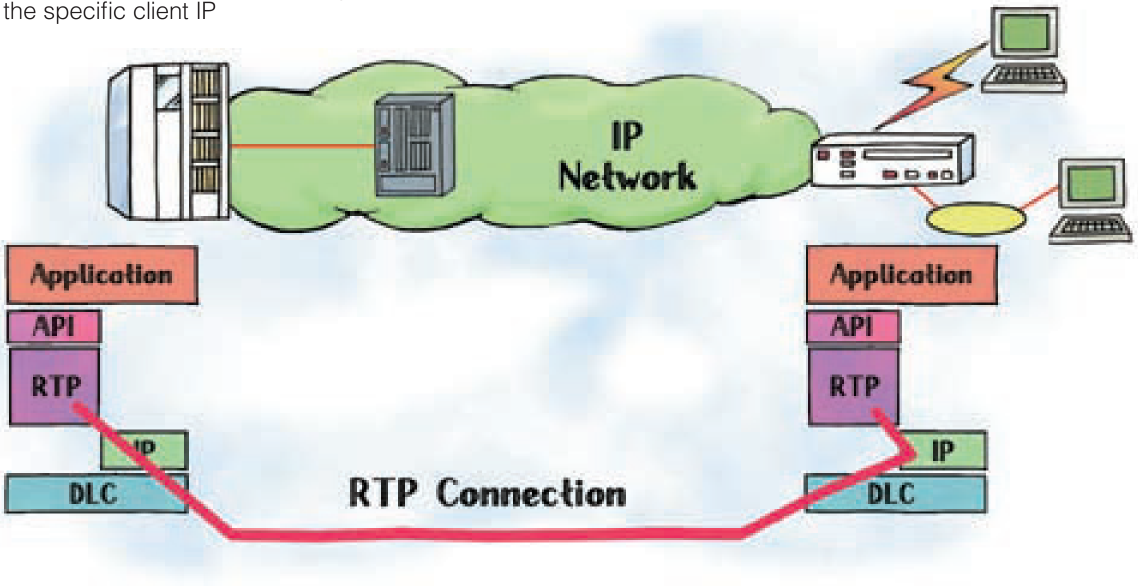
The IP layer handles packet forwarding for Enterprise Extender, providing the following advantages:

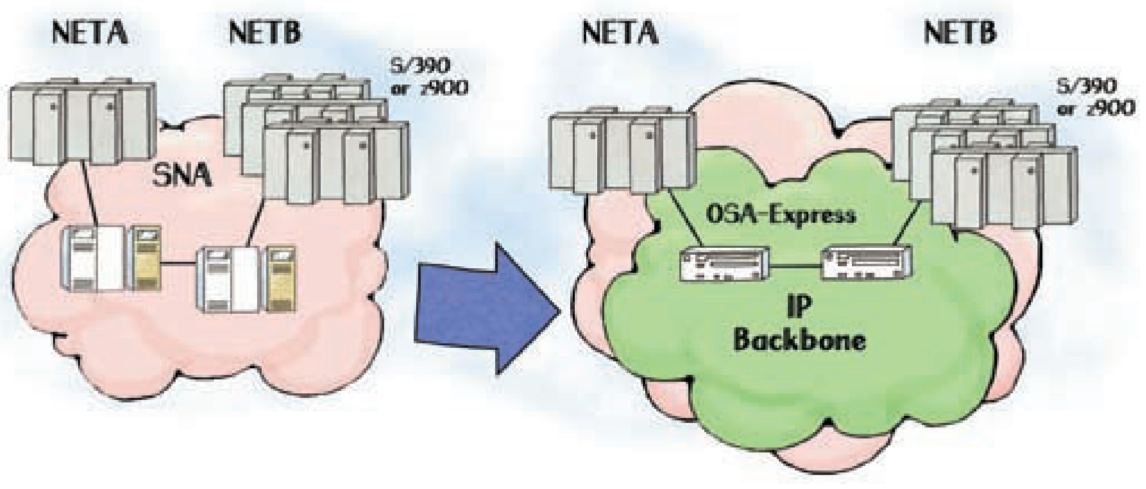
- The use of native IP routing maximizes router efficiency.
- By using Enterprise Extender, SNA applications are positioned to take full advantage of advances in IP routing technology.
- Enablement of a single network transport reduces costs and simplifies network management and network architecture.

Enterprise Extender versus Data Link Switching

Another commonly used SNA/IP integration technology is data link switching (DLSw). This technology was conceived by IBM and is widely used for transporting SNA data over IP networks. While DLSw is a well-established technology, it has a number of disadvantages when compared with Enterprise Extender:

- DLSw requires significant processing in routers acting as the DLSw endpoints, especially those used to terminate many DLSw connections in the data center. By using Enterprise Extender on S/390 or z900 servers, data-center router requirements are reduced, and can be eliminated by using OSA-Express.
- The data-center router serving as the DLSw endpoint provides an additional single point of failure. Loss of that router will cause session outages. With the Enterprise Extender endpoint in Communications Server, IP routing will reroute around a failed data-center router.
- DLSw solutions typically leave the SNA "boundary function" in the 3745, or use ESCON®-attached routers or OSA for SNA connectivity into the S/390 server (pushing the boundary function into VTAM®). Enterprise Extender solutions typically put the boundary function on the branch routers by using the APPN dependent LU requester (DLUR).
- DLSw implementations have various degrees of support for SNA session priority. Enterprise Extender solutions preserve session priority by mapping the SNA transmission priority field (TPF) to one of a set of reserved UDP ports, and setting the ToS byte based on the TPF.
- Although DLSw is an industry standard, many variations (often proprietary) exist. As noted above, Enterprise Extender is an AIW- and IETF-approved standard.





Enterprise Extender provides the following additional advantages:

- It provides superior scaling characteristics in the data center, with each Communications Server image capable of supporting thousands of Enterprise Extender connections.
- It enables the APPN/HPR high-availability functions of Generic Resources and multinode persistent sessions within the Parallel Sysplex® environment.

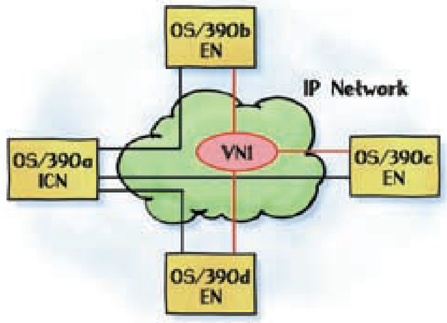
Enterprise Extender: Multiplatform, Multivendor, Industry Standard

As previously stated, Enterprise Extender is an industry-standard solution defined by the APPN Implementer's Workshop and the IETF. It has been available on IBM's Communications Server for OS/390 product since V2R7 (V2R6 via PTF), and is supported on a number of other products, including Communications Server for OS/2 WARP®, Communications Server for Windows NT®, and Cisco's SNASw feature for IOS. The multiplatform, multivendor support for Enterprise Extender was underscored by the successful conclusion of a joint IBM/Cisco scalability and interoperability test.

Connection Network

Connection network is an APPN mechanism by which a shared-access transport facility (SATF) is represented as a virtual routing node (VRN). APPN nodes connected to the SATF define a connection to the VRN instead of defining any-to-any connectivity to all nodes using the SATF. Much definition work can be avoided because each node replaces n-1 partner link definitions with a single definition of VRN connectivity.

Enterprise Extender exploits the connection network mechanism, allowing a VRN to represent the IP backbone. Connections will be dynamically activated as needed across the connection network.



Enterprise Extender for Inter-Enterprise Connectivity

Enterprise Extender provides an ideal migration tool to enable an alternative inter-enterprise connectivity path for existing users of SNA Network Interconnect (SNI).

The APPN replacement for SNI is Extended Border Node (EBN), a proven technology first shipped on VTAM in 1994. EBN is now used by numerous customers to facilitate inter-enterprise communication, and to ease network consolidation after mergers and takeovers. Unlike with SNI, which requires a Gateway NCP to act as the network boundary, the boundary between two EBNs is represented by the intersubnet link (ISL) itself. Therefore, the ISL is not limited to only NCP-based connections, but instead allows other options such as MPC+ and Enterprise Extender.

Because most enterprises today connect to an IP network (the Internet, intranet, or an extranet), Enterprise Extender emerges as an ideal way to connect multiple enterprises by using the existing IP connectivity. The two S/390 or z900 hosts that formerly acted as gateway system services control points (SSCPs) and ran the SNI protocol over NCP-based connectivity now act as APPN EBNs with an Enterprise Extender connection mapped over the IP connectivity. Thus, gateway SSCP/NCP functions are no longer required, and SNA applications can achieve inter-enterprise communication through any IP network attachment and connectivity supported by Communications Server.

With Enterprise Extender, the Rapid Transport Protocol (RTP) endpoint views its interface with the UDP layer of the stack as just another data link control, and treats the connection across the IP network the same as it would any SNA connection.

Enterprise Extender Enables High-Speed, Industry-Standard Host Connectivity

OSA-Express provides an economical, high-speed method for host access. The adapter provides access to industry-standard network attachment options using a direct memory access (DMA) model that utilizes a set of priority queues shared between the adapter and Communications Server's TCP/IP stack. Communications Server accesses OSA-Express through the Queued Direct I/O (QDIO) interface, which provides higher bandwidth, lower latency, and reduced CPU consumption.

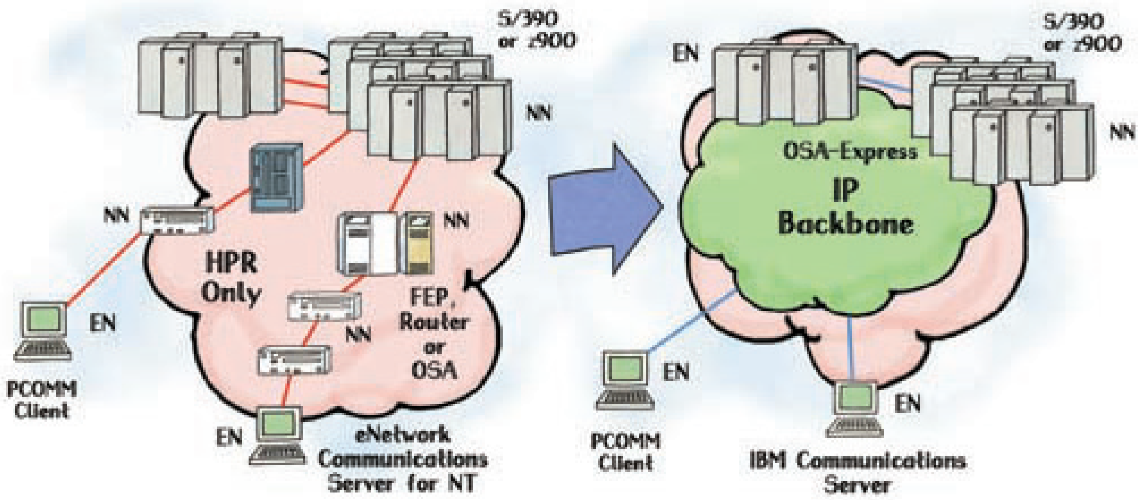
The QDIO interface is efficient because it treats the OSA-Express adapter as a logical extension of the TCP/IP stack, thereby allowing for an intelligent division of workload between the stack and the adapter. A consequence of this is that the adapter expects to receive only IP packets across the QDIO interface, and does not support native SNA communication when running in QDIO mode. However, while there is no support for native SNA with the QDIO interface, SNA applications can still utilize this high-speed path by using Enterprise Extender.

Combining the use of Enterprise Extender with OSA-Express provides an ideal way to preserve existing SNA applications, while using industry-standard network attachments with performance characteristics that exceed those of native HPR attachment. Furthermore, by relying on the IP layer for the underlying transport, the Enterprise Extender/OSA-Express combination positions existing SNA applications to take advantage of future advances in IP routing technology and protects customers' SNA device and endpoint investments.

Enterprise Extender and OSA-Express Simplify SNA Network Design

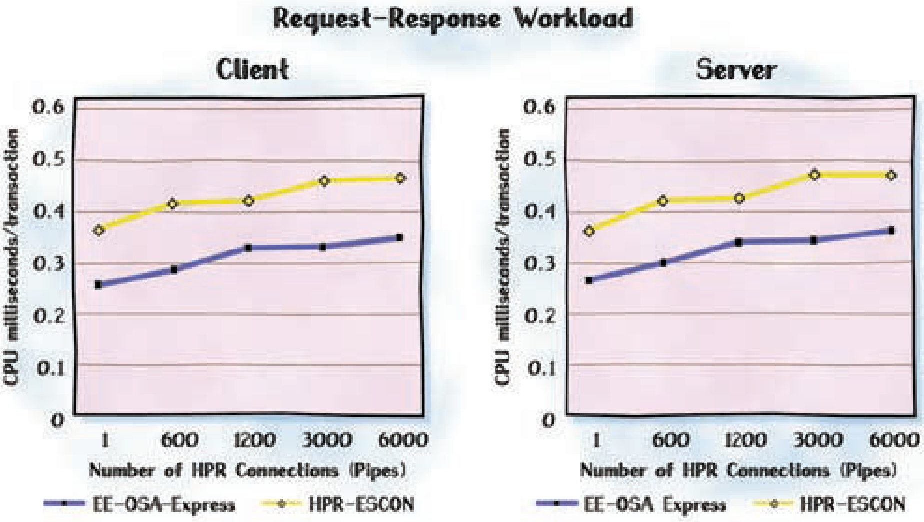
When APPN was first introduced, many people envisioned relatively complex networks with large numbers of network nodes. Now, with Enterprise Extender and a pervasive IP WAN, it is possible to design much simpler APPN networks. In the data center, a small number of Communications Server images would be designated as network nodes, typically the Communications Management Controller (CMC), the backup CMC, and a couple of images per sysplex to provide basic network node services within each sysplex. The remainder of the Communications Server images would be data hosts defined as end nodes. Branch routers such as Cisco's SNASw routers would be configured as Branch Extender (BX) nodes. A BX node presents an end node image to the upstream hosts in the data center, but can provide network services for end nodes and clients in the branch.

This results in a simplified and highly scalable design that greatly reduces the impact of APPN topology and search broadcast traffic, while fully enabling APPN/HPR access into the Parallel Sysplex environment, and allowing the exploitation of Generic Resources and multinode persistent sessions.



Scalability

Despite some early predictions to the contrary, HPR has been shown to be a very scalable architecture. The following graphs were generated from a performance test intended to show the reductions in CPU utilization that can be achieved when moving from native HPR over ESCON to Enterprise Extender over an OSA-Express Gigabit Ethernet. However, equally interesting is the scalability statement that these graphs make. The measurements were taken while running up to 6000 RTP connections (pipes) from a single S/390 host. (Subsequent tests have expanded that number to 10 000 pipes.) The slope of the curves between the 3000 and 6000 pipe data points is relatively low, indicating a highly scalable function.



CS for OS/390 V2R10 Enhancements

Despite the advantages of APPN and HPR, many HPR migrations have been delayed or inhibited by a restriction that has existed since the earliest releases of HPR on Communications Server for OS/390. Prior to V2R10, an interchange node (ICN) would not allow HPR on the first hop from the ICN for interchange sessions (those sessions from a subarea partner crossing into the APPN network). Sessions from applications on the ICN could enter the APPN network via HPR, but those entering the ICN from a cross-domain subarea partner (such as an SNI partner) could not use HPR for the first hop from the ICN into the APPN network.

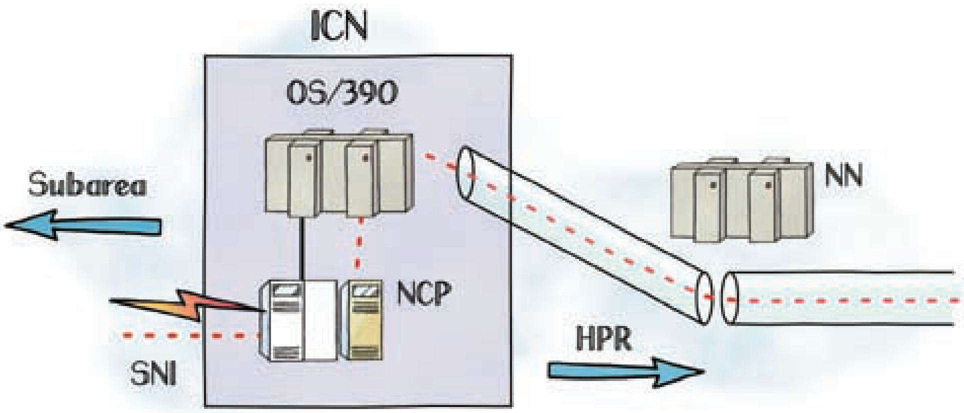
If all connection types allowed both HPR and Intermediate Session Routing (ISR) — the basic pre-HPR routing scheme — then this restriction would be little more than an irritation. However, some connection types, including Enterprise Extender and native ATM, require HPR and do not have underlying ISR support. Therefore, attempting to set up interchange sessions over these connections will fail.

There is one exception to this restriction. If the interchange session is going from the ICN to an end node through a single APPN hop, then HPR is allowed. This slight bit of relief from the restriction can be quite helpful, especially in sysplex configurations. Nonetheless, given the difficulty of completely eliminating subarea partners (especially SNI partners), this restriction has made it difficult (and sometimes impossible) to completely migrate to Enterprise Extender on the S/390 server.

Communications Server provides relief from this restriction. When an ICN, at the V2R10 level, receives a session initiation request for an interchange session, it examines the APPN session path to determine if HPR should be used for the ICN's entry hop into APPN. If so, the ICN ties the session to a one-hop pipe to the adjacent APPN partner node. Sessions for multi-hop APPN paths will use the one-hop pipe to the adjacent node, and then access a second HPR pipe for the remainder of the HPR-capable session path.

Of course, there are disadvantages to this approach. By setting up back-to-back pipes in the adjacent node, we introduce a small performance impact in the adjacent node for serving as the endpoint for the two HPR pipes. Furthermore, the adjacent node becomes an additional single point of failure on the session path. However, the benefit is clear: We can now establish interchange sessions using HPR, thereby enabling all connection types (including Enterprise Extender and native ATM) anywhere within the APPN network. Note also that the above disadvantages apply only to interchange sessions. For sessions originating on the ICN (or within its domain), or passing through the ICN as an intermediate node on an APPN path, the back-to-back pipe mechanism is not employed.

Due to the value of this function in enabling customer migrations to Enterprise Extender, Communications Server for OS/390 has also provided this function for prior releases (V2R6 - V2R8) by APAR OW44611.



There is still one remaining limitation for interchange sessions. While V2R10 now allows interchange sessions to enter the APPN network through HPR-capable connections, those connections cannot be over a connection network (VRN). The removal of this final restriction is being studied for a future release. In the meantime, the following recommendations apply if you deploy Enterprise Extender with a connection network:

- Define the Enterprise Extender VRN at all end nodes, migration data hosts (MDHs), branch extenders (BXs), and pure network nodes, but not at the ICNs.
- Use only defined (non-VRN) connections at the ICNs.

In most situations, the above recommendations do not imply a great deal of definition beyond what is already required. The ICNs are typically also functioning as network node servers (primary and backup) for the end nodes and branch extenders, and there must be defined connectivity between an end node (or a BX) and its network node server for the CP-CP session.

Summary and Future Directions

Enterprise Extender provides a highly scalable and reliable component for SNA/IP integration strategies. It allows you to preserve your SNA application and device investment, while maintaining the session prioritization and availability characteristics of SNA and HPR. Furthermore, Enterprise Extender allows for a simplified network architecture that positions applications for exploitation of future advances in IP networking technologies.

Enhancements to Enterprise Extender (and APPN/HPR in general) will continue in future releases of z/OS. Enhancements currently being planned include additional diagnostic tools as well as a function to further improve the usability and performance of Enterprise Extender as an inter-enterprise connectivity option. ■

About the Author

Sam Reynolds has been associated with the design and development of host networking products since joining IBM in 1990. He is currently a designer for z/OS Communications Server, a product that provides TCP/IP and SNA connectivity.
samr@us.ibm.com

The Rebirth of TCP/IP on OS/390

by Alfred Christensen

When TCP/IP for MVS was initially ported from VM, the main focus was IP network access and base server and client functions, including the ability to write your own socket programs on MVS. There was less demand for characteristics such as availability, scalability, performance, and reliability.

As the commercial importance of TCP/IP on MVS and OS/390 grew during recent years, so has the customer demand for more functions, improved availability and reliability, better performance, and the ability to scale socket workload across multiple processors and even OS/390 images in a Parallel Sysplex® environment. After evaluating the existing product back in 1996, it became clear that a new product would be required to satisfy these customer requirements.

During the development of the new TCP/IP product for OS/390, we wanted to make sure that all existing external functions and APIs remained as intact as possible. We removed only one external API, which was the low-level assembler API directly on top of the IUCV/VMCF subsystem. We decided to remove this API after extensive research into who used it at that time. Software vendors as well as MVS TCP/IP users were notified three years in advance that the API would disappear.

Since OS/390® V2R5, a new TCP/IP component has been included in the Communications Server for OS/390 as a complete replacement for the old TCP/IP for MVS™ V3 product.

Traditional TCP/IP V3R2 C-socket programs, sockets extended programs (callable or assembler macro), ONC RPC, XTI, X Windows, interpreted REXX socket programs, or Pascal programs run unchanged on OS/390 V2R5 or later. Application developers do not need to recode them, they do not need to change any external configuration, such as TCPIP.DATA or other resolver-type configuration files they have created, and they do not need to learn about OS/390 UNIX® to make them run with the new TCP/IP product. Depending on your current level of TCP/IP for MVS, some programs can be re-linked to achieve optimal performance, but that's it.

Existing applications will work on OS/390 V2R5 and later with one exception — applications written using the assembler IUCV API must be converted to the sockets extended assembler macro API or the UNIX assembler callable services API.

Performance and Scalabillity

In designing the new TCP/IP for OS/390, we made performance and scalability our top priorities. Additionally, a major part of the OS/390 program's future would be based on enabling popular UNIX applications — such as Web services, DCE-based services, and directory services — to run on OS/390. While it would have been possible to rewrite all of these applications to a new API, we decided not to do that because a POSIX-compliant API already exists on OS/390. So we decided to take advantage of UNIX sockets for these applications. However, to meet the performance criteria for the new TCP/IP product, it was obvious that the code path between the OS/390 UNIX applications and the TCP/IP stack had to be minimized, and therefore the UNIX System Services and TCP/IP stack became much more integrated than before.

Although the new TCP/IP uses some OS/390 UNIX functions that require you to do some configuration of the OS/390 UNIX element, you do not need to be a UNIX expert to run the new stack.

After the stack is initialized, you can move your existing MVS-based sockets workload to the new TCP/IP on your OS/390 V2R5 or later system without concern for UNIX. If your users require some of the new OS/390 server functions, you may have to learn more about the OS/390 UNIX environment. Servers such as the Dynamic Host Configuration Protocol (DHCP) server or the Trivial File Transfer Protocol (TFTP) server are ported UNIX servers, and they run in the OS/390 UNIX environment and behave somewhat as other UNIX programs do.

The whole idea with OS/390 UNIX is to allow programs that are written to the open standards as defined by POSIX and XPG4 to run on OS/390 side-by-side with traditional MVS programs.

Many Requirements Met by Porting Existing Code

The whole idea with OS/390 UNIX is to allow programs that are written to the open standards as defined by POSIX and XPG4 to run on OS/390 side-by-side with traditional MVS programs. Many of the requirements for new functions on OS/390 can be met by porting existing code that is developed according to those standards. That saves all of us time and money. Instead of developing a new DNS server for OS/390 from scratch, we were able to take the standard DNS/Bind source code and, in a relatively short time, make it run as a standard DNS server on OS/390. Same story with standard components such as Sendmail, the TFTP server, and the DHCP server. If you do not use any of these UNIX-style servers, you really do not need that much knowledge about the OS/390 UNIX environment.

For a straight migration from a standard TCP/IP for MVS V3R1 or V3R2 level, you wouldn't need any of these new servers. Then, as you move ahead and want to exploit functions, you will have to learn some more about the OS/390 UNIX environment, but you set the pace.

In some situations, TCP/IP will be the first IBM subsystem that requires that you familiarize yourself with OS/390 UNIX. There is no doubt that you will have to do so at some point during the next few years. Initially, TCP/IP for OS/390 V2R5 requires learning something new, but after acquiring that education, you will be able to offer your users a completely new set of OS/390-based services, often at a price that is significantly lower than what they currently pay for having the same type of services running on various UNIX boxes. ■

About the Author

Alfred Christensen is a senior designer in the z/OS Communications Server strategies and design group in Raleigh. He has been working with mainframe software for 25 years and with TCP/IP on OS/390 and z/OS for the past 10 years.
alfredch@us.ibm.com

OS/390 and z/OS TCP/IP in the Parallel Sysplex Environment – Blurring the Boundaries

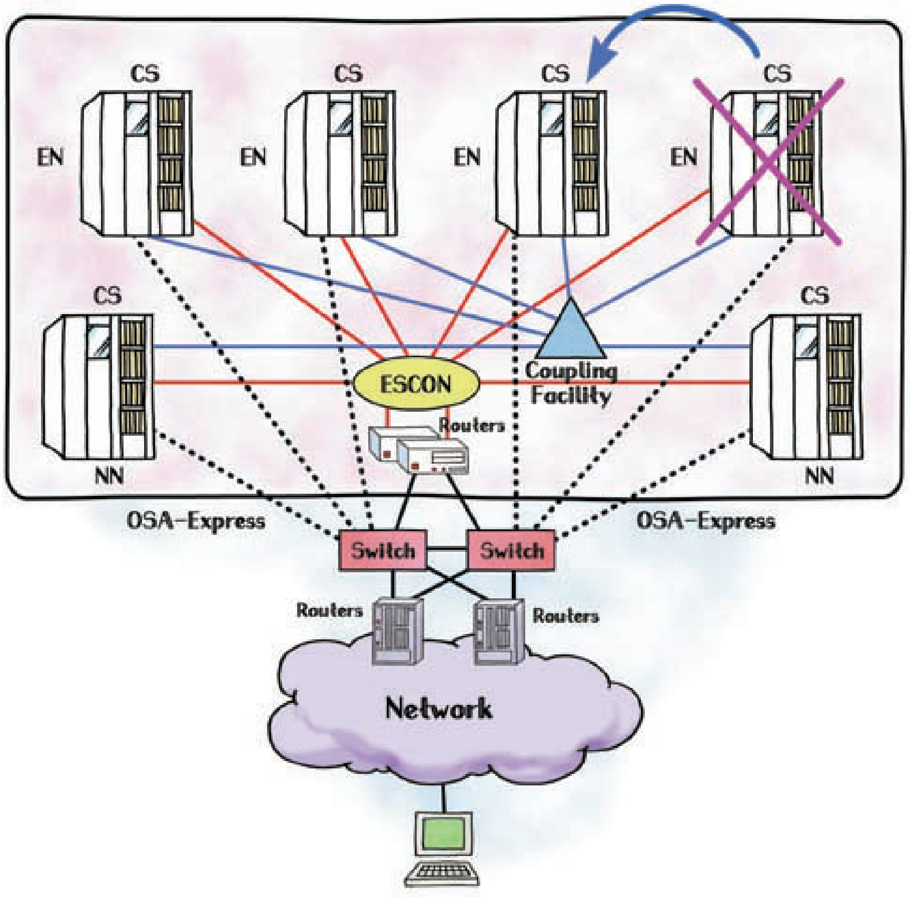
By Jay Aiken

Since OS/390® TCP/IP was released in March 1998, IBM has made continuing improvements in TCP/IP to exploit Parallel Sysplex® technology, evolving toward a true cluster IP server and distributed application platform.

With OS/390 V2R5, TCP/IP for MVS™ became an integral part of the OS/390 program, IBM's flagship operating system for the S/390® Parallel Sysplex implementation, and the premier clustering solution in the industry.

At that time, OS/390 TCP/IP capabilities and functions for the most part were analogous to TCP/IP functions. There was one difference, though: OS/390 TCP/IP runs on the S/390® server, the IBM server platform for Parallel Sysplex clustering that delivers industry-leading availability and scalability.

One of the aspects of both availability and scalability of a clustered solution such as a Parallel Sysplex cluster is that clients and the IP network should know as little as possible about the internal makeup of the cluster in terms of individual nodes and application instances. On the other hand, the IP way of assigning IP addresses to specific physical adapter links to some extent exposes the internal composition of the cluster to the network and clients. OS/390 TCP/IP, released in March 1998, provided much-improved scalability on a single multiprocessing system, but in other respects was similar to TCP/IP on other platforms. Since that time, IBM has made continuing improvements in TCP/IP to exploit Parallel Sysplex technology, evolving toward a true cluster IP server and distributed application platform.



Three Technologies Available

At the time of the OS/390 V2R5 TCP/IP availability, three technologies were available for use with S/390 Parallel Sysplex technology in support of clustered IP applications: Virtual IP Addresses (VIPAs), Workload Manager (WLM)-enabled Domain Name Server (DNS), and Network Dispatcher. The first two were delivered with OS/390 TCP/IP, while the last one executes on an outboard system (AIX® operating system or IBM 2216 Nways® Multiaccess Connector).

Virtual IP Addresses

VIPAs are IP addresses that are independent of any particular network interface. To an external router or other TCP/IP stack, a VIPA appears as simply an address (or a subnet) that is reachable through the hosting stack. When an OS/390 TCP/IP receives an IP packet whose destination is a VIPA that it supports, the packet is merely routed up the stack to the upper-layer protocol (TCP, UDP, or raw socket) as with any other IP packet to an address of a physical link hosted by the stack. VIPAs are advertised using static routes or dynamic routing protocols just as with any other IP address. The benefit of using a VIPA as an application address on an OS/390 TCP/IP with multiple physical links, however, is that a failure of any one link will not disrupt connectivity to the application. As long as there is a network path from a remote client to the TCP/IP hosting the VIPA and the server application, the client and the server application can interact without interruption. A VIPA thus provides independence from any particular adapter, but is still for the most part tied to a particular OS/390 TCP/IP stack.

With OS/390 V2R5, TCP/IP for MVS became an integral part of the OS/390 program, IBM's flagship operating system for the S/390 Parallel Sysplex implementation, and the premier clustering solution in the industry.

WLM-Enabled DNS

DNS/WLM, the WLM-enabled DNS, provides standard name-resolution services (converting a name into an IP address) for IP applications in the Parallel Sysplex cluster. OS/390 TCP/IP stacks register their IP addresses with WLM. Application instances register themselves under a generic application name (similar to SNA Generic Resources). The DNS/WLM associates the IP addresses with application names as appropriate.

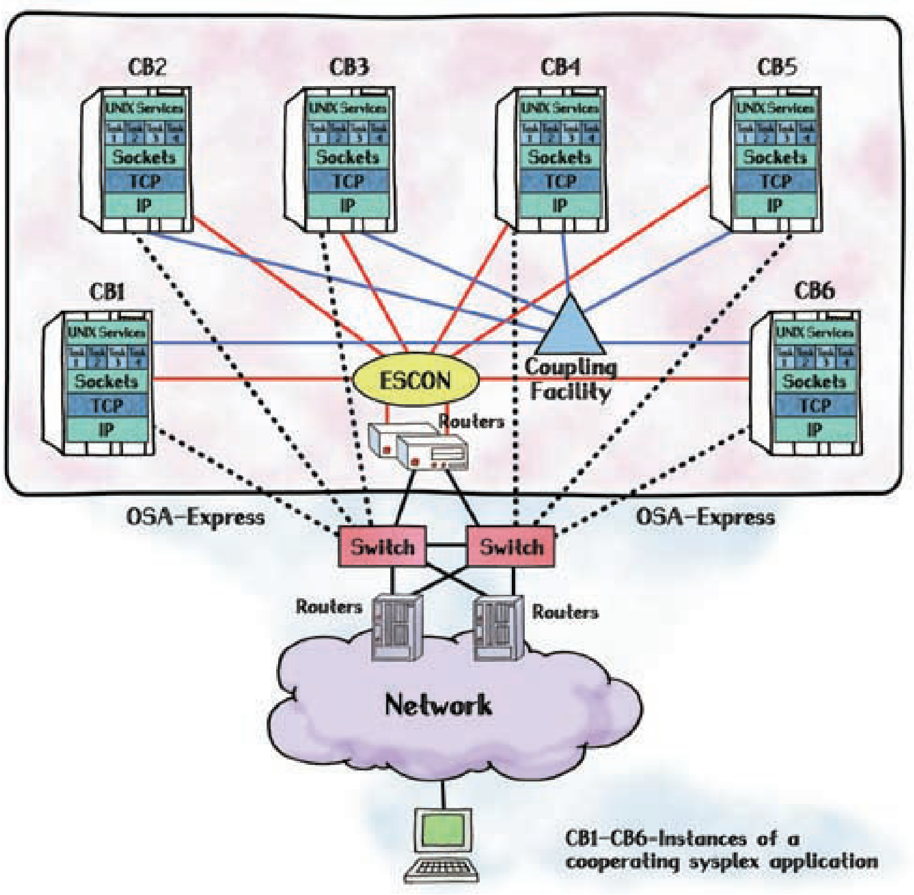
When WLM is being run in goal mode in the OS/390 images in the Parallel Sysplex cluster, and a resolution request for an application name arrives, DNS/WLM consults WLM information to determine to which TCP/IP the request should be routed to balance the workload appropriately.

If a client resolves the same name again, a different IP address may be returned if relative available capacity among the server instances has changed. If clients in the network resolve the name each time, and do not cache the IP address from a previous request, the workload will be balanced among the available server applications according to relative available capacity. Work can thus be spread across the cluster, and the Parallel Sysplex cluster presents the appearance of a single platform to well-behaved client applications.

Network Dispatcher

Network Dispatcher is also an approach to workload distribution across a set of application instances in a cluster, but it works with a "cluster IP address," rather than with name resolution. Network Dispatcher (and, for example, the similar function in the Cisco MultiNode Load Balancing [MNLB]) is an external entity adjacent to the Parallel Sysplex cluster, running on either a 2216 or the AIX operating system. An agent in the sysplex communicates workload capacities of the nodes hosting the application to the Network Dispatcher node. The Network Dispatcher advertises ownership of the cluster IP address (application IP address) to the routing network. OS/390 TCP/IP stacks that are hosting the application define the same address as a loop-back address, which is not advertised to the routing network. The Network Dispatcher must have a direct link (single IP hop) to each TCP/IP stack hosting the IP application. When a new TCP connection request arrives, Network Dispatcher consults the most recent capacity information received from the WLM agents, selects the appropriate TCP/IP for this request, and forwards the request over the direct link to the selected stack.

For selected applications such as Web serving, an application advisor function in the Network Dispatcher will periodically query the application to be sure that it is available and responding, to ensure that requests are routed only to functioning server applications. Subsequent traffic from the client to the server is routed through Network Dispatcher to the same TCP/IP, though return traffic from the server application to the client need not flow through the Network Dispatcher. Clients thus see a network presence represented by a single IP address, and the servers that make up the processing that backs up the cluster address are hidden from the clients.



TCP/IP for OS/390 V2R7 — Sysplex Awareness

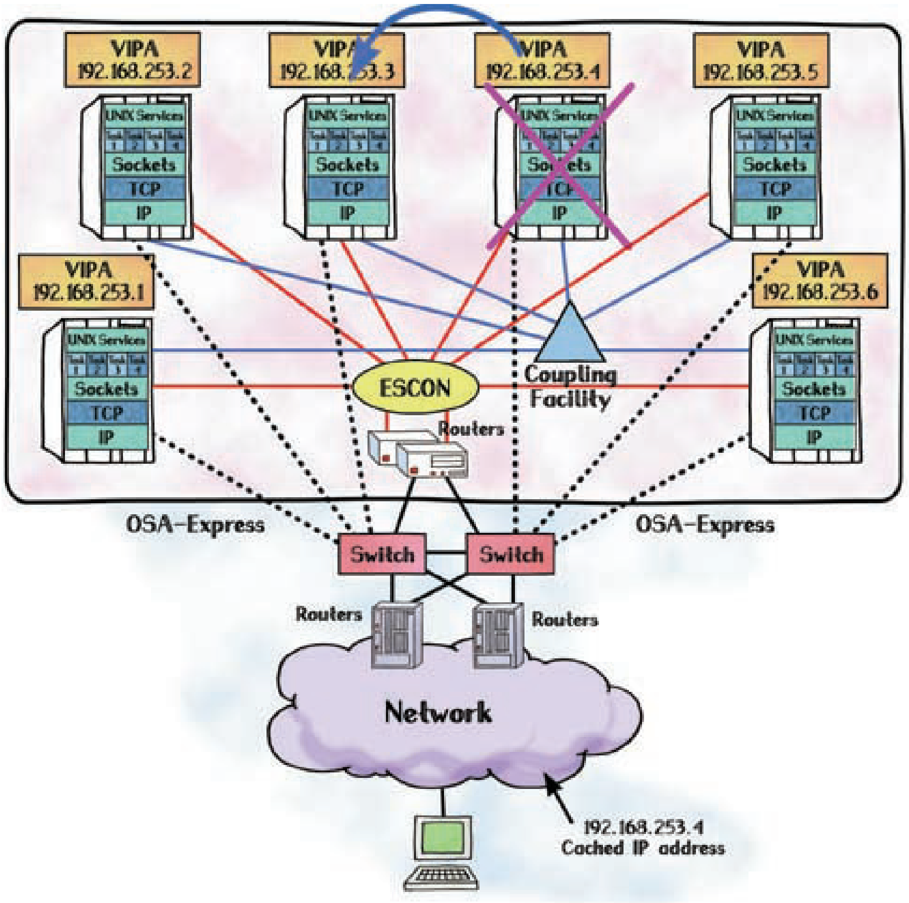
In OS/390 V2R7, TCP/IP began to exploit Parallel Sysplex functionality to become aware of other TCP/IPs in the sysplex, and to exchange information with these other stacks through MVS XCF Messaging, a basic service of OS/390 that is independent of network protocols. Functions provided by OS/390 TCP/IP as a part of this initial sysplex awareness include MVS System Symbols, Sysplex Sockets, and Dynamic XCF IP connectivity.

MVS System Symbols

MVS System Symbols allows a single configuration profile to be used for more than one TCP/IP. MVS System Symbols used within PROFILE.TCPIP is automatically resolved at TCP/IP initialization or during VARY OBEY processing. Items that are unique to each TCP/IP, such as device and link definitions, can be segregated into separate included files. This reduces the administrative workload for managing multiple TCP/IPs in a sysplex. Even other configuration files such as TCP.DATA that are read and used by applications can now also be managed as a single entity. A “source” file is maintained for the sysplex, and changes are made only to that file. A supplied utility can be executed on each OS/390 image (with unique definitions for the included symbols) that reads the “source” file and produces an output file with symbols resolved appropriately for the OS/390 image. While this is more cumbersome than with PROFILE.TCPIP, the process can be automated in a similar way to program compilation. Other than being an EBCDIC text file, there are no particular requirements on the “source” file, and both source file and the image-specific output files can reside in the Hierarchical File System (HFS), an MVS data set, or a member of a partitioned data set.

Sysplex Sockets

Sysplex Sockets provides a way for collaborating applications in an OS/390 Parallel Sysplex cluster to determine that both partners reside in the sysplex, to allow for programmed optimizations that might not be possible when the partner application resides elsewhere in the network. For example, information exchanged need not be converted to and from a common format, but can be exchanged in native S/390 format. Similarly, when the traffic flows over a link that is protected with the same physical security as the data center containing the sysplex, such as ESCON®, CTC, or XCF IP links, additional security such as Secure Sockest Layer (SSL) or other encryption might not be necessary, in which case the overhead can be avoided. TCP/IP stacks in the sysplex use MVS XCF messaging to exchange their supported IP addresses, so each TCP/IP in the sysplex is aware of all IP addresses active in the sysplex, and can therefore tell when the TCP partner application is using an IP address supported by one of the OS/390 TCP/IP stacks in the sysplex. As new TCP/IP stacks are added with new OS/390 images (or in existing images), information is exchanged with all other TCP/IPs. The information is updated whenever an IP address is deleted from or added to a TCP/IP in the sysplex via VARY OBEY, and the sysplex TCP/IPs are also notified when a TCP/IP is halted or suffers an outage, so the respective IP addresses may be removed from their tables of sysplex IP addresses.



XCF has been available as an IP transport between TCP/IP stacks in the sysplex since OS/390 V2R5 (and earlier in TCP/IP V3R2). *Dynamic XCF* builds on both the automatic XCF connectivity of VTAM® and the use of MVS XCF messaging by TCP/IP to define XCF IP links between TCP/IPs automatically, removing the need for manual definition and updates as new TCP/IPs are added to the sysplex. A single configuration statement (IPCONFIG DYNAMICXCF) on each TCP/IP is sufficient to enable this function. Whenever a new OS/390 TCP/IP is added to the sysplex, information exchanged through MVS XCF messaging allows the existing TCP/IPs to discover this new stack. Appropriate DEVICE, LINK, HOME, and BSDROUTINGPARMS statements are then created and activated automatically — with no additional manual definition to the existing TCP/IPs.

If dynamic routing protocols such as RIP or OSPF are in use in the sysplex and the attached routing network, a new application host can be added to the sysplex without external physical connectivity (other than the coupling facility links or ESCON links necessary for sysplex MVS XCF messaging), and IP applications can be reached from clients by DNS name resolution to the Dynamic XCF IP address on the new TCP/IP stack. Seamless horizontal growth of the sysplex is thus made much easier for IP applications, especially when combined with DNS/WLM as already described, or with technologies soon to follow.

TCP/IP for OS/390 V2R8 — Dynamic Virtual IP Addresses

Because VIPAs are not associated with any physical link, there is no particular reason why a VIPA should be associated with one, and only one, TCP/IP in the sysplex, other than the normal IP requirement that only one stack can advertise ownership of a given IP address to the attached routing network. When a TCP/IP or its hosting OS/390 suffers an outage (for example, a power failure), access via the physical links and their associated IP addresses is of course lost. However, a VIPA may be moved to another TCP/IP manually or via automation, and dynamic routing protocols can be used to notify the routing network of the new location of the VIPA without additional manual configuration change in the routers. In OS/390 V2R8, the concept of a *Dynamic VIPA* was introduced. Simplified configuration definitions allow Dynamic VIPAs to be activated either continuously or on demand from an application.

Other TCP/IPs in the sysplex can also be configured as backup for a continuously active Dynamic VIPA, such that the Dynamic VIPA is automatically activated on a backup stack whenever the normally owning TCP/IP suffers an outage. This automatic Dynamic VIPA backup is called *VIPA Takeover.* VIPA Takeover is applicable when multiple application instances exist, each of which can respond appropriately to any client request. However, some applications need to associate a particular IP address (or VIPA) with a particular application instance, such that client requests to that IP address always go to the same application instance. OS/390 V2R8 supports this as well, with application-initiated Dynamic VIPAs.

VIPA Takeover allows multiple identical application instances to be deployed in different sysplex nodes, each with a unique IP address, while preserving the appearance of almost continuous availability to the clients in the event of an outage on one of the nodes. Web serving is an example of such an application, where multiple instances of a Web server are distributed among sysplex nodes for availability and scalability. Each instance serves the same set of static and dynamic Web pages, so it really does not matter to the client which server handles a particular request. (Client requests can be distributed among the server instances by DNS/WLM, for example.)

Each server instance accepts requests to any local IP address (binding to INADDR_ANY, for the technically-inclined). Each TCP/IP serving such an application is configured as the primary owner of a Dynamic VIPA with a simple configuration statement (VIPADEFINE). This single statement causes activation of a virtual device and link, and adds the IP address to the HOME list for the stack. Other stacks are configured as backup for that VIPA with a VIPABACKUP configuration statement, in addition to having their own respective Dynamic VIPAs activated by VIPADEFINE. When an outage occurs at the server side, one of the backup stacks will automatically activate (take over) the Dynamic VIPA from the stack that suffered the outage. The active client connections will be disrupted, but the normal client recovery action is to attempt to reconnect to the same IP address, so the client will very quickly be able to establish a new connection to the backup stack.

As an added benefit, if the backup stack receives client traffic for a connection to the stack that is no longer available, the backup stack will immediately notify the client that the connection has been terminated, and the client will not have to wait for normal TCP timeouts to discover the connection outage. (Performance studies have shown up to 60% faster TN3270 failure detection and session reestablishment with this mechanism.)

Application-Initiated Dynamic VIPAs

Application-Initiated Dynamic VIPAs were provided in OS/390 V2R8 to address a different set of applications. Some applications define a relationship between client and server that requires the client to reconnect to the same server instance, which means that the IP address of the server instance must remain constant, and be different from the IP address of other server instances. When such an application suffers a failure, or the underlying OS/390 suffers an outage, automation such as OS/390 Automatic Restart Manager may in fact restart the application on a different OS/390 image, depending on available capacity and the restart policy defined for the application. In such a scenario, the IP address must move with the application, and this can be accomplished only with a VIPA.

TCP/IP supports this by allowing the application to bind to an IP address that is not currently active on the TCP/IP, or on any other TCP/IP in the sysplex. When the application binds to a nonexistent IP address, the stack will automatically activate a Dynamic VIPA with that IP address. A configuration statement (VIPARANGE) is provided to ensure that a misconfigured application will not be able to activate an IP address that is not appropriate for the installation. As long as all stacks where the application can be started have an appropriate VIPARANGE configuration statement, the application can be restarted on any of the TCP/IPs in the event of a failure, and the Dynamic VIPA will be activated when the application is initialized and issues the bind.

In some cases of this class of application, particularly with purchased, off-the-shelf products, the application cannot be configured to bind to a specific IP address. OS/390 V2R8 provides a utility, which can be added to the JCL, or OMVS shell script that starts the application. The utility issues a command to the stack to activate the Dynamic VIPA (subject to the same VIPARANGE considerations). As long as the application is always started with the same JCL package or shell script — even in automated restart scenarios — the VIPA will be activated on the TCP/IP that will be hosting the application instance.

In terms of presenting a single-node appearance of a clustered IP platform to the network, Dynamic VIPAs are a step forward, because the application address is no longer tied to a single processing node (OS/390 image) in the sysplex. However, when multiple application server instances are deployed for availability and scalability, each instance still has its own IP address, which means that the composition of the cluster still shows through to the client population to some degree. Also, moving an active application server instance from one OS/390 node to another within the sysplex is still disruptive, in that either client access to a server must first be quiesced, or active client connections to the server at the time of the move will be terminated.

Because Dynamic VIPAs are considered application addresses, rather than stack addresses, they are not automatically reported to DNS/WLM. If they were, then DNS/WLM might return a Dynamic VIPA to clients for other applications such as Telnet or FTP, which would greatly complicate the problem of relocating the Dynamic VIPA automatically, requiring coordinated movement of a disparate server population. Applications assigned to use Dynamic VIPAs should be configured statically in DNS/WLM.

TCP/IP for OS/390 V2R10 and z/OS — Nondisruptive VIPA Movement, Sysplex Distributor, and Server Bind Control

One of the limitations of VIPA Takeover in OS/390 V2R8 is that when the TCP/IP that normally hosts the VIPA is restored after a failure, the VIPA cannot be moved back to the normal stack immediately without disruption to application connections to that VIPA on the backup stack. Because new client connections can be established to the backup stack while existing connections are being serviced, it may be a very long time before there are no active connections to the VIPA on the backup stack — if ever. OS/390 V2R10 and z/OS provide the ability to move a Dynamic VIPA back to the normal hosting TCP/IP immediately, without disrupting connections to applications on the backup stack. New connection requests are serviced by applications on the normal hosting stack, while IP traffic for existing backup stack connections continues to be routed to the backup stack.

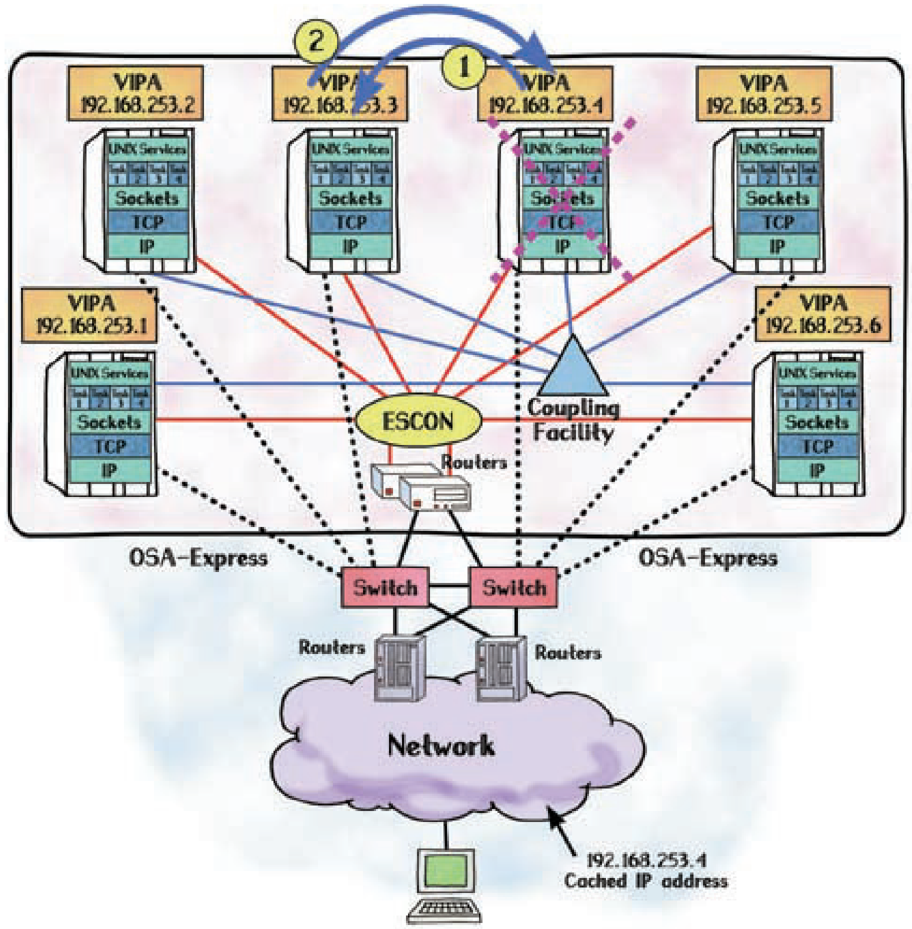
z900 and S/390 enable multiple OS/390 and z/OS images to be run in a single processor complex, or central electronic complex (CEC), using the logical partitioning function. High-speed connectivity functions like the Open Systems Adapter-Express (OSA-Express) can be shared between TCP/IPs in different logical partitions, and the OSA performs a routing function on incoming IP traffic based on destination IP address. This can cause a problem for Network Dispatcher and other outboard solutions, because two such TCPs cannot both be using the same cluster IP address — the OSA would not be able to determine to which TCP/IP to route the incoming traffic. Also, the cost of direct physical connectivity from Network Dispatcher to all IP application hosts, and the difficulties of configuring the external Network Dispatcher node for new application hosts, make the external distribution function undesirable in some configurations. The Sysplex Distributor function in OS/390 V2R10 and z/OS address all of these problems by bringing the distribution function into OS/390 and z/OS TCP/IP.

Note: Such OSA-related restrictions are also be addressed by Generic Resource Encapsulation (GRE), an industry-standard mechanism for encapsulating packets in other IP packets with a different destination IP address. GRE requires that the outboard workload-balancing solutions also support it, as well as the receiving TCP/IP stack. GRE was made available by PTF on OS/390 V2R10 TCP/IP when it became generally available in September 2000.

Both nondisruptive movement of Dynamic VIPAs and Sysplex Distributor are oriented toward TCP connections. They do not apply to UDP applications, where the relationship between client and server over time, if any, is maintained solely at the application level, if at all. The combination of the fact that most server applications allow requests from any IP address (binding to INADDR_ANY in socket terms) and the prevalent use of SOURCEVIPA means that many UDP server applications may not be able to use Dynamic VIPAs unaided. This is because the source address of a response may not match the destination address of the corresponding request, a situation that many UDP clients will not tolerate. Server Bind Control, introduced in OS/390 V2R10 and made available via PTF on V2R8, allows OS/390 TCP/IP to restrict such servers to a single IP address through TCP/IP configuration, without server program modification, thus allowing such servers (for example, DNS/WLM) to share in the benefits of Dynamic VIPAs.

Nondisruptive VIPA Movement

Nondisruptive VIPA Movement in OS/390 V2R10 and z/OS is useful for both of the Dynamic VIPA scenarios. For VIPA Takeover, when the normally hosting TCP/IP is restored, the Dynamic VIPA is immediately activated on the restored stack. The backup stack notifies the restored stack of all active TCP connections, so traffic for those connections will continue to be routed to the backup stack as long as the connections are active. The normal stack, not the backup stack, handles new client connection requests for the Dynamic VIPA. Assuming normal connection duration of minutes or hours (or even days), eventually all connections to the backup stack for the VIPA will end normally, and the VIPA can be deactivated.

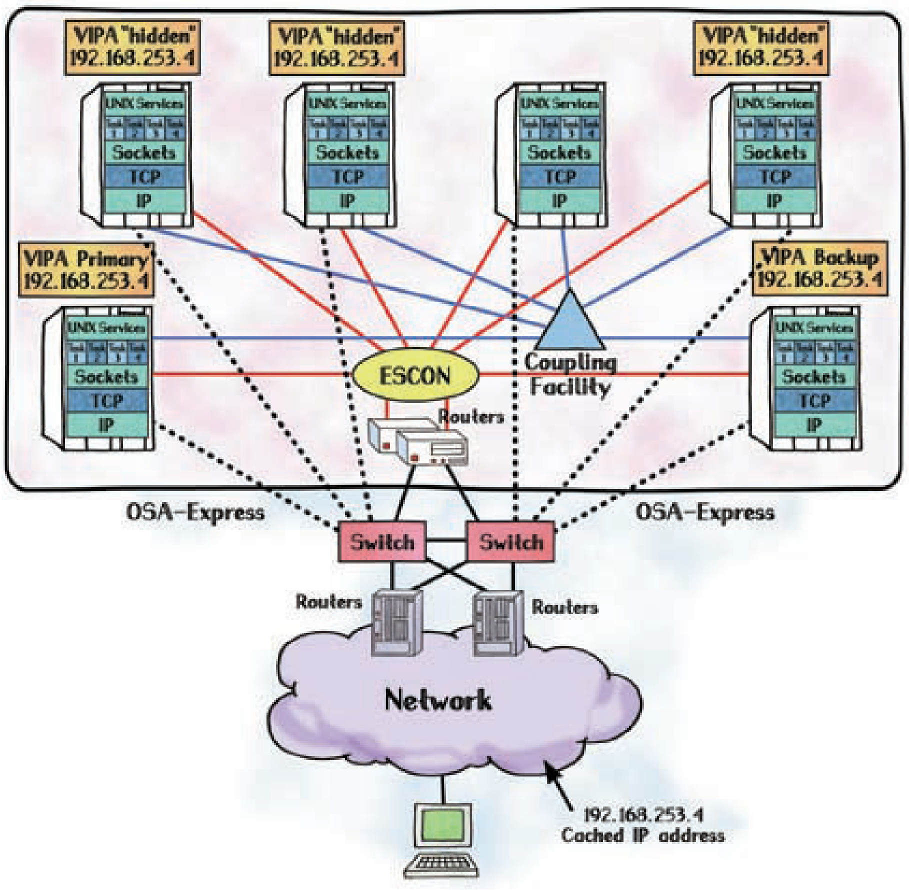


As described above, application-initiated Dynamic VIPAs are tied to a specific application instance. In addition to short-term fluctuations, application traffic patterns and workloads may vary over time, such that it may be desirable to move an application instance to a different OS/390 or z/OS in the sysplex to provide better long-term balance. Adding new application host processors to the sysplex is another case where moving existing application instances may be desirable. Nondisruptive VIPA Movement in V2R10 and z/OS allow such planned application movement to take place nondisruptively, as far as the clients are concerned, simply by starting another copy of the same application instance in another OS/390. The Dynamic VIPA is activated by the application (or JCL or OMVS shell script) on the new TCP/IP, and advertised using dynamic routing protocols to the routing network, but TCP traffic to the existing instance continues to be routed to the former stack until all connections there end naturally. The former application instance can then be shut down, and the new instance services all the client requests.

Sysplex Distributor

Sysplex Distributor was made available in OS/390 V2R10 and in z/OS V1.1 to provide additional network configuration flexibility, particularly as related to network attachment costs and sharing of OSAs among logical partitions. In addition, Sysplex Distributor provides more real-time consultation with Workload Manager for cluster node capacities, as well as consulting the Service Policy Agent to allow the server selection to be influenced by policies governing Service Level Agreements.

Sysplex Distributor builds on Dynamic VIPAs from V2R8. A particular TCP/IP is configured as a routing stack with a Dynamic VIPA, and one or more other stacks can be configured as backup routing stacks for that Dynamic VIPA. An additional configuration statement (VIPADISTRIBUTE) on the primary stack designates the identifying server port numbers and which TCP/IPs (target stacks) in the sysplex will be hosting server applications. This configuration information is distributed automatically to all the target stacks by MVS XCF Messaging, so only the routing stack needs the additional VIPADISTRIBUTE statement. The target stacks each activate the same VIPA address, but do not advertise it to the routing networks. When a server application binds to the designated application port on one of the target stacks, the target stack notifies the routing stack by MVS XCF Messaging that the application is available for work.



When the routing stack receives a connection request from the client for a Distributed VIPA, the routing stack first determines which target stacks have active server applications listening on the port. The routing stack then consults Workload Manager information on relative capacities, and adjusts those capacity values with information from the Service Policy Agent (for Network Quality of Service [QoS] considerations), before selecting the target stack. The connection request is then routed via the appropriate Dynamic XCF IP link to the target stack and the server application. The routing stack keeps track of active connections, so that future client IP traffic for a connection is routed to the same target stack and application. The target stack notifies the routing stack when the connection ends, to allow the routing table entry to be deleted.

If the routing stack should suffer an outage, normal VIPA takeover mechanisms will move the Dynamic VIPA to a backup routing stack. The remaining target stacks notify the backup routing stack of active connections, so that clients connected to those target stacks will not see an outage at all (other than possible TCP retransmissions during the takeover process). When the primary routing stack is restored, nondisruptive VIPA movement is used to restore the routing function to the restored stack, and the backup stack no longer needs to serve as the routing stack, again with no visible disruption to the clients.

The benefits of Sysplex Distributor over external IP workload-balancing solutions are thus as follows:

- Only the primary routing stack and the backup routing stack or stacks need to be connected to the routing network. Target stacks can be connected to the routing network, and traffic from server to client will take the least-cost route, but application hosts need not have physical connectivity to the external routing network.
- Because only one routing stack receives inbound IP traffic from the external network for the Distributed VIPA, and distributes traffic to target stacks using Dynamic XCF IP links, there are no problems with OSA adapters shared among two or more TCP/IP stacks. The OSA always routes inbound traffic to a single (routing) stack.
- Target stacks notify the routing stack automatically when a server application binds to the designated application port, and also when the server closes the listening socket, so the routing stack is always aware of which target stacks actually have applications listening. All application types are thus covered without the need for application-unique advisor programming at the routing stack.
- In addition to capacity information from Workload Manager, Sysplex Distributor uses policy and Quality of Service information from the Service Policy Agent to determine how to distribute the work. For example, if an individual server is not meeting its Service Level Agreement (SLA), new work is directed at other server instances that are meeting their SLA instead, rather than making the lagging server's problem worse by directing additional work to it.

Server Bind Control

Server Bind Control actually addresses two different requirements. The first requirement concerns different and incompatible server application instances identified by the same well-known port number, when no such server instance can be configured to use a specific IP address. One example would be OTELNET and TN3270, both of which use well-known port 23 and accept requests from any IP address (binding to INADDR_ANY in socket terms), but which use different application protocols. If the servers use the same application protocol, normal port-sharing functions would allow the server instances to coexist on the same TCP/IP stack.

However, the stack balances connections between server instances sharing the same port, and has no way to know that a client needs to use one particular server over another. Before Server Bind Control, this requirement was addressed through the use of different TCP/IP stack instances, associating one server instance type with one stack, and the other server instance type with the other stack, such that DNS/WLM returns only addresses appropriate for the name of the particular server type. Note that if the server instances themselves could bind to different IP addresses, there would be no problem running both server instances on a single stack, because the destination IP address would uniquely identify the type of server, and the DNS could be configured to supply the correct IP address for each server type.

Server Bind Control allows the OS/390 TCP/IP stack to be configured to address this. A modification to the PORT configuration statement allows an IP address to be specified and associated with a particular job name. If the server job binds its listening socket to a particular IP address, nothing new occurs. However, if the server job binds its listening socket to INADDR_ANY, OS/390 TCP/IP converts the bind to use only the specified IP address. In the example above, OTELNET would be configured for one IP address, and TN3270 would be configured for a different address, with an appropriate DNS name to IP address configuration. Both servers can now run on the same TCP/IP stack using the same well-known port, and clients are automatically directed to the required server instance.

Note that this also addresses the concern identified above with some UDP server applications. If such a server application is configured to the TCP/IP stack to be bound to a specific IP address, the problem of SOURCEVIPA causing a response to use a source address that is different from the one specified on the corresponding previous request goes away, because the socket is bound solely to the designated address.

The specified address can also be a Dynamic VIPA, and need not already be active on the stack at the time that the application establishes its listening socket. In other words, applications that bind to INADDR_ANY can now use application-initiated Dynamic VIPAs, with all the benefits of such use, including the ability to restart the application on another OS/390 image (with an appropriate TCP/IP configuration already in place) and have the Dynamic VIPA move with the application. This configuration is available for both TCP and UDP applications, because the PORT statement differentiates between the two.

Future Sysplex TCP/IP Functionality

TCP/IP for OS/390 V2R10 and z/OS has come a long way towards becoming a clustered IP application platform, but the work is not yet done. Numerous additional functions are planned to enhance OS/390 and z/OS TCP/IP to move it toward a true single-IP-platform cluster appearance. Stay tuned for news of these new functions.

Parallel Sysplex TCP/IP — In Summary

OS/390 TCP/IP has made great strides since V2R5. It is no longer a single-node stack in the traditional sense. OS/390 TCP/IP uses Parallel Sysplex facilities to maintain awareness of the existence of all other TCP/IPs in the sysplex, and exchanges information to reduce the problem of configuring multiple stacks in a Parallel Sysplex cluster. TCP/IP can automatically establish IP connectivity to other TCP/IPs in the Sysplex using Dynamic XCF, and there are plans to enhance this with higher-speed HiperSockets. Use of MVS System Symbols in configuration files allows common profiles to be maintained with less effort. Configuration of new Dynamic and Distributed VIPAs has been simplified, often to a single profile statement, and the need for coordinated definition changes has been reduced greatly through Dynamic XCF and Sysplex Distributor.

Sysplex Distributor offers additional IP-workload distribution function and flexibility. Fast Connection Reset ensures that clients of server applications using Dynamic VIPAs and Sysplex Distributor are notified quickly of a connection failure. Server Bind Control removes restrictions that in the past resulted in deploying multiple TCP/IP stacks in the same OS/390 image, and extends the benefits of Dynamic VIPAs to additional UDP-based servers.

Future efforts such as Server-Controlled Affinity and Content-Based Routing may allow new application workloads to be distributed for availability and scalability while maintaining proper server-client relationships with reduced or minimal cost. Connection Recovery may some day provide true continuous availability to clients even in the face of application and endpoint TCP outages.

OS/390 TCP/IP and z/OS TCP/IP are well on the way to providing a single-system image to clients for clustered IP server applications. ■

About the Author

John (Jay) A. Aiken, Jr. has been with IBM in a number of different assignments, including almost five years with IBM Japan and a brief 16-month assignment to Corporate Technical Strategy Development at IBM Corporate Headquarters in Armonk. Since 1992, Jay has been with VTAM and TCP/IP for MVS, and is currently the focal point for exploitation of Parallel Sysplex functionality in TCP/IP for OS/390 and TCP/IP for z/OS. Jay was inventor or co-inventor on 17 granted patents or patent applications in his career with IBM.
jaaiken@us.ibm.com

TCP/IP Application Workload Balancing in the S/390 Parallel Sysplex

By Mac Devine

For years, application workload balancing in the S/390® Parallel Sysplex® environment was considered a function of SNA's Generic Resources, but recent enhancements now make application workload balancing a reality for TCP/IP applications as well.

Although the task of workload balancing in a TCP/IP environment is not nearly as straightforward as it is in an SNA environment — due to the wide variety of TCP/IP applications — the value of workload balancing is consistent across both application protocols. By distributing workload among replicated applications in a Parallel Sysplex, you can make efficient use of network resources while significantly enhancing the overall availability of the data center. To fully appreciate these benefits, two concepts, as well as their relationship with one another, need to be understood thoroughly.

Horizontal Growth and Single-System Image

On the surface, you might think that the only goal of application workload balancing in a Parallel Sysplex is simply to "spread the work" among replicated applications. However, the true goal is to spread that workload in a manner that factors in the "health" of individual application instances while allowing additional application instances to be added nondisruptively to the Parallel Sysplex. In other words, you do not want to distribute workload to application instances that are not available or that are performing poorly, and you do not want to be forced to bring down your Parallel Sysplex in order to add needed capacity.

By distributing workload among replicated applications in a Parallel Sysplex, you can make efficient use of network resources while significantly enhancing the overall availability of the data center.

Because the S/390 Parallel Sysplex looks like a single system to the client, the burden of knowing the physical makeup of the data center and choosing the appropriate application instance is shifted from the client to the Parallel Sysplex itself. The Parallel Sysplex provides a single-system image by having either a single name, which represents the replicated application instances (DNS method), or a single IP address, which represents the replicated application instances (cluster method). By allowing applications to be moved or added to new locations within the sysplex while still being considered part of the same single-system image application group, the data center is allowed to grow without impact to the other sysplex members, the WAN, or the clients.

The cross-system coupling facility (XCF) dynamics function, delivered in OS/390® V2R7, allows for this horizontal growth of the data center. XCF dynamics introduced the concept of a single local IP address for connectivity to all other stacks in the Parallel Sysplex. In other words, XCF IP links under XCF dynamics look very much like a shared-medium LAN such as Ethernet or token ring but all point-to-point connectivity between stacks is dynamically defined using the MVS™ XCF facility keeping track of the members of a particular XCF group and allowing the members to communicate with one another.

DNS Method

A DNS method for TCP/IP application workload distribution uses a single name to represent the replicated application instances within the Parallel Sysplex. There are two types of DNS methods available in the marketplace today: DNS weighted round-robin, which is available on a variety of platforms, and Domain Name Server/Workload Manager (DNS/WLM) support, which is provided by the S/390 server. The DNS/WLM support available on OS/390 V2R5 and later uses input from the MVS Workload Manager (WLM) to factor in the current health of the application instance in terms of CPU consumption and availability (that is, is it currently up and running on a particular stack?). This provides a competitive advantage over the round-robin approach, which does not factor in these variables.

Because of the time needed to perform a DNS resolution for every connection request from the client in the DNS/WLM model, it is generally recommended only for long duration connections like Telnet and FTP. Another important factor to consider is whether clients or the primary DNS (that is, a DNS that is more local to the client than the S/390 DNS), or both, cache the IP address even though the S/390 DNS supplied a time-left value of zero on the initial DNS resolution. Without the honor of the time-left value of zero, the availability of the Parallel Sysplex is severely compromised, because the client might attempt to reconnect using the IP address of a "down" application server. The Virtual IP Address (VIPA) takeover function in OS/390 V2R8 significantly improves the availability in this scenario, because the cached IP address of the server can be a dynamic VIPA address that can automatically be recovered on a backup stack. However, the WLM information is not factored into the reconnect decision.

Cluster Method

A cluster method for TCP/IP application workload distribution uses a single IP Address that represents the replicated application instances within the Parallel Sysplex. There are a variety of implementations in the marketplace, including IBM's Interactive Network Dispatcher (IND) and Cisco's MultiNode Load Balancing (MNLB) feature of Local Director. Both implementations use an IP address advertised as being owned by a particular router, and therefore, all inbound connection requests for that IP address are directed to that particular router. Upon arrival at the distributing router, the connection is forwarded to the destination server that is chosen based on MVS WLM information obtained from agents running on the S/390 server and providing feedback to the router. The destination server processes the inbound connection request because the S/390 TCP/IP stacks have also been defined with a loopback alias address that corresponds to the cluster address owned by the router.

Cisco's MNLB has the advantage of being able to physically reside anywhere in the IP network (that is, the IND must be on the same IP subnet as the S/390 servers in the Parallel Sysplex) as long as there are only Cisco routers upstream to the data center. This is because the Cisco routers communicate through their Dynamic Feedback Protocol (DFP) to enable forwarding of the data packets to the "real" destination server instead of all inbound data packets being sent to the router owning the cluster address. In a future release of OS/390, a function called the Sysplex Distributor will also use the cluster method to allow for sysplex-wide VIPAs corresponding to individual ports, thus allowing application-centric workload distribution (that is, total independence from network connectivity having to be factored into the workload-distribution decision). The Sysplex Distributor will also be able to factor in Quality of Service (QoS) and policy information into the workload-distribution decision, in addition to WLM information.

Which TCP/IP Application Workload Distribution Method do I Choose, DNS or Cluster?

Before answering this question, other questions need to be answered concerning the applications that will be included in the workload-distribution model:

- Are connections to these applications of short duration (for example, Web) or of long duration (for example, Telnet)?
- Do these applications support WLM registration, DNS registration, or both?
- Do these applications require QoS information, policy information, or both, to be factored into the distribution decision in addition to WLM information?

In addition, questions concerning types of clients, network resources and network connectivity involved in the workload-distribution model must also be addressed:

- Do these clients cache the IP address of the application server?
- Will S/390 DNS be used?
- Is there enough capacity on the system that will own the cluster IP address?
- What is the network attachment to the data center?
- What is the desired location of the distribution point in the network?
- Do all clients have access to the network resources needed for the distribution model?
- Do different groups of clients require different QoS or policy? ■

About the Author

Mac Devine is a 1989 graduate of Clemson University with an MS in Mathematical Sciences. He spent 7 years in VTAM development working on a variety of APPN/HPR functions and was the Chief Programmer of several major releases. He spent 3 years as part of the Communications Server for OS/390 Strategy and Design group that is responsible for providing the strategic direction for the Communications Server component of OS/390. He served as the Chief Designer for Communications Server for OS/390 V2R7 and for Communications Server for OS/390 V3R10. He has spent the past year as a WebSphere® Networking Solutions Architect responsible for developing joint solutions between WebSphere and networking vendors.
wdevine@us.ibm.com

SNA and IP Workload Balancing in the S/390 Parallel Sysplex

By Mac Devine

Although SNA and IP application servers are very different, the requirements for effective workload distribution among a clustered group of replicated servers are consistent across both application protocols.

These requirements go well beyond finding a simple methodology for spreading the workload. More importantly, they include the ability to maximize the overall availability of the network, make efficient use of network resources, allow for non-disruptive growth, and balance workload in accordance with business goals. The attributes of the S/390® Parallel Sysplex® architecture and its exploitation by the Communications Server for OS/390® are the keys to reaching these goals.

SNA Workload Balancing

The Generic Resources function is the key to effective workload balancing in an SNA environment. Generic Resources allows replicated SNA applications to be known by a single generic name. Communication Servers for OS/390 uses the sysplex coupling facility to provide a data structure that is accessible to all sysplex images, so that real-time information about the applications registering under a particular generic name can be shared. This allows sessions to be balanced across all the applications within a generic resource group according to goals defined within the S/390 Workload Manager (WLM). In addition, many SNA applications, such as the CICS® and IMS programs, use shared message queues in the coupling facility, which allow them to also balance transaction workload across the replicated applications. Therefore, in many ways, Generic Resources is more important as an availability function than it is as a workload-balancing function. To fully appreciate the availability benefits of Generic Resources, some concepts need to be understood.

Role of APPN/HPR

Generic Resources relies on the dynamics and networking capabilities of Advanced Peer-to-Peer Networking® (APPN®) to ensure that sessions can be balanced across a multi-instance application with the assurance that sessions are never distributed to an unavailable instance and without any dependencies on the client or network. Therefore, Generic Resources requires at least one network node (NN) residing in the sysplex. As in any good sysplex design, it is important to remove any single points of failure, so it is wise to have two NNs. For all sessions originating from the network and destined for a generic resource in the sysplex, the NNs in the sysplex will access the coupling facility structure to retrieve the registered application instances, interface with the WLM to determine the best application instance, and handle all of the routing and directory services needed to complete session setup to the best application instance.

Although High-Performance Routing (HPR) itself is not required for session balancing with Generic Resources, it is a very valuable because it will allow non-disruptive path switches of the session around intermediate failures (for example, links). This additional availability can be significantly increased by extending HPR further out into the network so that networking solutions like Enterprise Extender (that is, HPR/IP) are also valuable additions for Generic Resources.

Role of the Application Instance

It is also important to understand that the functions provided by Generic Resources are shared by the Communications Server for OS/390 and the multi-instance application itself. The responsibilities of each application might only include the registering and deregistering of the generic name across its VTAM® application programming interface (API) or it might also include managing a session affinity. This session affinity is used for certain application protocols to ensure that a particular end user will be reunited with a particular application instance on a re-logon attempt following a session failure. Although APPN/HPR provides significant availability for Generic Resources, it does not address outages of an application itself or its underlying subsystems (that is, VTAM, MVS™, or hardware failure).

If the SNA session is broken due to such an outage, then the success of a re-logon attempt by the end user is directly related to the presence of a session affinity in the Generic Resources structure of coupling facility. If the affinity is owned by the application instance, which is the case for LU 6.2 sessions, then it survives the outage and therefore, a restart of the application instance, either manually or by the automatic restart manager, is required for re-logon success. However, if the affinity is owned by the Communications Server for OS/390 (for example, SNA 3270 and TN3270), the end user is allowed to log on to another available instance of the application, because the affinity went away as soon as the original session was broken. This usually means that the session can be restarted in seconds instead of the 10 or more minutes it usually takes to recover the application instance.

IP Workload Balancing

The task of workload balancing in an IP environment is not nearly as straightforward as in an SNA environment, due to the wide variety of IP workload-distribution choices. To choose the best method for your network, it is important to understand your current and future network requirements as well as your network configuration. For example, a cluster IP address method for workload distribution, like Cisco's MultiNode Load Balancing (MNLB) or IBM's Interactive Network Dispatcher, uses a single IP address to represent the replicated application servers, and therefore requires that all clients have the same access to the node advertising the cluster IP address.

In contrast, a DNS method for workload distribution, like S/390 DNS/WLM, uses a single name to represent the replicated application servers, and therefore has no dependency on the network access of the clients. However, it does depend on the clients (and other DNS nodes) honoring the time-left value of zero returned on the resolution (that is, no caching is allowed). Although each of these methods supplies workload balancing based on input from the S/390 WLM, they do not factor client/server-specific policy or Quality of Service (QoS) into the workload distribution decision. These additional factors are vital in a service-provider environment where differentiated services for different client groups and adherence to service-level agreements are required. To alleviate any dependence on network access or specific client behavior, as well as to address the service-provider environment, the Communications Server for OS/390 has developed the Sysplex Distributor in OS/390 V2R10.

Sysplex Distributor

The Sysplex Distributor function is actually shared among the TCP/IP stacks in the Parallel Sysplex by utilizing Release 7's XCF dynamics support for inter-sysplex communication and Release 8's Dynamic Virtual IP Address (VIPA) support for configuration and recovery. The role of each stack is established by configuring a Dynamic VIPA that has been defined with a distribution server list for a particular port or ports. When the ALL keyword is specified in the distribution server list, any TCP/IP stack on an existing or new sysplex image automatically becomes a candidate for workload distribution. This can reduce the administrative burden significantly in a rapidly changing environment (e-commerce, for example) by allowing the complete flexibility to move application servers or add new application-server instances to new locations within the sysplex and still be considered part of the same single system-image application group. Once a dynamic VIPA becomes a sysplex-wide VIPA, workload can be distributed to multiple server instances without requiring changes to clients or networking hardware and without delays in connection setup, thus allowing the data center and the customer's business to grow nondisruptively.

Configuration

The stack defined as the primary owner of a Dynamic VIPA (by using a VIPADEFINE statement in the TCP/IP profile) receives all IP datagrams destined for that particular VIPA. If this Dynamic VIPA definition includes a distribution server list for a particular port or ports, then the Sysplex Distributor code running in the primary owning stack is activated so that it can distribute connection setup requests destined for that particular VIPA and port or ports.

The stacks identified as candidates for workload distribution for particular ports require no configuration, because they are notified of their role through inter-sysplex communication with the primary owner. In order to avoid the primary owning stack being a single point of failure, it is recommended that you define one or more backup stacks (by using a VIPABACKUP statement in the TCP/IP profile).

The backup stacks or stacks can inherit the distribution server list through inter-sysplex communication with the primary owner, or can specify an entirely different distribution server list to be used if the primary owner experiences an outage. It is also possible to specify a distribution server list at the backup stack so that distribution occurs only during an outage of the primary owner. This allows the additional workload originally destined to the primary owner to be spread across multiple servers, thus lessening the overall impact to the data center during the outage.

Current Functionality

Because the Sysplex Distributor resides in the Parallel Sysplex itself, it has the ability to factor real-time information regarding policy, QoS and application status into the distribution decision. By combining these "real-time" factors with CPU utilization information, the Sysplex Distributor has the unique ability to ensure that the best destination server instance is chosen for a particular client connection, while ensuring that client/server-specific service-level agreements are maintained.

Unlike other workload distribution methods, the Sysplex Distributor uses a variety of sources to obtain its distribution decision criteria. In addition to using information obtained from the S/390 WLM, it also uses information from the Communications Server for OS/390 Service Policy Agent and information directly obtained from the target stacks themselves. Although it is very desirable to factor in the CPU utilization supplied by WLM to understand the workload on a particular system, it's not enough, because it does not consider the network performance (that is, QoS) in its workload-balancing algorithm.

Network performance is often the bottleneck in the overloaded Internet/ISP network, and is a critical factor in the end-to-end response time. Also, enterprise networks often have alternative paths to address network availability and reliability, and yet they're not taken into consideration in the optimization of end-to-end response time and availability/reliability. This makes it difficult for the service provider to adhere to service-level agreements. For example, it might be desirable to route more incoming connection requests to a more-loaded server (higher CPU utilization) with better network performance than to a less-loaded server with much worse network performance. Therefore the Service Policy Agent will inform the Sysplex Distributor whenever a particular server instance is not adhering to the QoS specified in its service-level agreement.

The Sysplex Distributor has also been updated to include policy concerning the clients' characteristics into the workload-distribution decision. This policy can include the IP characteristics of the client (IP address and port, IP subnet, and so on), time of day, day of week, and any other policy rule supported by the Service Policy Agent. An example of the usefulness of this function is in application hosting, or Internet service provider (ISP) marketplace, where clients accessing the same application can be assigned to different servers having different capability, connectivity, or both.

The target stacks also assist the Sysplex Distributor in making the best distribution decision possible by supplying immediate server status information through inter-sysplex communication. Whenever an application server binds and listens to a port on a target stack being serviced by the Sysplex Distributor, the target stack sends a notification through inter-sysplex communication to the primary owner, indicating that an application server exists and is ready to accept connections. When the application terminates or closes the listening socket, a corresponding notification is sent to the primary owner so that no additional connection requests will be distributed to this stack. The Sysplex Distributor has up-to-date information on available servers on target stacks, so there is no need for application-specific advisors to issue periodic null application requests to determine existence of functioning servers, as is the case with many other workload-distribution methods.

In addition to providing workload distribution, the Sysplex Distributor also enhances the availability provided by the Dynamic VIPA support. The Dynamic VIPA support in Release 8 allowed a backup stack to take over the VIPA in cases where the primary stack experienced a system outage. It did not, however, allow nondisruptive movement of the VIPA during normal operations. There was no way to preserve existing connections while relocating an application server that was using the VIPA through BIND or IOCTL DVIPA, or while recovering the VIPA ownership at the primary stack upon its recovery from the outage. The nondisruptive VIPA takeover function, which will also be available in Release 10, allows for this freedom of movement by maintaining the active connections established with the backup stack and allowing the VIPA to move immediately back to the primary owning stack. The primary owner will then be allowed to accept all new connection requests and internally use the Sysplex Distributor function to forward the IP datagrams to the backup stack for connections that were active at the time of the nondisruptive takeover. This ensures minimal impact for planned or unplanned system outages, because workload can be redirected without affecting existing connections.

Future Functionality

Watch this space for more exciting functions involving the Sysplex Distributor, coming your way in the not-so-distant future. The best is yet to come! ■

About the Author

Mac Devine is also the author of "TCP/IP Application Workload Balancing in the S/390 Parallel Syplex" in this edition of *NCP and 3745/46 Today.*

Policy-Based Quality of Service in OS/390 V2R10 and z/OS V1.1

By Lap Huynh

In the past few years, it has become evident that the Internet is becoming the information marketplace where individuals go to conduct their lives and companies go to conduct their business. The Internet has brought about an explosive growth in the number of new applications designed to enable services and facilitate business transactions.

The rate and the nature of data (for example, time delay-sensitive vs. loss delay-sensitive) generated by these applications put a tremendous strain on the network IT infrastructure despite advances in high-speed transmission technology. Different data types (for example, FTP, Web browsing, enterprise resource planning [ERP], interactive transactions, and so on) require different levels of Quality of Service (QoS), and necessitate a means for service differentiation to guarantee that important e-business traffic gets better service than less important Web browsing or FTP data transfer. This is especially critical considering that business on the Web happens in real time. It's no longer a question of when a service can be delivered but rather how fast it can be delivered to meet customers' demands.

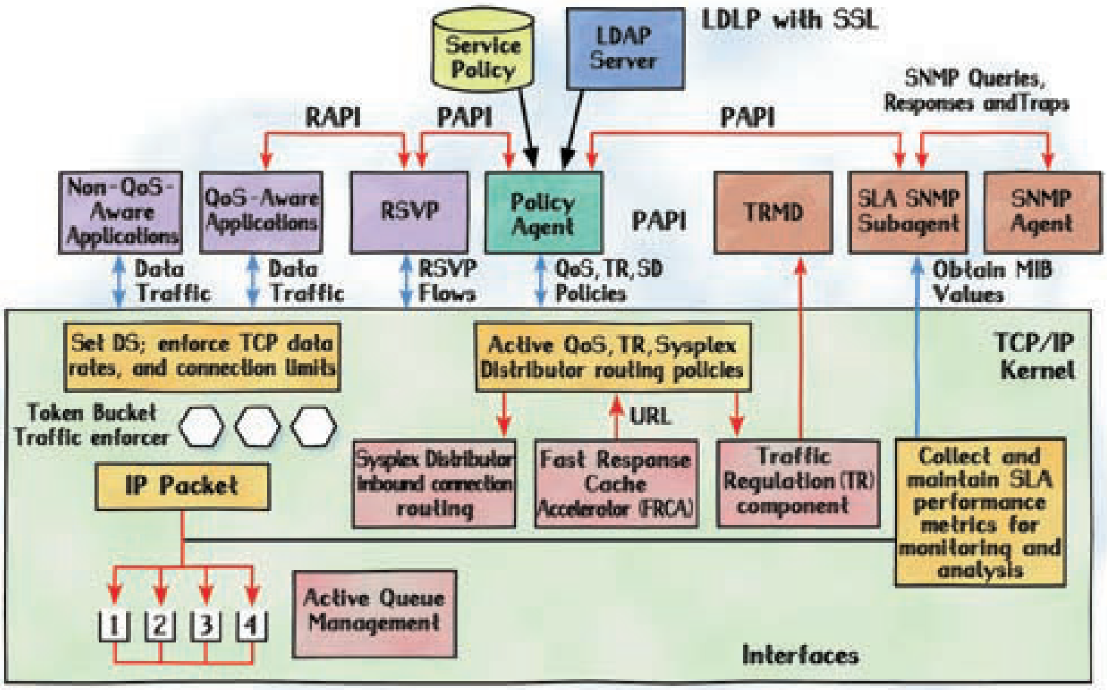
As the growth of Web traffic puts a strain on the IT infrastructure, it also creates a strain on a company's IT organization due to an increase in infrastructure complexity. This complexity comes from server and network capacity planning, configuring and managing servers, configuring network devices and application response times, and planning and installing new software. It comes as no surprise that IBM's enterprise customers have dealt with these issues effectively over the years using IBM's rich product architectures, services and functionality (for example, Workload Manager [WLM] to prioritize application workloads, SNA and APPN®/HPR for network prioritization, and so on). But as they move to enable their IT infrastructure to support e-business, they will require a new set of products and functions to face this ever increasing complexity. While these large companies can afford an in-house IT staff and infrastructure, it is much harder for smaller companies to justify the cost.

This leads to a new breed of service providers, the *application service providers (ASPs).* These ASPs manage complex IT infrastructure that provides application- and data-hosting services to customers who need to manage their business on the Web. The service-level agreement (SLA) requirements that ASP customers put on the ASPs are likely to be more stringent than if the customers had their own IT infrastructure. These requirements include high-speed connectivity, QoS, scalability, reliability, availability, and security. ASPs that can differentiate themselves from others by having an IT infrastructure capable of addressing these issues will prevail in this new but crowded market.

Communications Server for OS/390 and z/OS has defined a set of Lightweight Directory Access Protocol (LDAP) directory-based policy schemas to enable the administering of the QoS policy.

The IBM S/390® and the new zSeries 900 servers are known to be among the best server platforms for reliability, scalability, availability and security. The OS/390® V2R10 software, which became available in September 2000, and the new z/OS V1.1 will bring S/390 and z900 servers to the next level in functionality, enabling enterprise customers to provide predictable services, and ASPs to differentiate themselves in providing intelligent hosting services. There are many features in OS/390 V2R10 and z/OS; however, the focus of this article is on the policy-based QoS functions and how they can help enterprise customers and ASPs in addressing critical issues in hosting services. The following five major topics in policy-based QoS are discussed:

1. Service-provisioning policy and QoS enforcement
2. Web content-based QoS
3. Sysplex Distributor (SD) routing policy and load-balancing mechanism
4. Traffic Regulation (TR) management
5. Service-level agreement (SLA) management information base (MIB)


Service Policy
LDAP Server
LDLP with SSL
RAPI
PAPI
PAPI
SNMP Queries, Responses and Traps
Non-QoS-Aware Applications
QoS-Aware Applications
RSVP
Policy Agent
PAPI
TRMD
SLA SNMP Subagent
SNMP Agent
Data Traffic
Data Traffic
RSVP Flows
QoS, TR, SD Policies
Obtain MIB Values
Set DS; enforce TCP data rates, and connection limits
Active QoS, TR, Sysplex Distributor routing policies
TCP/IP Kernel
Token Bucket Traffic enforcer
URL
IP Packet
Sysplex Distributor inbound connection routing
Fast Response Cache Accelerator (FRCA)
Traffic Regulation (TR) component
Collect and maintain SLA performance metrics for monitoring and analysis
1
2
3
4
Active Queue Management
Interfaces

The above figure shows the different components that together form the foundation of policy-based QoS function in OS/390 V2R10 and z/OS. Note that these components and this figure will be referred to throughout this article.

Overview of IP QoS

The Internet Engineering Task Force (IETF), the standards body for the Internet, has defined two mechanisms for providing QoS:

- The first mechanism, the *Integrated Services (IntServ),* is an end-to-end reservation-based service that uses explicit *Resource ReSerVation Protocol (RSVP),* which is a signaling protocol, to request an appropriate level of service for specific traffic “sessions/flows.” Each networking device along an RSVP path from the sender to the receiver needs to maintain the individual reservation state until the reservation is explicitly torn down or the RSVP refresh timer expires. This reservation state corresponds to how many resources (for example, bandwidth, buffer space) are allocated to the reservation. IntServ is appropriate for traffic types that require bandwidth and delay guarantee such as Voice over IP (VoIP) or video streaming.
- The second mechanism, the *Differentiated Services (DiffServ),* provides service differentiation between broad classes of users and applications. In other words, it is a form of aggregation of traffic class with the same network service provision (for example, applications such as ERP, CRM, HTTP, and so on).

DiffServ uses the Differentiated Services Code Point (DSCP) in the IP header to indicate different QoS levels. The IETF has recently defined a standard format for the DiffServ field in the IP packet header in RFC 2474 (the same byte that was used to indicate the type of service [ToS] byte specified in RFCs 791 and 1812) to carry different DSCP values along with some initial defined classes (for example, Assured Forwarding, Expedited Forwarding). Each networking device will, based on the DSCP value, provide appropriate QoS treatment to the corresponding traffic class.

Without explicit signaling like IntServ, DiffServ depends on the per-hop behavior of the networking devices to provide consistent end-to-end QoS treatment. Networking devices can utilize a variety of queueing and scheduling services in provisioning QoS. Some examples of these services include class-based queueing (CBQ), weighted fair queueing (WFQ), weighted round robin (WRR), weighted random early discard (WRED), and so on. OS/390 V2R10 and z/OS support both forms of QoS provisioning, IntServ and DiffServ.

Overview of QoS Policy

With QoS provisioning comes a requirement for QoS level policy that controls service differentiation within a network by an enterprise network administrator (or an ASP). Additionally, a centralized mechanism must be provided for storing and retrieving these QoS level policies in order to guarantee consistency across network devices. Because the directory is an important component in creating intelligent networks, Communications Server for OS/390 and z/OS has defined a set of Lightweight Directory Access Protocol (LDAP) directory-based policy schemas to enable the administering of the QoS policy. These schemas closely follow the work that is being conducted by the Directory Enabled Network (DEN) working group, which is under the auspices of the Distributed Management Task Force (DMTF) and the IETF.

QoS policy consists of, essentially, one or more rules that describe the actions to be taken when specific conditions exist. The semantic can be simply illustrated as follows:

If condition then action

Condition specifies a set of filters to identify certain network activities or traffic. For example, a filter can identify traffic from a client's node IP address attempting to access a server's node IP address to request specific service from an application (a Web-application server). *Action* specifies how the identified activity or traffic must be treated. For example, a Web-application request from an important client should be assigned to high priority with a minimum throughput.

Policy rules, conditions, and actions together define the finite-state machine for networks to operate in a predictable manner to enable service differentiation that is device-independent and consistent end-to-end (for example, traversing administrative boundaries). As such, it must be unambiguous. Only one rule and associated action can be applied to a particular unit of network activity or traffic at a time. Another important attribute of policy is that it must allow a hierarchical construction of rules and grouping capability to enable a network administrator to build complex policy from a simple set of policy objects. QoS policy, in short, serves as a foundation upon which enterprise network administrators and service providers can build SLA for a given client or user.

QoS Policy and the Policy Agent

OS/390 and z/OS allow network administrators to define QoS policy in two ways. One way is through a policy configuration file and the other way is through an LDAP directory-based repository. The Policy Agent (Pagent) component shown in the first figure can either access the service-policy configuration files or go to an LDAP server, or both, to retrieve QoS policy entries or objects. Because policy might contain sensitive information regarding network-resource allocation and SLA information, access to an LDAP server can be protected by enabling Secured Socket Layer (SSL) security in the Pagent configuration file. Note that the Pagent configuration file can itself be protected by the IBM Resource Access Control Facility (RACF®) program.

The OS/390 and z/OS QoS policy core schema closely follows to the core policy schema effort that is currently ongoing in the IETF. This core schema contains all the functions that are necessary for the construction of complex policy (for example, nesting of policy groups, rules, containment, and so on). The Internet Draft (a working document within the IETF), which contains a snapshot of this core schema on which the implementation of OS/390 and z/OS is based, is shipped as a sample file. Refer to this file for further information.

The following core schema classes are supported in OS/390 and z/OS:

```
ibm-policyGroup
ibm-policyRule
ibm-policyCondition
ibm-policyTimePeriodCondition
ibm-policyAction
```

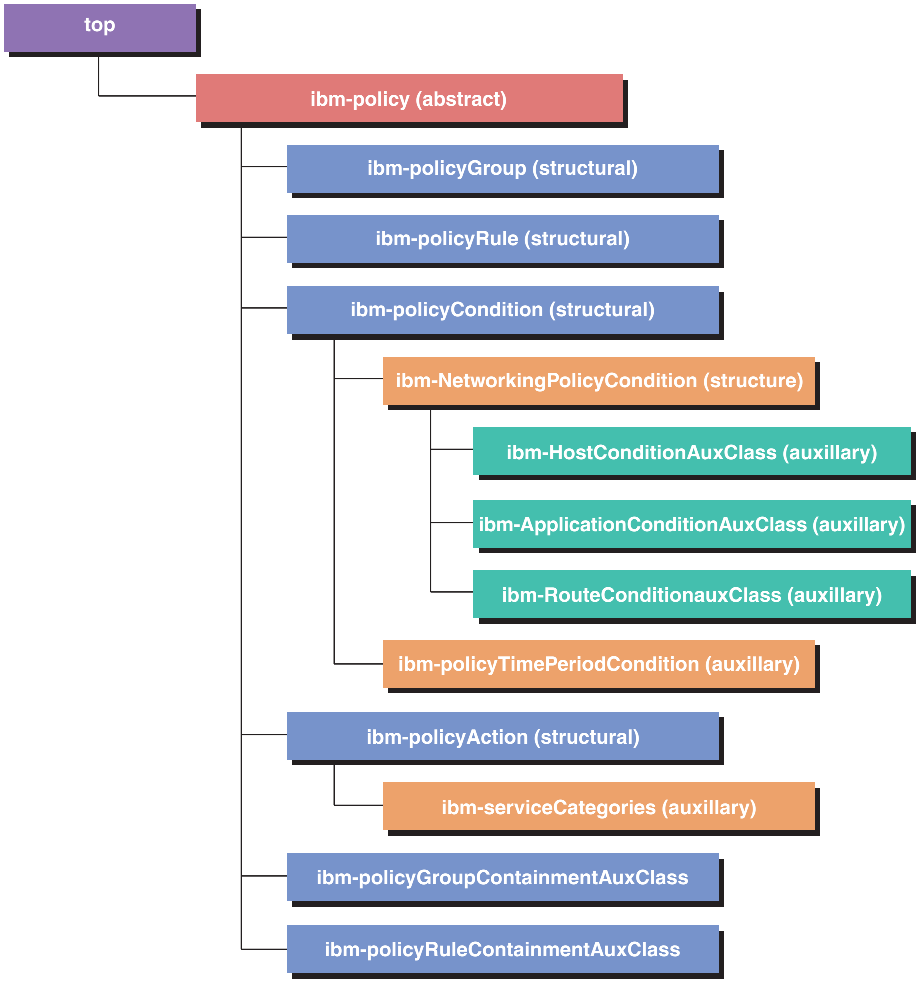
The core schema defines general structural classes for defining policies, but specific applications, such as QoS policy, require specialized subclasses and auxiliary classes that are derived from the core `policyCondition` and `policyAction.` OS/390 and z/OS have defined the following subclasses and auxiliary classes for use in defining QoS policy:

```
ibm-networkingPolicyCondition
ibm-HostConditionAuxClass
ibm-RouteConditionAuxClass
ibm-ApplicationConditionAuxClass
ibm-serviceCategories
```

The diagram on the right shows the LDAP QoS classes and their hierarchy.

As shown, the **`ibm-NetworkingPolicyConditions`** subclass has three auxiliary classes that are designed to categorize the criteria that network administrators or service providers commonly use to control access to network resources and services (that is, the QoS level specified as part of the action):

- **`ibm-HostConditionAuxClass`** has attributes to specify source or destination IP addresses/range, or both. They're used to identify client and server nodes to which a particular policy rule is to be applied.
- **`ibm-ApplicationConditionAuxClass`** has several attributes to specify specific applications to which a policy rule is to be applied. These attributes include source/destination port numbers, protocol ID, application name (also referred to as job name), and application data. The application data is used to classify HTTP Web request universal resource locator (URL) pages for the purposes of assigning different QoS levels for different content retrieval (we'll discuss this feature later on in more detail. It's also known as *transactional QoS*).
- **`ibm-RouteConditionAuxClass`** has attributes to specify inbound and outbound interfaces on which a policy rule is to be applied. It is used mainly to limit the scope of a rule or rules to a particular subnet or subnets (for example, traffic entering different Internet service providers [ISPs] may need to be marked with a different ToS/DSCP value).


top
ibm-policy (abstract)
ibm-policyGroup (structural)
ibm-policyRule (structural)
ibm-policyCondition (structural)
ibm-NetworkingPolicyCondition (structure)
ibm-HostConditionAuxClass (auxillary)
ibm-ApplicationConditionAuxClass (auxillary)
ibm-RouteConditionauxClass (auxillary)
ibm-policyTimePeriodCondition (auxillary)
ibm-policyAction (structural)
ibm-serviceCategories (auxillary)
ibm-policyGroupContainmentAuxClass
ibm-policyRuleContainmentAuxClass

The **`ibm-policyTimePeriodCondition`** subclass specifies when a corresponding policy rule is active. This enables network administrators or service providers to schedule tasks and assign them different QoS levels depending on the time of day. For example, a client's data backup can run at a high priority after working hours when more resources are available. The time specification supported by OS/390 and z/OS allows networks that cross time zones to sync up on their time with respect to when a set of policy rules should be active or idle.

The **`ibm-policyAction`** contains the **`ibm-serviceCategories`** auxiliary class, which has several attributes that are applicable to either RSVP or DiffServ QoS level. These attributes will be explained further in the next section that describes OS/390 QoS traffic provisioning enforcement.

As mentioned above, Pagent can retrieve QoS policy information either from a configuration file or from an LDAP server. Defining QoS policy in a configuration file can be accomplished by following the format of the policyRule and policyAction statements described in the Pagent sample configuration file (also documented in *OS/390 V2R10.0 IBM CS IP Migration Guide, SC31-8512 and OS/390 V2R10.0 IBM CS IP Configuration Guide, SC31-8725*) shipped with OS/390 and z/OS. In most cases, there is a one-to-one mapping of QoS function attributes that can be specified either through LDAP schema or through the configuration statement. However, there are specific functions that can be defined only with the Pagent configuration file, such as those policies that apply to the Traffic Regulation (TR) component. These exceptions are discribed later in this article.

With QoS provisioning comes a requirement for QoS level policy that controls service differentiation within a network by an enterprise network administrator (or an ASP).

QoS Enforcement

The **`ibm-serviceCategories`** auxiliary subclass of the **`ibm-policyAction`** class contains attributes that describe QoS actions in terms of observable or measurable behaviors for both RSVP and DiffServ. The following sections discuss these attributes and how they are enforced.

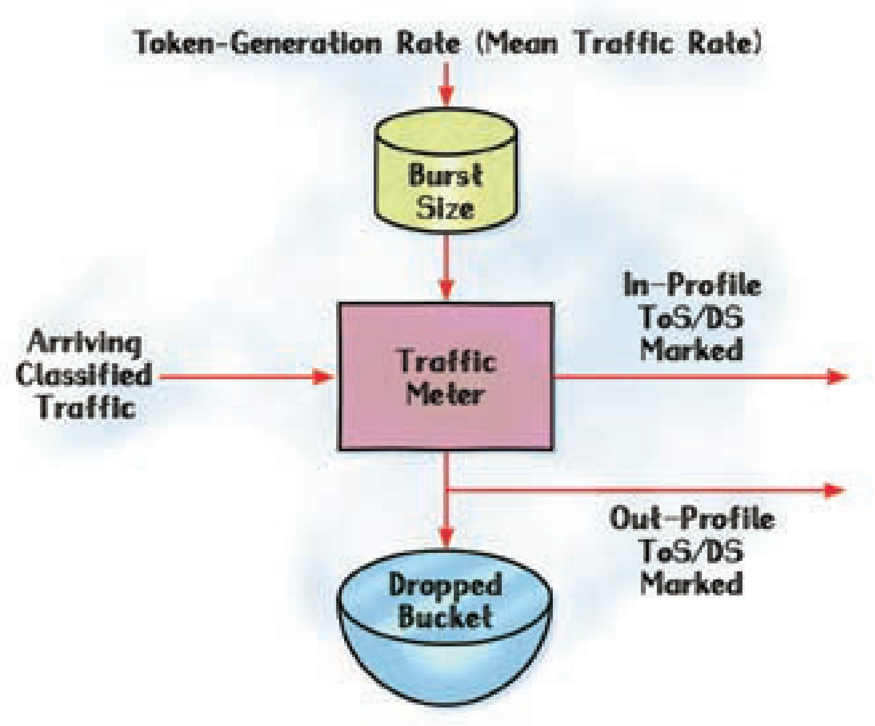
DiffServ Enforcement

There are many attributes in the **`ibm-serviceCategories`** subclass that network administrators or service providers can use to control DiffServ QoS service levels. Available functions include setting the ToS/DSCP value, which affects how traffic is treated end-to-end, controlling individual TCP connection throughput and assigning aggregated committed bandwidth for all traffic sent to clients from the S/390 or z900 server. Within S/390 and z900 servers, the ToS/DSCP value determines which of the four priority queues a packet will be sent to when it is transmitted over a Queued Direct Input/Output (QDIO) device. As the packet traverses the network, the packet's ToS/DSCP value will determine its appropriate treatment by the network devices along the path. By manipulating the ToS/DSCP or setting the limit on TCP connection throughput, or both, network administrators or service providers can effectively control the application response time for respective clients.

There are two ways in which enterprise network administrators or service providers can control the total committed bandwidth that is allocated to applications or clients (for example, limiting bandwidths for different level of premiums, and so on). Limiting the committed bandwidth available to certain applications or clients can avoid the negative effect of "greedy" clients on other, more-critical clients. There are two ways in which enterprise network administrators and service providers can control the total committed bandwidth that is allocated to applications or clients (for example, to limit bandwidths based on different levels of premiums).

- The first method limits the number of TCP connections and each individual TCP connection throughput. With this method, network administrators can set the maximum number of connections that clients (or a set of clients from the same subnet) are allowed to open, as well as the maximum throughput for each connection to send or receive data. For example, during working hours, FTP connections and throughput can be limited such that their total bandwidth usage will have minimum impact on other traffic types.

- The second method uses the DiffServ Token Bucket traffic-policing function. This is more flexible than the first method because it applies also to the User Datagram Protocol (UDP) and enables network administrators and service providers to specify how excess traffic will be handled. Generally, how excess traffic is handled directly affects how much a client is charged for access. With the Token Bucket traffic-policing function, excess traffic can be dropped or transmitted at a different QoS level. OS/390 and z/OS have a function that can simulate dropped packets, causing TCP to reduce throughput without packets actually having been dropped. This significantly increases overall efficiency. For instance, normal traffic can be sent at one QoS level that has less probability of being dropped in the network when congestion occurs, and excess traffic can be sent at another QoS level that has a higher drop probability. This way, excess traffic will be discarded first when congestion occurs, minimizing the effect on other traffic with better QoS levels. The following figure shows the functional components of a Token Bucket.


Token-Generation Rate (Mean Traffic Rate)
Burst Size
Arriving Classified Traffic
Traffic Meter
In-Profile ToS/DS Marked
Out-Profile ToS/DS Marked
Dropped Bucket

RSVP Enforcement

For applications that need to explicitly reserve bandwidth through RSVP using RSVP API (RAPI), the RSVP action attributes in the **`ibm-serviceCategories`** class allow network administrators or service providers to limit how much an application can reserve for an RSVP flow. In addition, the number of flows can be limited, or reservation can be denied altogether. QoS RSVP policy is communicated between the PSVP Daemon (RSVPD) component and Pagent through Policy API (PAPI), as shown in the first figure.

Once a reservation is made, traffic will be policed through a token bucket mechanism as described in the previous section. Traffic explicitly reserved will be transmitted with a specific ToS/DSCP value specified in the QoS RSVP policy so that it can be treated accordingly as it traverses the network. When a reservation is made over an asynchronous transfer mode (ATM) subnet, a separate virtual circuit (VC) is activated with QoS characteristics derived from the QoS parameters of the RSVP reservation.

Sysplex Distributor function manages the sysplex as a cluster of nodes, and dynamically monitors and balances workloads among these nodes.

Web Content-Based QoS

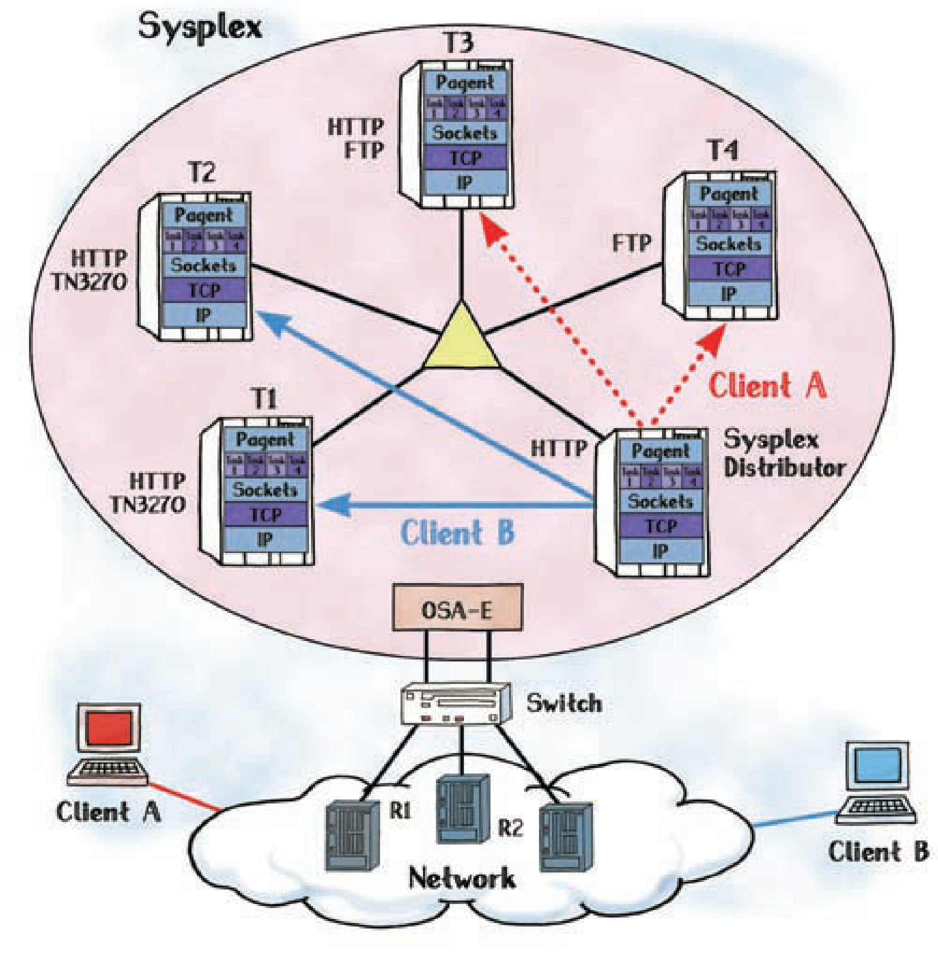
The Web content-based QoS feature in OS/390 V2R10 and z/OS V1.1, together with the IBM WebSphere® platform's use of the OS/390 and z/OS WLM function, offers enterprise network administrators and service providers a critical function that they can leverage for a competitive advantage in enabling their business or services on the Web. A variety of content categories is delivered over the Web, and each of them requires different QoS levels. For instance, an HTTP request for downloading a file (for example, browsing) should have a lower QoS level than a purchasing transaction that is being processed.

The WebSphere platform's use of WLM function routes incoming requests, based on the requested URL, to appropriate OS/390 and z/OS WLM enclaves, each of which has different WLM goals in terms of CPU resource allocation and delay time inside the server. Web content-based QoS classifies and assigns different network QoS levels based also on the requested URL that identifies the kind of HTTP transaction and the content to be delivered (for example, a URL to retrieve a file or a URL to conduct business). The two components — server processing and network QoS — together assure that a Web transaction is assigned an appropriate QoS level that would satisfy end-to-end SLA.

As shown in the first figure, the component that is responsible for URL classification is the Fast Response Cache Accelerator (FRCA). Upon receipt of an inbound HTTP request, FRCA parses the request and retrieves the URL (for example, the Universal Resource Identifier [URI] portion) and invokes the policy component to classify the request along with other criteria (that is, IP addresses, port number, and so on). The URL can be specified in the **`ibm-ApplicationConditionAuxClass`** (the **`ibm-ApplicationData`** attribute) of the policy schema.

Once the request is classified and a matching policy rule is found, the corresponding QoS level is assigned to the returned data of the requested URL. If no matching rule for the URL is found, the QoS level for the associated HTTP connection (if one exists) will be assigned to the returned data. Note that network administrators or service providers can use other policy condition attributes to differentiate QoS levels, for example, taking into account the client's or server's IP address. As a result, the same data that is requested by different clients can be assigned to different QoS levels — clearly a function that enterprises and service providers can leverage in delivering content based on types of clients and users (for example, premium versus regular clients and users).

Sysplex Distributor Policy-Based Routing and Dynamic Load Distribution



One of the major functions released in OS/390 V2R10 and z/OS V1.1 is the Sysplex Distributor function. Briefly, this function presents an S/390 and z900 sysplex as a single server node (that is, one IP address); upon receipt of an incoming connection request (TCP SYN message), Sysplex Distributor will route the request to the most efficient target server for processing. The most efficient server here means a server that has the most available processing resources and good overall network QoS performance (how network QoS performance is monitored at the target server is discussed in the following paragraphs). In other words, the Sysplex Distributor function manages the sysplex as a cluster of nodes, and dynamically monitors and balances workloads among these nodes.

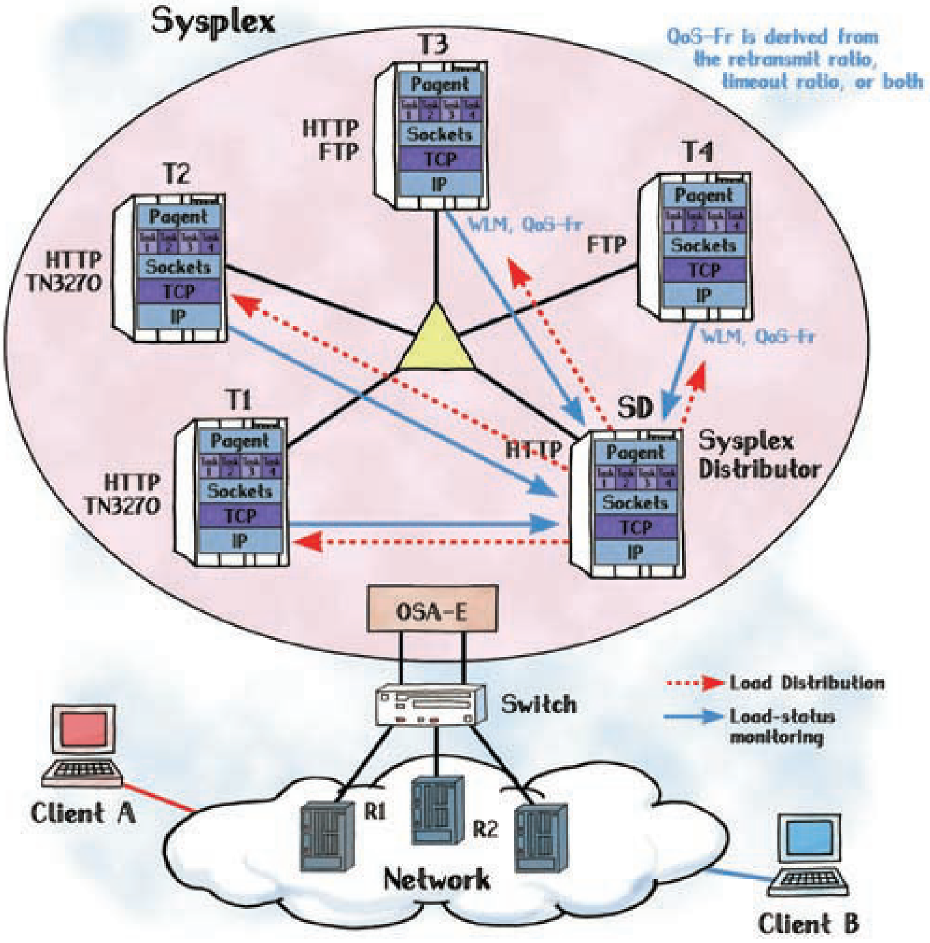
With the policy-based routing feature, Sysplex Distributor can also direct incoming requests to one or more target servers that are designated to serve those requests (in the case where there are more than one identified by policy, the best one is dynamically chosen). This feature can be particularly useful for service providers who want to differentiate a set of clients from other clients, who are assigned to a reserved server for better response time. For example, a premium client would be assigned to the most powerful node within a sysplex to get the best response time. The interaction between Sysplex Distributor and policy is shown in the first figure. The following figure illustrates how Sysplex Distributor with policy-based routing works. In this example, routing policies are defined at the Sysplex Distributor node such that requests from client A will be routed to target nodes T3 or T4 (the more efficient node is chosen at the time of the requests); whereas, requests from client B will be routed only to T1 or T2.

Once a set of potential target servers is identified either by the sysplex routing policy, or based on the main Sysplex Distributor configuration, Sysplex Distributor will choose one server with the best potential for shortest response time, and route the incoming request to it. The criteria for choosing the best target server include the target server's CPU availability and its corresponding network QoS performance. The CPU resource availability is obtained from the WLM, and the network QoS performance is obtained from the Policy Agent. The network QoS performance is monitored on a per-application basis (for example, HTTP and FTP, where each has its own performance status) at each target server node.

The algorithm for choosing the best server can be summarized as follows:

- Sysplex Distributor gets a weight number from WLM, which represents the CPU's availability at a target server node, and a QoS fraction from the Pagent for the corresponding application.
- The WLM weight is multiplied by the QoS fraction (0 to 1, with 1 being the worst).
- The weight that is derived from the QoS fraction is subtracted from the WLM weight.
- The best target server is the one that has the highest weight number.

It's critical to understand the importance of using both the CPU's availability and the network QoS performance in the Sysplex Distributor's load-distribution algorithm, because it is not often done by clustering workload-distribution mechanisms offered on other server platforms. For instance, if a target node has a high CPU resource but is constrained by network bandwidth and another node that has less CPU resource but is not network bandwidth-constrained, it is better to route to the latter target node because it will likely yield better overall end-to-end response time. The following figure depicts an overview of the Sysplex Distributor's workload-distribution algorithm.



Traffic Regulation Management

Traffic Regulation (TR) management is a function that enables network administrators and service providers to prevent two common threats that can bring network services to a halt. One threat is from "greedy" clients that send multiple requests to a server to get better service at the cost of preventing others from satisfying theirs. The other threat is "flooding attack" where a server is bombarded with connection requests such that little useful service can be rendered. This is also known as a type of *denial of service (DoS)* attack. This function is enabled by defining TR policies in the policy configuration file. Note that TR policies are not supported by LDAP in OS/390 V2R10 and z/OS V1.1. TR function is applied on an application basis, and TR policy can be defined as such. However, network administrators have the option to define a single TR policy to be applied to all applications.

The interaction between TR components, including TR Management Daemon (TRMD) and TR component in the TCP/IP stack, and policy components is shown in the first figure. Once TR policy is applied to an application (for example, HTTP), network administrators can perform any of the following options:

- Define the total number of concurrent connections that are allowed for an application (for example, HTTP port 80).
- Use the *Statistics* action to gather information about the normal number of connections to an application server. This can prove useful for monitoring clients' behaviors in accessing the corresponding application or for capacity planning purposes, or both.
- Use the *Limit* action to limit a client to a defined percentage of remaining available connections, denying a client's request when it exceeds this threshold percentage.
- Instead of using the Limit or Statistics action, or both, use the *Logging* action where a client's request that exceeds the percentage threshold won't be denied but is instead logged. This is useful for identifying greedy clients or "testing out" a TR policy, or both.

To continue the S/390 and z900 leadership as a server platform of choice for conducting e-business, future releases of the OS/390 and z/OS software will be enhanced with functions that address key critical areas.

QoS Service-Level Agreement Monitoring

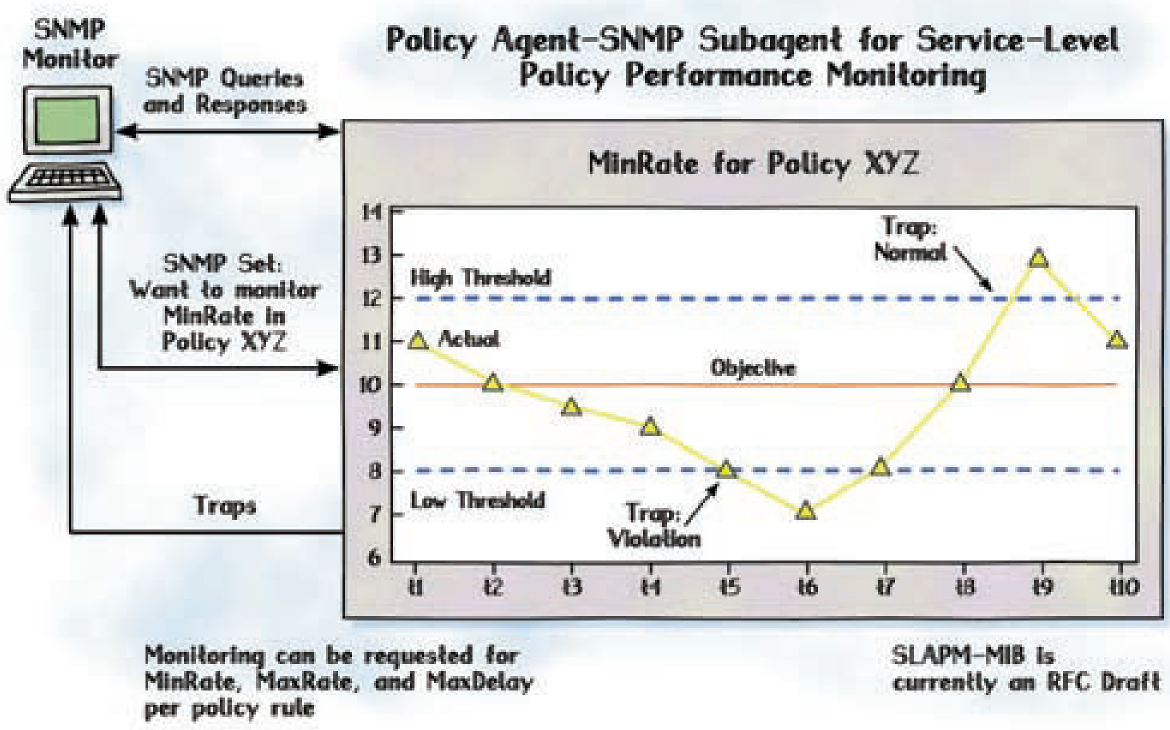
Once QoS levels are defined and enforced, it is important to monitor their performance for any deviations in order to be proactive in addressing clients' SLAs. OS/390 and z/OS provide the SLA Management Information Base (SLA MIB), which can be used to monitor the performance of respective QoS policies. Thresholds can be defined for different QoS performance attributes (for example, MaxRate, MinRate, MaxDelay) such that when a deviation occurs, traps will be sent to the attention of the right Simple Network Management Protocol (SNMP) manager for appropriate actions. The information on a per-policy basis that is reported in the SLA MIB can also be effectively used for accounting and billing data: information such as the bytes or packets sent or received, in-profile vs. out-profile counts. The following figure illustrates how a OS/390 and z/OS SLA MIB subagent monitors the minimum rate with the defined upper and lower thresholds for deviations.

A Look at the Future

To continue the S/390 and z900 leadership as a server platform of choice for conducting e-business, future releases of the OS/390 and z/OS software will be enhanced with functions that address key critical areas: connectivity, availability, scalability, reliability, and security. In particular, the policy-based networking feature is being considered for possible enhancements in areas besides QoS, such as virtual private network (VPN) IP Security Protocol (IPSec), intrusion detection (ID), and so on. Potential enhancements in dynamic load distribution will be able to take into account the network-performance status of each QoS level, in addition to the per-application performance status. Potential extensions to the SLA MIB such that integrated end-to-end delay performance can be taken into account in dynamic load distribution. ■

About the Author

Lap Huynh is currently with the Communications Server for z/OS Design and Strategy Group in RTP. He's been working in different areas of networking for over 15 years. His areas of interest include policy-based networking, Quality of Service, protocol/algorithm design and analysis, an performance evaluation.
LAPH@us.ibm.com


SNMP Monitor
SNMP Queries and Responses
Policy Agent-SNMP Subagent for Service-Level Policy Performance Monitoring
MinRate for Policy XYZ
SNMP Set: Want to monitor MinRate in Policy XYZ
Traps
14
13
12
11
10
9
8
7
6
High Threshold
Actual
Objective
Low Threshold
Trap: Normal
Trap: Violation
t1
t2
t3
t4
t5
t6
t7
t8
t9
t10
Monitoring can be requested for MinRate, MaxRate, and MaxDelay per policy rule
SLAPM-MIB is currently an RFC Draft

NCP and 3745/46

3746 Extended Functions 6 offers functional enhancements to IP, APPN/HPR and DLUR support, improved scalability and performance, increased APPN/HPR and DLUR connectivity, and enhanced network management.

A Message from Bill Cheng

Vice President Marketing, Storage Networking Division, Storage Systems Group

I have asked the Editor of *NCP and 3745/46 Today* Magazine to allot me this space so that I can thank all of our 3745, 3746, and NCP customers for your patience, and especially for staying with IBM even while experiencing delays in delivery of controllers and upgrades.

These delays were caused primarily by logic component availability. Our manufacturing team has worked hard with the suppliers to improve the situation, and I am confident that we will continue to improve.

In addition to improving fulfillment, we have also invested on the development side. In June, we announced 3746 Extended Functions 6 — a new offering for the 3746 Models 900 and 950. Extended Functions 6 consists of important management, networking, and usability enhancements that many of you have told us are important to you. Also, we announced a new release of ACF/NCP (V7R8.1) that complements the 3746 enhancements.

And to further strengthen your ability to manage your networks, we announced a new release of NTuneMON™, a monitoring and tuning tool running under the NetView® program. Also, to further simplify network administration, we have packaged NTuneNCP™ within the Network Control Program. This eliminates the mechanics of ordering and handling another product feature.

Recent statements from non-IBM sources that IBM will no longer continue to support previously discontinued 3745 Communication Controllers, such as the 210 and 410 models, are incorrect. Let me assure you that IBM will continue to protect your investment by providing worldwide maintenance for all the 3745 models.

As you may have seen from our recent announcements, we have made significant improvements in the performance, capacity, manageability, and usability of the 3745 and 3746 platforms. They are being made by IBM in recognition of the enormous investment that our customers have made in SNA applications, and in recognition of the requirement to maintain the stability of mission-critical applications during the transition to e-business environments. I think, as I hope you do, that this record of investment communicates to you, our networking customers, that IBM is your networking partner, and that we will continue to work to protect your past investments to see you through this time of transition to e-business with a focus on timely availability of product, premier technical support, and unparalleled services.

Bill

3746 Extended Functions 6 Offers Enhanced Connectivity, Management, and Networking Functions

By Pierre Planas Comas

The 3746 Nways® Multiprotocol Controller provides a single platform that consolidates IP, SNA, and APPN®/HPR routing over the same transmission media. Extended Functions 6 (Feature 5813) for the 3746 delivers:

Functional enhancements to IP, APPN/HPR and DLUR support

Improved scalability and performance

Increased APPN/HPR and DLUR connectivity

- 50% increase in the number of user sessions per 3746
- 33% increase in the number of control sessions per 3746

Enhanced network management

- New Tivoli® NetView® RUNCMD commands
- Processor load reports using NetView Performance Monitor (NPM)
- Enhancements in the areas of problem management and reporting

These new functions:

- Are available for new and installed 3746s equipped with a Network Node Processor (NNP)
- Are provided as a new licensed internal code option that can be enabled by installing the 3746 Extended Functions 6
- Complement the 3746 Extended Functions 5 (Feature 5812)

3746 Extended Functions 6 offers functional enhancements to IP, APPN/HPR and DLUR support, improved scalability and performance, increased APPN/HPR and DLUR connectivity, and enhanced network management.

APPN/HPR and Dependent LU Requester (DLUR)

30 000 LU-LU data sessions Controlled by the Network Node Processor, instead of 20 000

This enables simpler data-center design. For example, two 3746s (instead of three) can now support 30 000 sessions with full backup capability. The total number of LU-LU sessions that can be routed through the 3746 is now 45 000, including the intermediate sessions established by other network nodes.

This capability is supported also for 3746s configured with the Extended Functions 2, Feature 5802 (Session Services Extensions, Enhanced Topology Management — APPN Option Set 087, and the new Branch Extender and Adaptive Rate-Based Flow and Congestion Control [ARB2] functions).

80 000 SSCP-LU Control Sessions through DLUR Running in the Network Node Processor, instead of 60 000

This allows installations with up to 80 000 defined LUs, (but with much fewer LUs actually in use) to be migrated from NCP control to NNP control without definition changes in the VTAM® program.

This capability is supported also for 3746s configured with the Extended Functions 2 — Feature 5802 (Session Services Extensions, Enhanced Topology Management — APPN Option Set 087, and the new Branch Extender and ARB2 functions).

New Branch Extender option

This new option enables you to operate the 3746 as a Branch Extender node.

This mode of operation can be used to reduce the number of network nodes in the topology of large APPN/HPR networks, and to improve the scalability and performance of APPN/HPR networks. (Network bandwidth and Network Node Processor load are not affected by large topology database updates.) For details of the Branch Extender option, see page 96 of this issue.

Adaptive Rate-Based Flow and Congestion Control (ARB2)

Adaptive Rate-Based Flow (ARB) reduces the risk of frame loss and retransmissions by automatically detecting network congestion over the HPR path and slowing down the HPR traffic. Adaptive Rate-Based Flow and Congestion Control (ARB2), also called "responsive mode," can deliver higher performance than the original ARB. For example, it can provide:

- Faster ramp-up at transmission startup
- Enhanced bandwidth allocation
- Better utilization of high-speed links

A 3746 with ARB2 can now operate as an HPR/RTP end point in conjunction with other HPR/RTP end points, such as VTAM or IBM 221x routers, which support ARB2.

APPN/DLUR support for existing NCP functions (migration to 3746/NNP control)

- Connection of multiple low-entry networking (LEN) nodes having the same CP name (LEN-dependent LUs only). No changes are required in existing LEN nodes and there is no need to define the LEN nodes in Controller Configuration and Management (CCM).
- Support for LU-LU sessions that do not have session pacing, or for LUs that are currently configured without session pacing. No changes are required in the user LU when migrating to 3746/NNP control.
- Inclusion of the link station name (or PU name) in the XID3 frame. When the ESCON® APPN links are activated, the station name is included in the XID3 sent to the Transaction Processing Facility (TPF) end nodes. This allows TPF to differentiate links associated with the 3746 Network Node from those associated with the VTAM/NCP node (for example, an interchange node [ICN]).

Internet Protocol

Same Subnet Option (multiple IP ports to the same subnet)

By default, the IP addresses assigned to the network interfaces must each be in a different network or subnet. This can be changed by enabling the Same Subnet option. This option allows the 3746 IP router to have multiple IP ports, which can be ESCON, token ring, frame relay or X.25, to the same subnet. Multiple IP interfaces to the same subnet enable implementation of:

High-availability configurations (alternate IP routes)

- OSPF Point-to-Multipoint is configured on the IP interfaces, typically for user connections through a frame-relay or X.25 network.
- Next-Hop Awareness is enabled on the IP interfaces, and static routes are defined for the routes that go through the IP interfaces. This enables high-availability configurations over ESCON or frame-relay ports.
- For IP over token-ring, two ports (TIC3s) can be configured in the 3746 for the same physical token-ring LAN.

Load balancing over multiple IP ports

- The Equal-Cost MultiPath routing of the 3746 supports up to four routes of equal cost to the same destination.
- With the Per-Packet MultiPath option, the traffic to such a destination can be spread over multiple IP ports round-robin. This option is now enabled by default.

Next-Hop Awareness

Next-Hop Awareness allows the router to sense whether a neighboring router is active or inactive. When this option is enabled, the router makes a more accurate determination of whether a static route that uses the neighboring router as its next hop will function. It also allows the router to determine over which network interface a static route's next hop can be reached when that next hop is in an IP subnet that is defined on multiple network interfaces. Next-Hop Awareness is supported on frame-relay and ESCON ports.

Next-Hop Awareness over Frame Relay

When static routes are defined over frame-relay ports, this allows the 3746 to update the status, active or not, of the data link connection identifiers (DLCIs) in the OSPF tables. This ensures the use of active DLCIs for the route selection, and enables high-availability network design.

Next-Hop Awareness over ESCON

- Enables the reporting of TPF status, active or inactive, to the neighboring LAN-attached router. This allows end-user connection requests (especially those reconnection requests received after a failure) to be automatically routed through the 3746 to an operational TPF.
- Enables high-availability access to TPF applications, in conjunction with multiple ESCON ports (Same Subnet option).
- No operator intervention is required in case of TPF failure.

Important Migration Planning Information

The new 3746 Licensed Internal Code will be shipped, beginning October 17, 2001, for microcode maintenance and installation with new 3746s and selected 3746 field upgrades, including the 3746 Extended Functions 6. Compatibility of this level of microcode with the current reporting of 3746-900 statistics to NPM requires NCP Version 7 Release 3, or later, with PTF. The NCP PTF must be applied before the installation of the new Licensed Internal Code, otherwise the reporting of all statistics related to 3746-900 resources controlled by NCP will be interrupted.

Details about the required NCP PTF (and about any necessary NPM PTF) will be available by July 31, 2001 at ibm.com/networking/support/3746 (select Downloads).

"Not-So-Stubby-Area" (NSSA)

NSSA allows for external route aggregation, and reduced route table size, specifically in the OS/390® program (TCP/IP), thus resulting in more efficient IP network operations. NSSA, which is is one of the three general classes of area configurations defined in OSPF RFC 1587:

- Provides a mechanism to aggregate route information from many OSPF routers (routes from within a NSSA are summarized by the NSSA Area Border Routers).
- Prevents external routes from being advertised throughout the entire OSPF domain (no exchange of route information between a NSSA and other areas).

For full details of NSSA, see "New 3746 "Not-So-Stubby-Area" Function Addresses Important OSPF Limitations" on page 99.

Network and System Management

Tivoli NetView Program (New RUNCMD Commands)

- Sorts LUs in alphabetic order. This gives the operator faster access to LU information, especially in large 3746 configurations.
- Deletes LUs from APPN directory. This enables the operator to delete unexpected LU names, such as duplicate LUs, which might affect network operations.
- Displays the maximum number of incoming calls accepted on a token-ring port (TIC3) or frame-relay port. This information is provided in both the "Port List" and "Port Details" displays. It allows the network operator to determine whether the number of PUs actually connected to a TIC3 (NPM report) is reaching its maximum.
- Displays session counters, including the new set of SSCP-LU session counters (active sessions, pending sessions, down sessions, total sessions). This allows the operator to accurately monitor the number of SSCP-LU sessions controlled by the NNP.
- Activates a configuration with the option not to re-IML the ESCON channel adapters. When there is no ESCON configuration change, this avoids disruption of the NCP traffic and allows the APPN and IP network to restart more rapidly.

NetView Performance Reporting (NPM)

- Reporting of the processor and memory utilization of the NNP. This enables automatic triggering of a NetView alert and command list when the NNP load exceeds a defined threshold.
- Reporting of processor and memory utilization through the NNP for all the adapter processors (control bus and service processor [CBSP], ESCON processor [ESCP], token-ring processor [TRP], and communication line processor [CLP]). This enables load monitoring and capacity planning for processors that are not controlled by NCP.

VTAM Program

Provides DeactPU and DeactLU command support by the NNP. The NNP supports the DeactPU and DeactLU command facility of VTAM, thus allowing network operators the continued use of existing VTAM commands.

New Controller Configuration and Management (CCM) Commands

- Deletes LUs from APPN directory, equivalent to the NetView RUNCMD command mentioned previously.
- Provides a new set of SSCP-LU session counters, including the counter of sessions in pending state (between ACTLU received from VTAM and negative response received from PU), counter of down sessions, counter of active sessions, and total session counter.

Telnet Session Awareness

A user trying to start a new Telnet session is informed about the IP address of the user who is already in session with the 3746. This allows Telnet users to communicate and coordinate their actions (single Telnet session at a time).

Problem Reporting

- Storing of information (PUs, LUs) displayed using RUNCMD commands for retrieval by IBM support. This reduces the need to recreate problems, avoids potential related disruptions and shortens problem-correction time.
- Fast method of transferring files (problem-determination data) from the 3746 to IBM support. Shortens problem-correction time.
- NetView Alert and automatic call to IBM Support in case of uncontrolled NNP or Service Processor (SP) restart. Complements existing reporting of NNP and SP problems for faster problem resolution. ■

The 3746 Nways® Multiprotocol Controller provides a single platform that consolidates IP, SNA, and APPN®/HPR routing over the same transmission media.

About the Author

Pierre Planas Comas is the 3745/3746 Business Line Manager and is based in La Gaude, France. He is responsible for product-plan contents and product announcements, provides technical marketing support, and is a contributor to marketing documentation and product publications. In his career with IBM, Pierre has worked in various roles with networking hardware and software products and has been involved in the IBM Communications Controller product line for 20 years.
planas1@fr.ibm.com

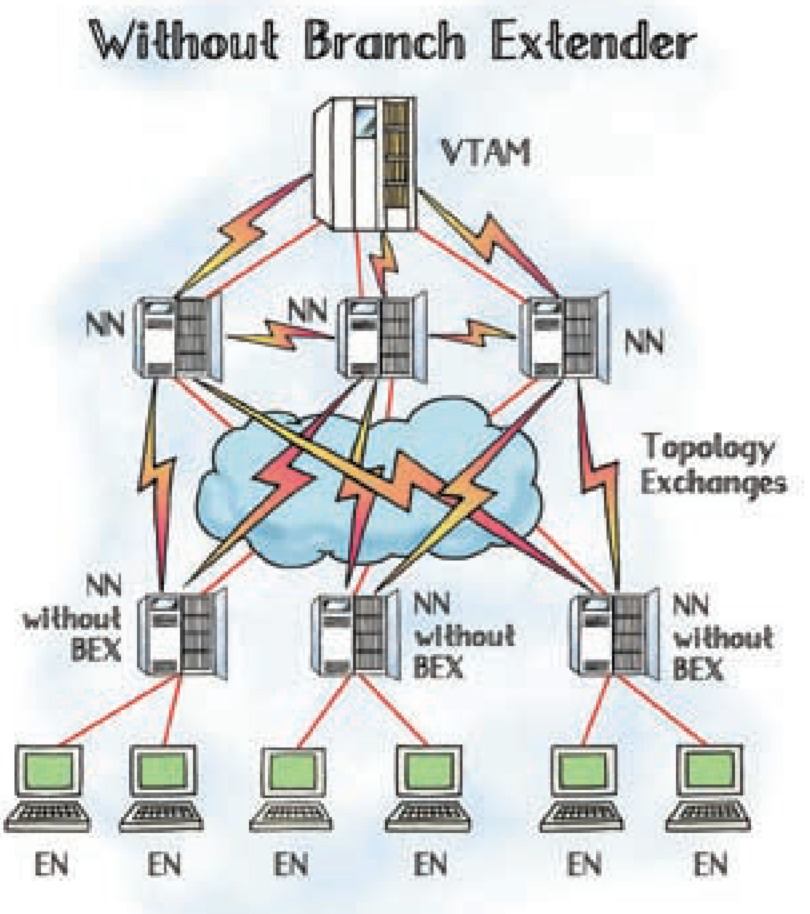
New 3746 Branch Extender Function Offers Performance and Cost Advantages in Large APPN Networks

By Denis Esteve and Pierre Planas Comas

Branch Extender technology helps save on network costs for enterprises that have a large number of branch sites, and have Systems Network Architecture (SNA) applications in the data center and SNA clients in the branches. Branch Extender allows even the largest subarea SNA networks to exploit High-Performance Routing (HPR) cost-effectively. Although primarily targeted at large organizations, Branch Extender can also benefit smaller networks.

Large enterprises often have HPR in the data center; some also have HPR in branch sites. The number of networks with end-to-end HPR connectivity is growing as more HPR products become available. Branch Extender was developed for large enterprises that have many branch sites. Banks, retail chains, car dealers, and insurance agencies typify the kind of enterprise that is structured in this way. The data network connecting the branches of a very large enterprise is often spread over a vast geographic area. In the past, sharing the cost of bandwidth over several sites with multidropped lines was the most economic solution. But more recently, many phone tariffs fell low enough to allow a dedicated line to service each remote branch. In many regions, frame relay has now become the most cost-effective way to interconnect branches. Even with these decreased telecommunication costs, bandwidth may still be the largest single item in the IT budget. Yet bandwidth is essential for operating the networked applications that provide a competitive edge.

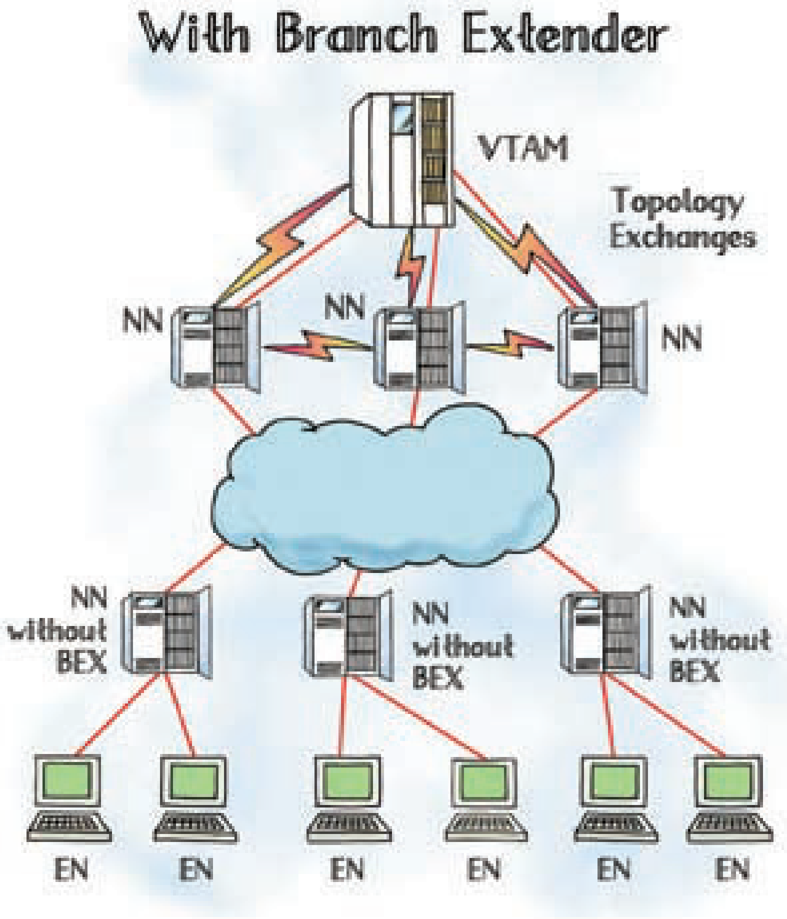
A new 3746 Multiprotocol Controller function called Branch Extender (BEX), included in the new 3746 Extended Functions 6 (Feature 5813), offers improved performance and reduced costs in large networks that use Advanced Peer-to-Peer Networking® (APPN®).



Large enterprise networks can benefit greatly from advanced HPR functions such as automatic resource discovery and route selection. However, the price of dynamic network operation is the need to advertise the network topology or search the network for resources. In a small network, the overhead of these network control messages is insignificant. But in a large network made up of hundreds to thousands of relatively slow lines, even a small overhead might be unacceptable. The 3746 Models 900 and 950 can now operate as Branch Extenders to enable such enterprises to gain the benefits of HPR throughout the entire network without these concerns.

Most of HPR's network control messages are restricted to network nodes (NNs), that is, systems such as front-end processors, routers, or communication servers that direct traffic in the network. PCs or servers that do not direct traffic are end nodes (ENs), and do not send or receive these messages. Previously, any HPR node that routed data between other systems (for example, a 3746 connecting a branch office to an S/390® server) was required to be a network node, and to send and receive such network control messages. See the figure on the previous page.

The Branch Extender filters out network control messages that are unnecessary to the branch. Technically, a Branch Extender poses as a network node with DLUR support to computers, workstations or 3270 terminals in its local branch. At the same time, it poses as an **end node** to network nodes in the wide area network (WAN).



By impersonating an end node, a Branch Extender gains several advantages:

- It does not receive topology data or directory broadcasts from the WAN. Instead, it lets a network node in the WAN choose routes for destinations in the branch.
- It relays data between the branch and the WAN, and enables HPR traffic to flow to end nodes in the branch. See the above figure.
- It can register all of the branch's resources to a directory server in the data center, so that no manual definitions are needed there. Central registration handles adds, moves, and changes automatically. It improves network efficiency by notifying the central directory server of each resource's location. This greatly reduces or even eliminates broadcast searches in the WAN.
- It is automatically discovered by the Tivoli® NetView® for OS/390® program. This enables entry-point/focal-point relationships to be established without configuring them in the NetView program.

Branch Extender significantly increases the number of branches that can participate in a single network. It was specifically designed to fulfill the needs of large single-enterprise networks. In contrast, **border node,** an existing HPR scalability feature of VTAM, has firewall features such as security, accounting, and filtering that are useful when interconnecting different enterprises. Branch Extender does not need firewall features because it is intended for trusted environments.

Branch Extender avoids the need to divide a single-enterprise network, thereby substantially reducing line costs compared with other solutions available today. It also makes a large network act like a small network by sending fewer network control messages. These two types of savings are cumulative.

Branch Extender enables individual APPN and HPR networks to grow very large. It uses cost-effective methods to reduce network control messages and switch traffic between branches. It also maintains the key functions of HPR — including congestion control and class of service. This enables full utilization of available bandwidth, while quickly and efficiently forwarding time-critical traffic. ■

About the Authors

Denis Esteve joined the IBM 3746 Software Development team at La Gaude, France, in 1989. He has developed some parts of the 3746 Token-Ring and Frame-Relay Outboard Data Link Control (ODLC). He was Development Team Leader for the 3746 X.25 ODLC and is now System Designer and Development Project Manager for the 3746-9x0.
esteve@fr.ibm.com

Pierre Planas Comas is also the author of “3746 Extended Functions 6 Offers Enhanced Connectivity, Management, and Networking Functions” on page 92 of this edition of *NCP and 3745/46 Today.*

New 3746 “Not-So-Stubby-Area” Function Offers Performance Advantages in Large OSPF Networks

By Denis Esteve and Pierre Planas

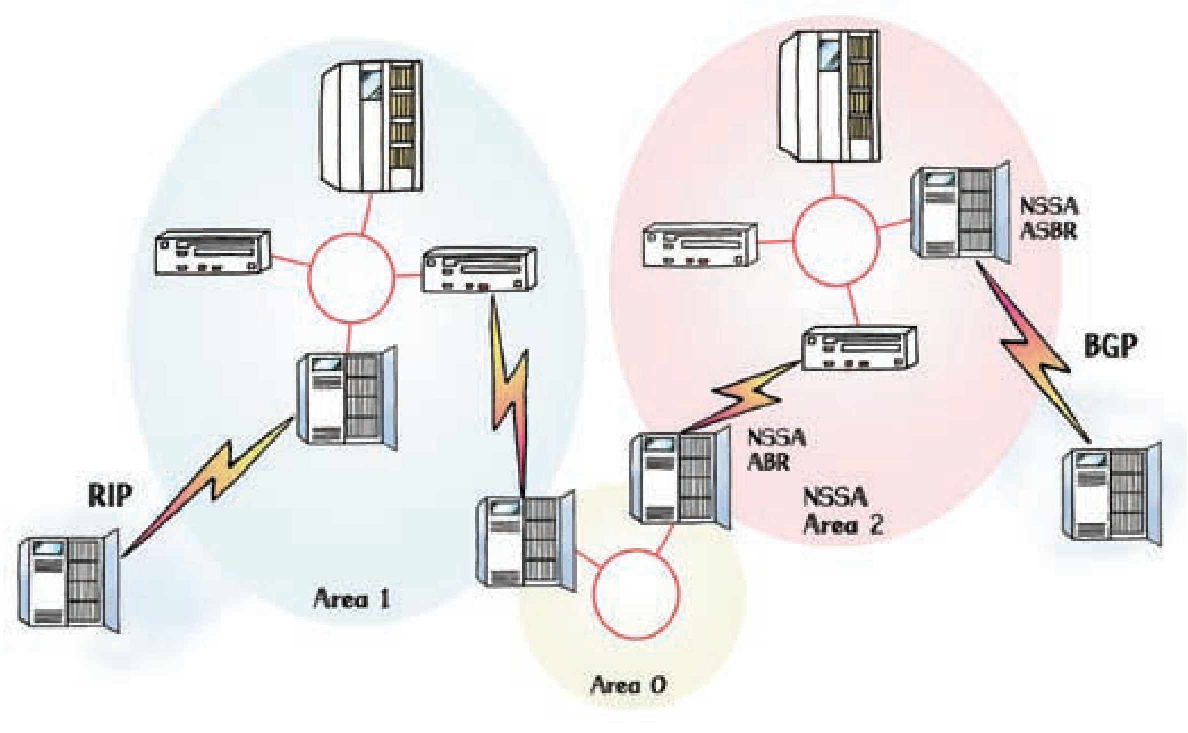
A new 3746 Nways® Multiprotocol Controller function called “Not-So-Stubby-Area” (NSSA), included in the new 3746 Extended Functions 6 (Feature 5813), offers significant improvements in large IP router networks that use Open Shortest Path First (OSPF).

For the 3746 Models 900 and 950, the primary motivation for implementing NSSA was the need for external route aggregation to limit Route Table size, specifically in OS/390® (TCP/IP). Before, there was no mechanism to aggregate external advertisements from many OSPF routers, nor any way to keep external routes from being flooded throughout the entire OSPF routing domain.

NSSA, one of the three general classes of area configurations defined in OSPF RFC 1587, addresses both these limitations and offers additional benefits:

- A mechanism to aggregate route information from many OSPF routers (routes from within a NSSA are summarized by the NSSA Area Border Routers).
- No advertisement of external routes throughout the OSPF domain (no exchange of route information between a NSSA and other areas).
- A separate default route to the firewall per area. In many enterprise networks, it is desirable to have a separate Internet firewall in each location. You can accomplish this using NSSAs without having the area-specific default route advertised throughout the routing domain.
- Extranet implementation. Multiple enterprises could use an OSPF backbone with attached NSSAs (one per enterprise) to implement an extranet (network administered by multiple cooperating parties) and still have flexible routing within the NSSA.

The redistribution into an NSSA creates a special type of link-state advertisement (LSA) known as type 7, which can exist only in an NSSA. An NSSA autonomous system boundary router (ASBR) generates this LSA and an NSSA area border router (ABR) translates it into a type 5 LSA, which gets propagated into the OSPF domain. The figure demonstrates this principle.



In the figure, Area 2 is defined as a stub area. Border Gateway Protocol (BGP) routes cannot be propagated into the OSPF domain because redistribution is not allowed in the stub area. However, if you define Area 2 as an NSSA, you can inject BGP routes into the OSPF NSSA domain by creating type 7 LSAs. Redistributed Routing Information Protocol (RIP) routes will not be allowed in Area 2 because NSSA is an extension to the stub area. The stub area characteristics still exist, including no type 5 LSAs allowed. Type 5 LSAs are not allowed in NSSAs, so the NSSA ASBR generates a type 7 LSA instead, which remains within the NSSA. This type 7 LSA gets translated back into a type 5 by the NSSA ABR.

There are two flavors of NSSA, just like stub areas. There are NSSAs that block type 5, but allow type 3 and 4 LSAs, and there are NSSA "totally stub areas," which allow only summary default routes and filter everything else.

There are two ways to have a default route in an NSSA. When you configure an area as an NSSA, by default, the NSSA ABR does not generate a default summary route. In the case of a stub area or an NSSA totally stub area, the NSSA ABR does generate a default summary route. ■

About the Authors

Denis Esteve and **Pierre Planas Comas** are also joint authors of "New 3746 Branch Extender Function Offers Performance and Cost Advantages in Large APPN Networks" on page 96 of this edition of *NCP and 3745/46 Today.* Pierre Planas Comas is also the author of "3746 Extended Functions 6 Offers Enhanced Connectivity, Management, and Networking Functions" on page 92 of this edition.

New ACF/NCP and SSP Enhancements Underscore IBM's Commitment to SNA

By Preston Johnston

IBM is firmly committed to supporting SNA, as demonstrated by the latest releases of NCP (V7R8.1) and SSP (V4R8.1).

We are committed because we fully understand the importance of SNA networks and the dependency that you, our customers, have on NCP to manage your critical information assets. This article provides you with an overview of the new NCP and SSP functions, many of which are the direct results of your suggestions.

Network Management Enhancements

NTuneNCP Shipped with SSP and NCP

For years, NTuneNCP™ was available only as an NCP feature. Then, last year, we included NTuneNCP at no additional charge with NTuneMON™ V3R1. Now, we are including it with NCP V7R8.1 and SSPV4R8.1, which will simplify the installation of NTuneNCP and make it easier for this valuable management tool to be propagated into networks. NTuneMON, when used by itself, is a great product to monitor SNA networks, but the addition of NTuneNCP will now allow you to make changes to network parameters without having to regen NCP, thus giving you greater flexibility and control over your network.

Enhanced 3746-900 TIC3 and NCP Flow Control for Token-Ring Networks

For token-ring environments in which both subarea and peripheral data traffic are supported (ECLTYPE=PHYS,ANY) and NCP is becoming congested, we have modified NCP to allow poll responses to a subarea station and continue to accept subarea traffic until NCP reaches critical congestion (CWALL). Previously, the subarea station's connections were broken as the station's retry limit was reached because NCP did not respond, creating a disruptive multiple-session outage, particularly if the TG had a single link and multiple virtual routes.

Maintaining 3746-900 Sessions during NCP Congestion

One of the flow control mechanisms used by NCP is to deactivate a physical line (IOH Halt Cause Code = 8412) attached to a 3746-900 when service clear has been withheld for more than 5 minutes. This prevents a deadlock condition from occurring. However, there is no guarantee that issuing an IOH to a particular link will either prevent or resolve a deadlock condition, and the IOH could result in many sessions being unnecessarily terminated. By no longer issuing IOHs to deactivate a physical line, NCP V7R8.1 allows 3746-900 sessions to be maintained and normal data flow to resume once NCP congestion has cleared.

Enhanced 3746-900 and NCP Flow Control for Token-Ring Subarea Connections

If congestion causes the outboard data link control (ODLC) pacing window to close, data can be discarded and receive not ready (RNR) responses will be sent to adjacent link stations. After congestion has cleared and the ODLC window opens, the combination of RNRs being sent and the discarded frames being retransmitted can cause throughput to be degraded. In NCP V7R8.1, we made the ODLC pacing window larger to accept more data frames, thus reducing the need for RNRs and retransmissions, and improving the chances that throughput will not be reduced.

Transmission Group Alert Enhancement

When a Transmission Group is congested or hung and throughput has degraded, NCP generates an alert notifying the owning host SSCP of the condition. In the alert, there is message referring the host operator to NCP documentation for additional information. We have made this reference more specific, and it now refers to the NCP/SSP/EP Diagnostic Guide, which provides detailed information about this alert condition.

Serviceability Enhancements

SNA networks and NCP have historically had a reputation for stability and dependability. An important element of this reputation has been our commitment to service excellence. To continue this excellence, we have included two service aid enhancements in NCP V7R8.1.

NCP Dump Enhancement

On rare occasions, the first record of an NCP dump, which contains information needed by the SSP dump formatter, is corrupted and the dump cannot be formatted. The usual remedy is to take another dump, but this is time-consuming and slows progress towards problem resolution. We modified the dump formatter so that it continues to process the dump file and to print it in hex format. This will allow IBM service personnel to continue problem investigation until a second, uncorrupted dump can be taken.

We are committed because we fully understand the importance of SNA networks and the dependency that you, our customers, have on NCP to manage your critical information assets.

Interactive Problem Control System Enhancements

The interactive problem control system (IPCS) has been used for years to analyze NCP dumps. In V7R8.1 we have made a number of enhancements to IPCS that provide more detailed failure data or user options to speed up the dump analysis process. New user options include the ability to:

- Select full or partial analysis of NCP slowdown first failure data capture (FFDC) information
- Choose between displaying control blocks for all interface addresses or for only a specific interface address
- Display current CCU and buffer utilization as well as histogram data, and expand dump analysis to include NCP packet switching interface (NPSI) resources

What About Our Future Commitment?

Although the era of major NCP and SSP enhancements has passed, it is our objective to continue to offer you incremental enhancements, because we understand the importance of NCP to you. We encourage you to continue suggesting enhancements as your needs evolve, and we remain committed to being your partner in managing your critical information assets on SNA networks. ■

About the Author

Preston Johnston joined IBM in 1969, retired in 1997 and returned part-time in 1998. Preston's first involvement with NCP was in 1971 with testing of the initial releases on NCP. After many years' absence from NCP, he has returned and is now part of the Business Line Management organization.
prestonj@us.ibm.com

What's New in NTuneMON

By Carl Marlowe

The latest release of NTuneMON™ (V3R2) brings new features as well as changes designed for ease of use and installation.

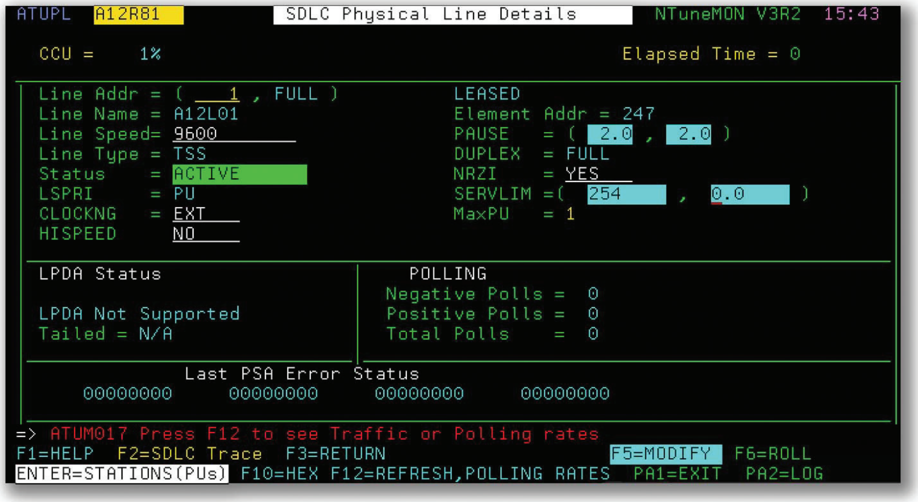
Starting with Version 3, the tuning feature is no longer a priced feature but is included in the base price of the product. To simplify installation, we are including the tuning feature in the current release of NCP. It will no longer be necessary to link NCP and NTuneMON libraries to access the tuning feature. We also recognize that some customers might have earlier versions of NCP, so we are still including the tuning feature with NTuneMON as well. For these users, the installation process will not change.

Prior releases of NTuneMON have included a number of new service aids. In this release, we have added two new service features. The following new functions have been added to the ATUSX panel:

- The capability for NTuneMON to start, stop, and modify the Channel Adapter IOH (CAIOTRC) trace.
- A new module name search function (Service Level and Address Display).

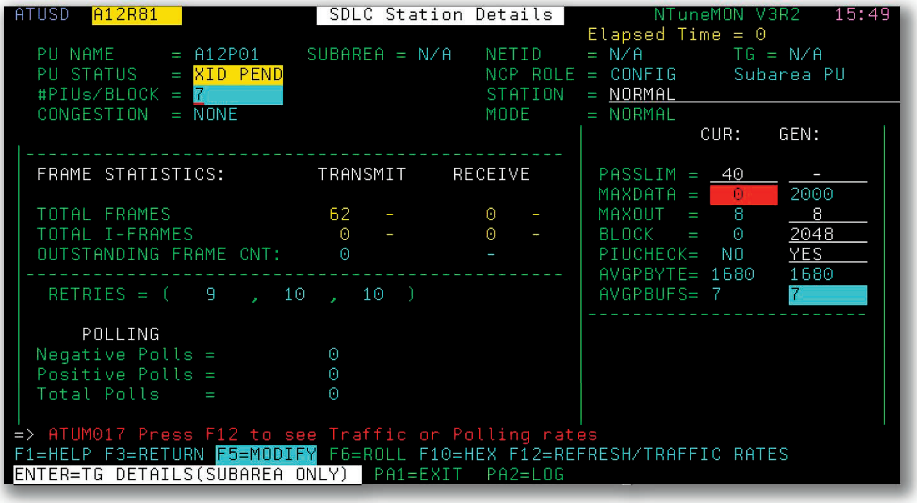
For details of these two new NTuneMON functions, see "New and Enhanced NTuneMON and SSP Functions Add to NCP's Reputation for Serviceability" on page 105 of this edition *of NCP and 3745/36 Today.*

We have also added one new display-only field. MAXPU, the maximum number of physical units allowed on a line, will be displayed on the SDLC Physical Line Details panel (ATUPL).



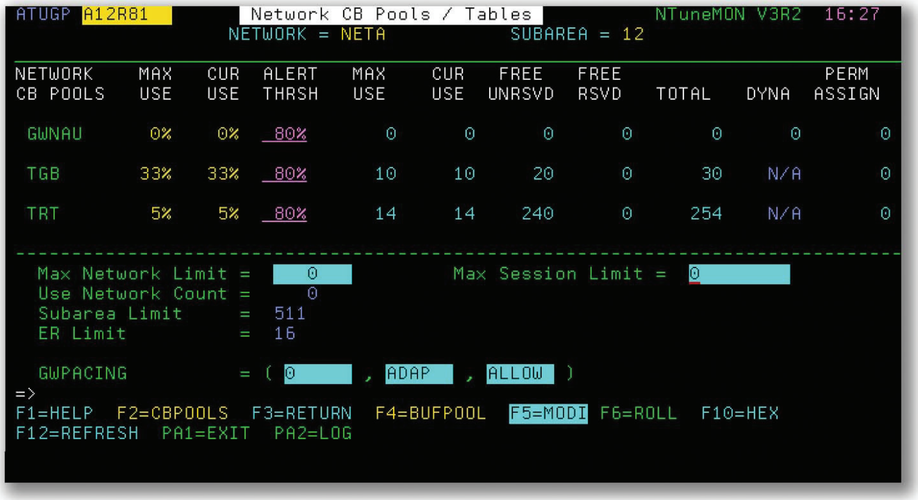
```
ATUPL  A12R81            SDLC Physical Line Details       NTuneMON V3R2  15:43

  CCU =     1%                                           Elapsed Time = 0

  Line Addr = (    1 , FULL )             LEASED
  Line Name = A12L01                      Element Addr = 247
  Line Speed= 9600                        PAUSE    = (  2.0 ,  2.0 )
  Line Type = TSS                         DUPLEX   = FULL
  Status    = ACTIVE                      NRZI     = YES
  LSPRI     = PU                          SERVLIM =(  254     ,  0.0      )
  CLOCKNG   = EXT                         MaxPU    = 1
  HISPEED     NO

  LPDA Status                   POLLING
                             Negative Polls =  0
  LPDA Not Supported         Positive Polls =  0
  Tailed = N/A               Total Polls    =  0

                   Last PSA Error Status
      00000000      00000000      00000000      00000000

=> ATUM017 Press F12 to see Traffic or Polling rates
F1=HELP  F2=SDLC Trace  F3=RETURN                    F5=MODIFY  F6=ROLL
ENTER=STATIONS(PUs) F10=HEX F12=REFRESH,POLLING RATES  PA1=EXIT  PA2=LOG
```

We have added three new fields that you can display and modify. As a result of a request to display the average poll bytes (AVGPB) parameter, we are displaying both the current and maximum (genned) values in terms of bytes and buffers. Only the maximum value can be changed. These fields will be on the SDLC Station Details panel (ATUSD).



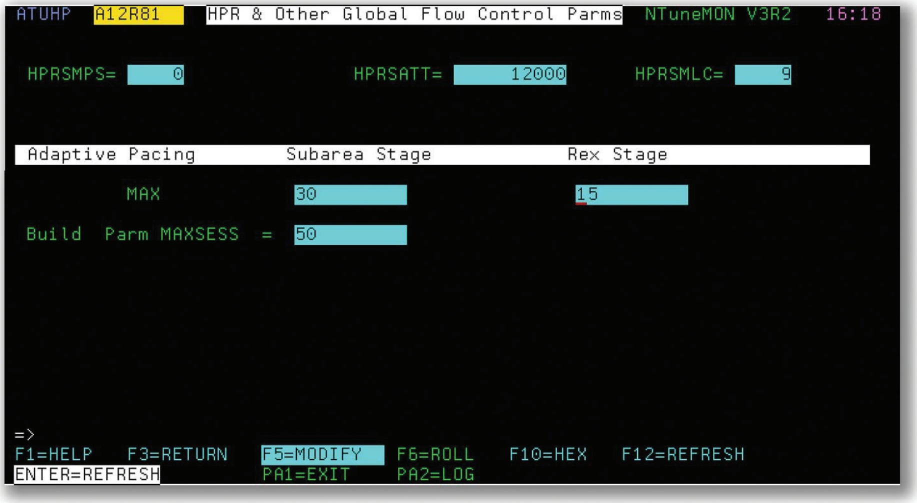
```
ATUSD  A12R81              SDLC Station Details          NTuneMON V3R2   15:49
                                                  Elapsed Time = 0
  PU NAME      = A12P01    SUBAREA = N/A   NETID     = N/A        TG = N/A
  PU STATUS    = XID PEND                  NCP ROLE  = CONFIG     Subarea PU
  #PIUs/BLOCK  = 7                         STATION   = NORMAL
  CONGESTION   = NONE                      MODE      = NORMAL
                                                             CUR:   GEN:
 ---------------------------------------------------
  FRAME STATISTICS:           TRANSMIT    RECEIVE        PASSLIM =  40      -
                                                         MAXDATA =   0    2000
  TOTAL FRAMES                  62   -       0   -       MAXOUT  =   8      8
  TOTAL I-FRAMES                 0   -       0   -       BLOCK   =   0    2048
  OUTSTANDING FRAME CNT:         0           -           PIUCHECK=  NO    YES
 ---------------------------------------------------     AVGPBYTE= 1680   1680
   RETRIES = (   9   ,  10   ,  10   )                   AVGPBUFS= 7      7
                                                        -------------------------
      POLLING
  Negative Polls =            0
  Positive Polls =            0
  Total Polls    =            0

=> ATUM017 Press F12 to see Traffic or Polling rates
F1=HELP F3=RETURN F5=MODIFY F6=ROLL F10=HEX F12=REFRESH/TRAFFIC RATES
ENTER=TG DETAILS(SUBAREA ONLY)  PA1=EXIT  PA2=LOG
```

The SESSLIM parameter from either the BUILD or NETWORK statements will be displayed on the Network CB Pools/ Tables panel (ATUGP). SESSLIM is the maximum number of sessions that can be assigned to an address in the network. You can modify this field.



```
ATUGP A12R81            Network CB Pools / Tables             NTuneMON V3R2   16:27
                       NETWORK = NETA              SUBAREA = 12
----------------------------------------------------------------------------------
NETWORK     MAX   CUR  ALERT    MAX    CUR   FREE   FREE                    PERM
CB POOLS    USE   USE  THRSH    USE    USE  UNRSVD  RSVD   TOTAL    DYNA   ASSIGN

 GWNAU       0%    0%   80%       0      0      0      0       0       0       0

 TGB        33%   33%   80%      10     10     20      0      30     N/A      0

 TRT         5%    5%   80%      14     14    240      0     254     N/A      0

----------------------------------------------------------------------------------
  Max Network Limit =      0             Max Session Limit =  0
  Use Network Count =      0
  Subarea Limit     =   511
  ER Limit          =   16

  GWPACING          = ( 0       , ADAP     , ALLOW   )
=>
F1=HELP  F2=CBPOOLS  F3=RETURN  F4=BUFPOOL  F5=MODI F6=ROLL  F10=HEX
F12=REFRESH  PA1=EXIT  PA2=LOG
```

The HPR & Other Global Flows panel (ATUHP) will display the MAXSESS parameter from the BUILD statement. MAXSESS is used to set the maximum sessions for any logical units (LUs) that are genned but do not have a MAXSESS value coded on the LU statement. You can modify this field as well.



```
ATUHP  A12R81      HPR & Other Global Flow Control Parms  NTuneMON V3R2    16:18

 HPRSMPS=       0               HPRSATT=     12000        HPRSMLC=      9


 Adaptive Pacing              Subarea Stage             Rex Stage

              MAX             30                        15

 Build  Parm MAXSESS  =       50








=>
F1=HELP    F3=RETURN    F5=MODIFY   F6=ROLL    F10=HEX    F12=REFRESH
ENTER=REFRESH           PA1=EXIT    PA2=LOG
```

We have also made ease-of-use enhancements to NTuneMON. On rare occasions, it has been necessary to change the controller timeout value that was hard-coded in NTuneMON. This value has been placed in the ATuneLST CLIST as PM.TMOUT for easier access. Its default value of 25 seconds has not been changed. ■

About the Author

Carl Marlowe joined IBM in 1980 and has worked as a development programmer on various products for the finance and manufacturing industries. He joined the NCP Design and Development Team in 1997 and is currently working on the NTuneMON program.
cmarlowe@us.ibm.com

New and Enhanced NTuneMON and SSP Functions Add to NCP's Reputation for Serviceability

By Vernel Shaw

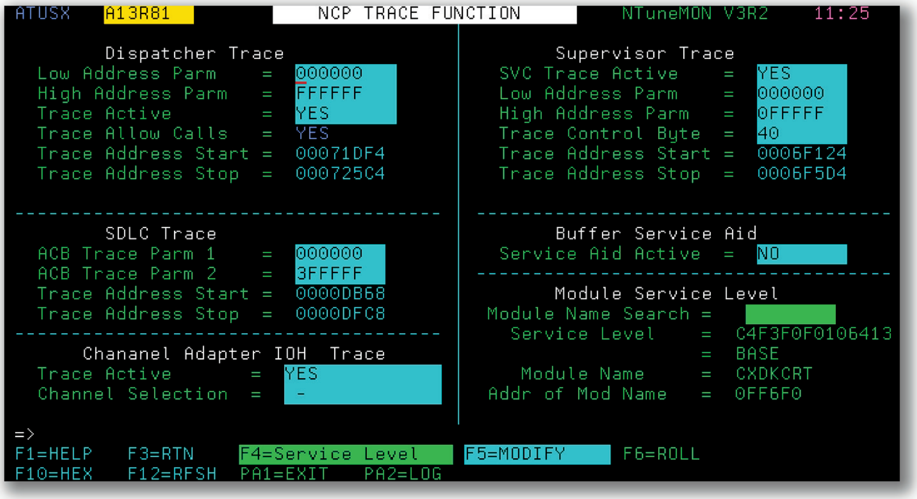
The latest releases of NCP (V7R8.1), SSP (V4R8.1) and NTuneMON™ (V3R2) provide:

- Two new service functions to the NTuneMON ATUSX panel. These are "CAIOTRACE" and "Service Level and Address Display."
- Three enhanced SSP IPCS CLISTs.
- New interactive problem control system (IPCS) CLISTs.
- Enhanced multilink transmission group (MLTG) Performance Degraded Alert reporting to provide you with references to additional documentation needed to analyze transmission group (TG) performance problems.

New Service Functions on the NTuneMON ATUSX Panel

The Channel Adapter I/O Halfword Trace (CAIOTRACE) function allows NCP to trace the flow of IOH instructions done for the 3745 hardware channel adapter. This trace can provide vital information needed to isolate problems between NCP and the Channel Adapter. You will need the latest release of NCP (V7R8.1) to activate this function. You can activate the trace by coding CAIOTRC=YES in the BUILD statement (as in prior releases), from the Maintenance and Operator Subsystem (MOSS) console, or by using the NTuneMON ATUSX panel.

NCP has an excellent serviceability record. Over the life of the product, we have been relentless in our efforts to build service aids into the product that allow NCP Support and users to quickly diagnose problems. Currently, there are more than 45 such functions.



```
ATUSX    A13R81         NCP TRACE FUNCTION          NTuneMON V3R2    11:25

        Dispatcher Trace                          Supervisor Trace
  Low Address Parm     =  000000          SVC Trace Active     =  YES
  High Address Parm    =  FFFFFF          Low Address Parm     =  000000
  Trace Active         =  YES             High Address Parm    =  0FFFFF
  Trace Allow Calls    =  YES             Trace Control Byte   =  40
  Trace Address Start  =  00071DF4        Trace Address Start  =  0006F124
  Trace Address Stop   =  000725C4        Trace Address Stop   =  0006F5D4

--------------------------------------  -------------------------------------
        SDLC Trace                                Buffer Service Aid
  ACB Trace Parm 1     =  000000          Service Aid Active   =  NO
  ACB Trace Parm 2     =  3FFFFF         -------------------------------------
  Trace Address Start  =  0000DB68                Module Service Level
  Trace Address Stop   =  0000DFC8       Module Name Search =
--------------------------------------    Service Level      =  C4F3F0F0106413
      Chananel Adapter IOH  Trace                            =  BASE
  Trace Active        =  YES               Module Name       =  CXDKCRT
  Channel Selection   =   -             Addr of Mod Name     =  0FF6F0

=>
F1=HELP    F3=RTN     F4=Service Level   F5=MODIFY     F6=ROLL
F10=HEX    F12=RFSH   PA1=EXIT   PA2=LOG
```

The new module name search function (Service Level and Address Display) allows a module's service level to be displayed without requiring an NCP dump. This new function is intended to be used with the help of IBM Service. The service representative will provide you with a module name to search for, and NTuneMON will return the service level and address of the module.

Previously, when NCP Support needed to verify the maintenance level of a specific module in an operating NCP, a dump of that NCP was required. This new function not only provides service-level verification that can be vital when troubleshooting, but in some situations it will allow support personnel to write TRAPs or ZAPs for the NCP without requiring a dump of NCP storage.

Enhanced SSP IPCS CLISTs

We enhanced the SSP IPCS CLISTs Internet Protocol (IFWIIP), Slowdown (IFWISLOW) and VRB (IFWIVRB). IFWIIP now provides a more comprehensive picture of the IP router contained within the NCP. We improved IFWISLOW so that FFDC information is provided prior to the lengthy analysis process done by the CLIST. In addition, we enhanced the IFWIVRB CLIST so that you have the option of displaying all VRB control blocks. Prior releases displayed only the VRBs that were active.

New IPCS CLISTs

We also included new IPCS CLISTs in this release. Performance Measurement Facility Display (IFWIPMF) will display the Performance Measurement Facility control block information. Storage Key Display (IFWISTK) will allow you to display all of the 3745 storage keys, or provide a storage key for a user-requested storage address. We also included a number of CLISTs to support NPSI.

Enhanced MLTG Performance Degraded Alert Reporting

Mixed-media multilink transmission groups (MMMLTGs) are transmission groups that can contain one or more links. Up to 255 links can be included in an MMMLTG. The links can be any combination of token-ring, SDLC, ISDN and frame-relay lines. A typical MMMLTG in today's environment will transverse a wide area network (WAN) containing multiprotocol components such as routers, converters, switches and hubs.

With the need for high availability in today's networks, it is important for you to be notified of MMMLTG performance problems as soon as possible. MLTG Performance Degraded Alert reporting is a vehicle for such notification. It has been improved to provide your with references to additional documentation needed to analyze the TG performance problem.

When an MMMLTG is experiencing degraded throughput, NCP will send a generic Alert to assist you in identifying the TG that is hung or potentially hung. NCP considers the throughput to be degraded when either of the following conditions lasts for more than the value of the TGTIMER keyword on the BUILD or NETWORK statement:

- The TG sweep function remains continuously active (see Note 1).
- The TG has path information units (PIUs) on the Resequence Queue and the next expected Inbound TG Sequence Number does not change (see Note 2).

This is an excellent tool for obtaining information on TG problems, especially now that many TGs cross WANs that may contain a variety of components.

The TG Performance Degraded Alert now refers you to the NCP Diagnosis Guide, which provides a detailed explanation of the data in the Alert.

The Service Aids described in this article are evidence of our commitment to NCP serviceability. You can be confident that where future serviceability needs exist, the NCP team will ensure that NCP remains a leader in serviceability.

You can be confident that where future serviceability needs exist, the NCP team will ensure that NCP remains a leader in serviceability.

Notes:

1. The sweep function suspends the transmission of PIUs over the MMMLTG until all PIUs in transit are acknowledged at the DLC level. The sweep function is invoked for the following conditions:
   - When sending the following control PIUs: Activate Virtual Route and its response, Deactivate Virtual Route and its response, Activate Explicit Route, Activate Explicit Reply, Explicit Route Operative, and Explicit Route Inoperative. This prevents these PIUs from overtaking each other and any other session PIUs that may precede them.
   - When TG sequence number field rollover occurs.
   - When a Receiver Not Ready (RNR) is received on a link that indicates that the adjacent subarea is congested.
2. Session protocol requires that all requests and responses arrive at their destinations in the same order in which they were sent from the session origin. To maintain this order, a TG sequence number is placed in the transmission header of each PIU that is sent across the MMMLTG.

At the receiving end of the MMMLTG, this sequence number is checked against the expected sequence number. If the received PIU sequence number is higher than the expected sequence number, the PIU is placed on the resequence queue until the PIU with the expected sequence number is received. PIUs are then inserted in the correct sequence and sent on to the correct destination. ■

About the Author

Vernel Shaw joined IBM in Memphis, Tennessee in 1970 as a Customer Engineer. He has worked in various customer support organizations during his career with IBM. He has been a member of the NCP L2 support team since 1983.
vernelsh@us.ibm.com

Speed Your NCP/EP/NPSI Problem Resolution by Helping Us to Help You

By Sharon Wick

Do you want to shorten the time you spend on the phone with the Support Center next time you have an NCP/EP/NPSI problem? Support Center personnel are eager and able to help you, and if you prepare for your call using the following guidelines, your problems will be resolved more quickly and more smoothly.

Prior to calling the Support Center:

1. Determine the release, version, and service levels of NCP or related products you are running.
2. To the best of your ability, validate that the problem is in NCP or a related product, and identify the component in which the problem occurred.
3. Look for and note the first symptom of the problem.
4. Determine whether the problem can be recreated.
5. If the problem has just appeared, determine what has changed since the system ran successfully.
6. Determine the problem type based on the table "Determining the Problem Type" in Chapter 2 of the *Diagnosis Guide.*
7. Use the Maintenance and Operator Subsystem (MOSS) report tool to list the chronology of error activity. The IBM 3745 MOSS maintains an error log called the box event record (BER) file. If an error is detected in the communication controller or the program, a record is stored in the BER file.
8. Check the NCP error and statistics.

Record maintenance statistics (RECMS) records are created for the host. These records are sent to a data set called LOGREC in the host and to a network problem determination application, if one is installed. To get a printout of the records kept in LOGREC, use the Environmental Record Editing and Printing Program (EREP).

If you have the NetView® program or its equivalent, check network problem determination aid (NPDA) or equivalent for alerts, error messages, inoperative messages, and so on. If these messages do not help you resolve the problem, then have the information available for IBM support personnel.

Some Hints for Specific Error Conditions

If there is an **error message,** make sure that the complete message text is available. If there is a group of related messages, make sure that complete text from all message is available.

If an **abend** occurs, take one of these actions:

- If the abend is in the NCP or EP, dump the NCP, obtain the abend code from the dump, and search on the IBMLink™ service (**ibm.com**/ibmlink) for symptoms to see if this is a known problem that has a PTF available. Using the *Diagnosis Guide,* perform diagnostic steps.
- If the abend is in the NCP Packet Switching Interface (NPSI), refer to the *NPSI Diagnosis Guide.* If you need additional assistance, report the problem to NCP/EP/NPSI support.
- If the abend is in user-written code, obtain on-site assistance.

If devices are connected through the NCP and a **sense code** is reported, find the NCP (as opposed to the VTAM®) sense code. Use *NCP/SSP/EP Messages and Codes,* SC31-6222 for sense codes issued by NCP.

If an **activate** or **deactivate failure** occurs when a resource fails to repond or returns an exception response to an activate or deactivate session request:

- Make sure that the device is powered on.
- Check for lack of resources (gen more resources or free existing ones if necessary).
- Check for a configuration mismatch and correct it if one exists.
- Run traces:
  a. Start the appropriate traces (PIU, NCP line trace, SIT trace, and so on).
  b. Recreate the problem.
  c. Stop the traces.
  d. Format the traces using ACF/TAP.
- Dump the NCP, the access method (if involved), and the Communication Subsystem (CSS) (if involved). If the failing resource is a remote NCP, also dump that NCP.
- If virtual routes (VRs) are blocked, refer to the *Diagnosis Guide,* which contains many steps and procedures for various hung-session symptoms. The *NCP and EP Reference,* LY43-0029 has additional information. One tip is to remember to obtain a dump at both ends of the communication (both NCPs or, if the VTAM program is involved, a dump of VTAM as well). The NTuneNCP™ and NTuneMON™ programs, if available, are also valuable for monitoring of blocked VRs.

Note: This is a small subset of the many possible conditions that might occur when running NCP. It is meant to serve as a starting point for obtaining information needed to diagnose or obtain assistance in diagnosing problems associated with the IBM products.

The NTuneMON™ program can be a great tool for assistance in diagnosing problems. Refer to the *NTuneMON User's Guide,* SC31-6266 and to "New and Enhanced NTuneMON and SSP Functions Add to NCP's Reputation for Serviceability" on page 105 of this issue for new enhancements to this tool.

The NCP/EP/NPSI Support Center is eager to partner with you to resolve any problems associated with these products. Together, we can resolve issues more quickly and effectively. We look forward to our joint endeavors.

For more detailed information on resolving your problem, refer to the *Network Control Program, System Support Program, Emulation Program Diagnosis Guide,* LY43-0033. ■

About the Author

Sharon Wick joined the IBM Federal System Division in 1966. After working in Washington, DC, she relocated to Boca Raton and worked in the Series/1 area. She has since worked in CICS level 2 and TCP/IP level 2, and is now a member of the NCP level 2 team. She is a retired IBMer and is working as a contractor.
sswick@us.ibm.com

NCP Trace Impact on CCU Utilization

By Sue DeMarrais

Your network is critical to your business — but maintaining a robust communications network occasionally requires system diagnostics.

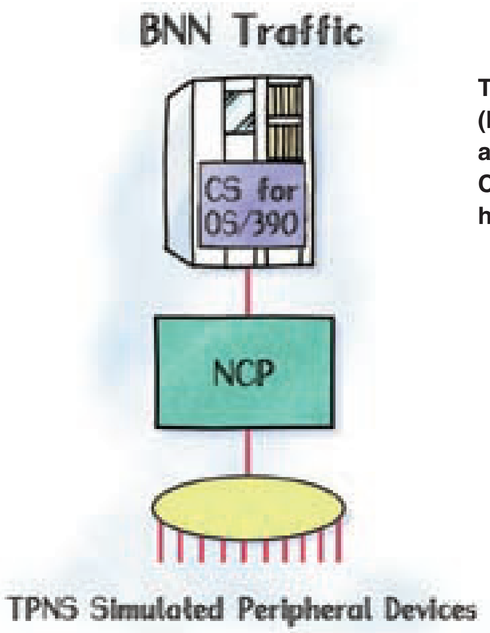
NCP supports several traces to accomplish the task of problem determination and data-flow analysis, such as the NCP line trace, NTRI SNAP trace, ODLC SNAP trace, NCP transmission group trace, NCP generalized PIU trace, channel adapter I/O trace, VTAM® buffer contents trace, and scanner interface trace (SIT).

When tracing a resource is necessary, it's nice to know what kind of impact the trace overhead might have on NCP's performance. In our System Verification Test environment, we measured the effects of running several of the most requested NCP traces in conjunction with various types of network traffic.

We measured the impact of traces on central control unit (CCU) utilization for three scenarios:

- Token-ring boundary network node (BNN) traffic through NCP to an application in the owning Communications Server for OS/390® host
- Mixed-media multilink transmission group (MMMLTG) intermediate network node (INN) traffic between applications in two CS for OS/390 hosts, through two NCPs
- Mixed BNN and INN traffic, combining elements of both of the above environments

We generated BNN traffic with Teleprocessing Network Simulator (TPNS). Session traffic consisted of 100-byte text messages sent between the simulated BNN devices and an echo application.

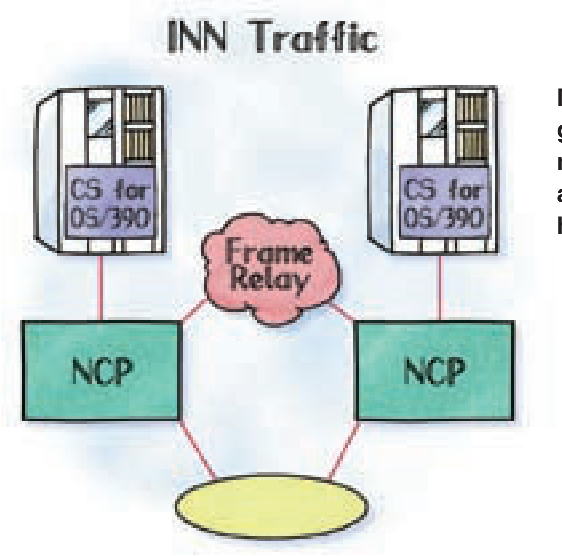


Token-ring boundary network node (BNN) traffic through NCP to an application in the owning Communications Server for OS/390® host

Impact of Traces on NCP Running BNN Traffic

Trace Type	**CCU Utilization (%)**		
No trace running	55	70	85
NTRI SNAP	65	82	99
ODLC SNAP	61	76	99
Channel adapter I/O	63	77	99
SIT on ESCON® physical line	76	91	99*
Line trace on ESCON physical line	69	83	99

* There was a 10% reduction in message rates while running this trace.



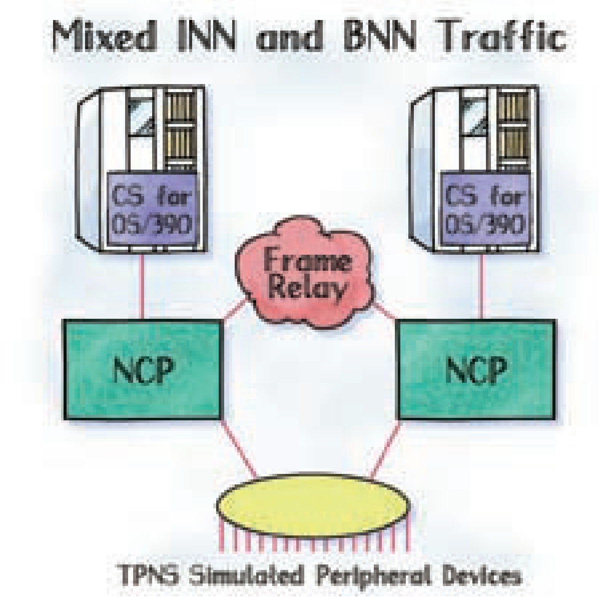
Mixed-media multilink transmission group (MMMLTG) intermediate network node (INN) traffic between applications in two CS for OS/390 hosts, through two NCPs

Impact of Traces on NCP running INN Traffic

Trace Type	CCU Utilization (%)		
No trace	55	70	85
NTRI SNAP	70	88	99
ODLC SNAP	65	78	99
Channel adapter I/O	62	78	99
SIT on ESCON physical line	78	94	99*
Line trace on ESCON physical line	71	85	99

* There was a 12% reduction in message rates while running this trace.

For INN traffic, we again used the echo application, in this case sending a mix of 64-byte to 2000-byte messages between two CS for OS/390 hosts. We varied the number of sessions to achieve the targeted utilization before starting the specified trace.



Mixed BNN and INN traffic, combining elements of both of the above environments

Impact of Traces on NCP running Mixed INN and BNN Traffic

Trace Type	CCU Utilization (%)		
No trace running	55	70	85
NTRI SNAP	69	85	99
ODLC SNAP	62	75	99
Channel Adapter I/O	63	77	99
SIT on ESCON physical line	76	90	99*
Line trace on ESCON physical line	71	86	99

* There was a 10% reduction in message rates while running this trace

System configuration, traffic rates, and message sizes in a production environment might affect the actual results you will see when running a specific NCP trace. And although your results may not be identical to these, we hope that this information will help in anticipating the availability and performance of network resources while gathering crucial diagnostic information. ■

About the Author

Sue DeMarrais joined IBM as a System Verification Tester for the NCP product set in 1990. She has participated in testing several NCP, 3746 Model 900, and S/390 releases.
demar@us.ibm.com

Find IBM Around the World

The following puzzle contains the names of 32 cities that have either an IBM manufacturing or research facility. (There may be a few other cities as well.)

S	Z	E	K	E	S	F	E	H	E	R	V	A	R	Y	A	M	A	T	O
A	B	S	A	N	J	O	S	E	C	D	E	F	G	A	H	A	I	J	G
K	L	S	A	N	T	A	P	A	L	O	M	B	A	S	M	I	N	O	U
B	R	O	M	O	N	T	R	P	R	R	A	S	T	U	V	N	W	X	A
Y	C	N	Z	I	B	M	A	B	O	D	N	E	W	F	S	Z	O	O	D
P	H	N	E	W	O	L	C	A	C	L	A	U	S	T	I	N	E	R	A
O	A	E	I	D	E	L	H	I	H	H	S	H	A	R	N	E	A	U	L
U	R	S	D	U	B	L	I	N	E	P	S	N	A	D	G	W	S	E	A
G	L	O	C	K	S	A	N	M	S	H	A	I	F	A	A	T	T	S	J
H	O	M	E	S	H	B	B	C	T	E	S	T	U	V	P	O	F	C	A
K	T	E	B	C	E	T	U	T	E	E	C	H	R	I	O	N	I	H	R
E	T	O	K	Y	N	O	R	T	R	O	M	E	A	A	R	B	S	L	A
E	E	A	Z	Y	Z	U	I	T	E	H	R	A	L	O	E	O	H	I	X
P	L	E	A	S	H	E	L	P	N	G	R	E	E	N	O	C	K	K	S
S	U	M	A	R	E	A	Z	Y	D	I	B	M	I	B	M	A	I	O	H
I	B	M	A	I	N	A	B	E	I	J	I	N	G	A	S	R	L	N	A
E	F	U	J	I	S	A	W	A	C	A	B	O	H	O	M	A	L	N	R
I	B	U	R	L	I	N	G	T	O	N	M	H	A	L	O	T	I	M	O
B	O	S	T	O	N	M	O	N	T	P	E	L	I	E	R	E	B	O	N
V	I	M	E	R	C	A	T	E	T	W	A	N	G	A	R	A	T	T	A

IBM Research and Manufacturing Plants, Worldwide

Bromont, Canada

Dublin, Ireland

Essonnes, France

Burlington, VT

Greenock, Scotland

Charlotte, NC

Mainz, Germany

East Fishkill, NY

Montpelier, France

Endicott, NY

Santa Palomba, Italy

Manassas, VA

Szekesfehervar, Hungary

Guadalajara, Mexico

Vimercate, Italy

Poughkeepsie, NY

Raleigh, NC

Rochester, MN

Fujisawa, Japan

Prachinburi, Thailand

Singapore

Sumare, Brazil

Shenzhen, China

Wangaratta, Australia

Yasu, Japan

Haifa, Israel

San Jose, CA

Yamato, Japan

Austin, TX

Rueschlikon, Switzerland

Beijing, China

Delhi, India

Library Update

by Leyland King

A number of the NCP/SSP, EP and NTuneMON publications have been updated and made available in both hardcopy and softcopy for the NCP Version 7 Release 8.1 Supplementary Update and NTuneMON Version 3 Release 2. These include:

NTuneNCP Feature Reference	**LY43-0039-01**
NCP V7R8.1, SSP V4R8.1, and EP R14 Diagnosis Guide	**LY43-0033-08**
NTuneMON User's Guide	**SC31-6266-08**
NCP X.25 Planning and Installation Guide	**SC30-3470-13**
LPS for ACF/NCP Version 7 Release 8.1	**GC31-6226-08**
LPS for ACF/SSP Version 4 Release 8.1 for VM	**GC31-6227-07**
LPS for ACF/SSP Version 4 Release 8.1 for OS/390 and MVS	**GC31-6229-08**
LPS for ACF/SSP Version 4 Release 8.1 for VSE/ESA	**GC31-6230-06**
LPS for NTuneMON V3R2	**GC31-6267-06**

Program directories for the supplementary update of the current releases of NCP and SSP as well as NTuneMON V3R2 are also available in softcopy on the NCP publications library CD as well as the Internet in PDF format. You can optionally obtain them in hardcopy from Mechanicsburg as follows:

Program Directory for ACF/SSP V4R8 Modification Level 1 for MVS/ESA OS390	**GI10-6618**
Program Directory for ACF/SSP V4R8 Modification Level 1 for VM/ESA OS390	**GI10-6619**
Program Directory for ACF/SSP V4R8 Modification Level 1 for VSE/ESA OS390	**GI10-6620**
Program Directory for ACF/NCP V7R8 Modification Level 1 for MVS/ESA OS390	**GI10-6621**
Program Directory for ACF/NCP V7R8 Modification Level 1 for VM/ESA OS390	**GI10-6622**
Program Directory for ACF/NCP V7R8 Modification Level 1 for VSE/ESA OS390	**GI10-6623**
Program Directory for NTuneMON V3R2 for MVS/ESA OS390	**GI10-6624**
Program Directory for NTuneMON V3R2 for VM/ESA OS390	**GI10-6625**

Along with these updates, the ACF/NCP, ACF/SSP, EP, NTuneMON and NPSI Softcopy Library Collection Kit, LK2T-0414-08, contains all other NCP library publications on CD.

Besides the BookManager® format, NCP V7R8.1 publications included with the supplementary update appear in PDF format on our CD, IBM ACF/NCP, ACF/SSP, EP, NTuneMON and NPSI Softcopy Library Collection Kit, LK2T-0414-08, as well as on the Internet. These publications are accessible using your Web browser. A pamphlet in the collection kit's CD jewel case tells you just how to access the information.

With the exception of licensed publications available only in hardcopy or on the softcopy library collection kit, you can access these and all other available NCP publications on the Internet at **ibm.com**/networking/ncp

NTuneMON Demonstration Software CD

A CD containing 90-day NTuneMON demonstration software is sleeved in the back cover of this publication. Follow the directions on the CD to load it onto your PC. You can order additional copies of the CD with copies of this magazine. See the back cover for its order number.

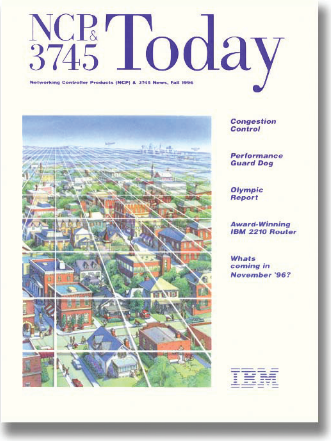
As always, we invite you to contact us conveniently by e-mail through the online version of this publication on the Internet at **ibm.com**/networking/ncp

NCP and 3756/46 Today – Yesterday

The order number for each hardcopy publication appears beside its period of publication. For soft copy versions go to our Web site at **ibm.com**/networking/ncp

Much of the information in *NCP and 3745/46 Today* remains relevant for some time after publication. For that reason, from time to time, our readers will request information covered in previous editions. This index addresses that request.

Fall '96, G325-3426-04
(available only at our Web site)


NCP & 3745 Today
Networking Controller Products (NCP) & 3745 News, Fall 1996
Congestion Control
Performance Guard Dog
Olympic Report
Award-Winning IBM 2210 Router
Whats coming in November '96?
IBM

Spring '98, G325-3426-05

NCP and 3745/46 Today
IBM
Networking Controller Products, 3745 & 3746 News, Spring 1998
Year 2000 – Ready or Not!
Extend Your 3746 Investment
In-Transport Routing
9729 Offers True Optical Networking
IBM Solution Beats Cisco!
See page 20

Fall '98, G325-3426-06

NCP and 3745/46 Today
IBM
Networking Controller Products, 3745 & 3746 News, Fall 1998
Networking Considerations for the S/390 Parallel Sysplex
See page 10
Two New Products!
The Network Utility and the ESCON adapter for 8265 ATM Switch
Do You Use a 3705, 3720, or 3725?
3745 NCP and 3746 Model 900 TCP/IP Routing
Connectionless Network Prioritization, Is it Real?

Focus on 3172 ICP Diagnostics

Addressing a New Networking Paradigm — High-Speed

Connection Networks and Quality of Service

NCP and 3745/46 Capacity Planning Case Studies

Network Traffic Analysis (NTA), a Service Offering Since 1989

What's New with ACF/NCP

Library Update

Connection Balancing for 3745 Frame Relay BAN

3745 NCP and 3746 Model 900 Supports Switched Frame Relay

New CIR Support for NCP-Controlled 3746 Model 900 Frame Relay Lines

NDF Changes Engines

Emulation Program R14 Now Supports All 3745 Models

Continued Investment in NTune Family Provides Exciting New Capabilities

3745 and 3746 Model 900 IP Routing

NCP Service Dates

Strategic Development of SNA Networks

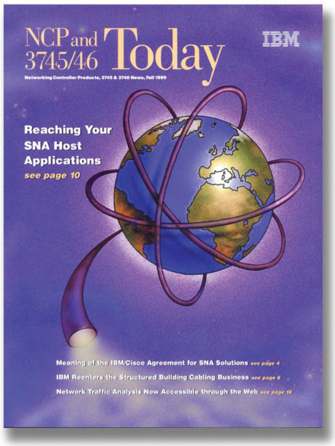
Did You Know?

Do You Use a 3705, 3720, or 3725?

NCP Word Search Puzzle

User Group Meeting Schedule as of this Issue

Fall '99, G325-3426-08


NCP and 3745/46 Today
IBM
Networking Controller Products, 3745 & 3746 News, Fall 1999
Reaching Your SNA Host Applications see page 10
Meaning of the IBM/Cisco Agreement for SNA Solutions see page 4
IBM Reenters the Structured Building Cabling Business see page 6
Network Traffic Analysis Now Accessible through the Web see page 18

About This Issue

Meaning of the IBM/Cisco Agreement for SNA Solutions

Why Is IBM back in the Cabling Business

Reaching Your SNA Host Applications

Network Traffic Analysis (NTA) Now Accessible through the Web

Multilayer, Multiprotocol Switched Networking for Powerful e-Business and Global Networking

3746 Nways Multiprotocol Controller Enhances Major Components in 1999

TPF adds CDLC/ 3746 Support for IP

Library Update

What's New with ACF/NCP (Carpe Diem)

User Group Meetings

NCP Product Set — Year 2000 Analysis Overview

NCP Service Dates

ACF/SSP Boasts Enhancements for V4R8

Connection Balancing for 3746-9x0 Token Ring and Frame-Relay BAN

Pool Association Change for 3746-900 Duplicate TICs

Non-Disruptive Route Switching for 3745 Subarea Token-Ring Connections

NTuneMON adds Support for Bisync and Start-Stop Lines along with New NCP Function

NPSI Update

Summer '00, G325-3426-10

NCP and 3745/46 Today
IBM
Networking Controller Products, 3745 & 3746 News, Summer 2000
NTuneMON CD Inside
Evolving your SNA Investment to e-business See page 12
IBM's Shark Enterprise Storage Server See page 42
What's New with ACF/NCP and SSP? See page 67
Get More NTuneMON for Your Money See page 69

About This Issue

SNA and IP Integration: Time to Move?

It's Time to Switch — OSA-Express Drives the TCP/IP Data Center and the Network Together

Evolving your SNA Investment to e-business

3745/3746 Server Access Scenarios with Technology Overviews

IBM Enterprise Storage Server

What's New with ACF/NCP and SSP

Get More NTuneMON for Your Money

How We Test New Program Enhancements

3746 Nways Multiprotocol Controller Year 2000 Enhancement Overview

NTA: A Service Tool for Analyzing Complex Performance Problems in TCP/IP and SNA

Library Update

User Group Meetings

Post-Sales Product Support — What's New?

NCP and 3745/46 Today — Yesterday

Download Free NTuneMON V3R1 Trial Software ■

List of Abbreviations

ABR	area border router
AFS	Andrew File System
API	Application programming interface
APPN	Advanced Peer-to-Peer Networking
ARP	Address Resolution Protocol
ASBR	autonomous system boundary router
ASP	application service provider
ATM	asynchronous transfer mode
BER	box event record
BEX	Branch Extender
CBQ	class-based queueing
CCM	Controller Configuration and Management
CHPID	Channel Path Identifier
CIFS	Common Interface File System
CLAW	Common Link Access to Workstation
CoS	class of service
CRH	Channel Request Handler
CS	communication server
CTC	Cluster transition configurator
DASD	direct access storage device
DEN	directory-enabled network
DFS	Distributed File System
DHCP	Dynamic Host Configuration Protocol
DLCI	Data Link Connection Identifier
DMA	direct memory access
DMTF	Distributed Management Task Force
DNS	Domain Name System
DNS/WLM	Domain Name Server/Workload Manager
DoS	denial of service
DSCP	(1) data services command processor (2) Differentiated Services Code Point
EBCDIC	Extended binary-coded decimal interchange code
EMIF	ESCON Multiple Image Facility
EN	end node
EREP	Environmental Record Editing and Printing Program
ERP	error recovery procedure
ESCON	Enterprise Systems Connection
FRCA	Fast Response Cache Accelerator
FTP	File Transfer Protocol
HBA	host bus adapter
HPDT	high-performance data transfer
HPR	High-Performance Routing
HTTP	Hypertext Transfer Protocol
ID	intrusion detection
IDC	International Data Corporation
IETF	Internet Engineering Task Force
IntServ	Integrated Services
IPCS	Interactive Problem Control System

IPSec	IP Security Protocol
IPX	Internetwork Packet Exchange
ISP	Internet service provider
IUCV	Inter-User Communication Vehicle
LAN	local area network
LANE	LAN emulation
LDAP	Lightweight Directory Access Protocol
LEN	low-entry networking
LSA	link-state advertisement
LU	logical unit
MAC	medium access control
MIB	Management Information Base
MLTG	multilink transmission group
MMMLTG	mixed-media multilink transmission group
MOSS	Maintenance and Operator Subsystem
MPC	multipath channel
MTU	maximum transfer unit
MVS	Multiple Virtual Storage
NFS	Network File System
NFS	Network File System
NN	network node
NPSI	X.25 NCP Packet Switching Interface
NSSA	Not-So-Stubby-Area
ODLC	Outboard Data Link Control
ONC RPC	Open Network Computing Remote Procedure Call
OSPF	Open Shortest Path First
PAPI	Policy API
PIU	path information unit
POSIX	Portable Operating System Interface for Computer Environments (IEEE standard)
QDIO	Queued Direct Input/Output
QoS	quality of service
RAPI	RSVP API
RECMS	record maintenance statistics
REXX	Restructured Extended Executor
RSVP	Resource ReSerVation Protocol
SNA	Systems Network Architecture
SNMP	Simple Network Management Protocol
STI	Self-Timed Interface
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol
TG	transmission group
ToS	type of service
TPF	Transaction Processing Facility
TR	traffic regulation
UDP	User Datagram Protocol
URI	Universal Resource Identifier
VC	virtual circuit
VIPA	virtual IP address
VoIP	Voice over IP
VPN	virtual private network
WAN	wide area network
WFQ	weighted fair queueing
WRED	weighted random early discard
WRR	weighted round robin
XCF	Cross-System Coupling Facility
XTI	X/Open Transport Interface

We want to hear from you

It has been our goal to make this magazine informative and interesting. We hope we have achieved that goal. Please send us your comments and suggestions.

Name: ____________________

Title: ____________________

Address: ____________________

City: ____________________ State: __________ Country: __________ ZIP: __________

Fax or e-mail address changes, comments, suggestions and subscription requests to:

NCP and 3745/46 Today
c/o Leyland King
NCP Product Development
919 254-0343

NCP and 3745/46 Today at:
ibm.com/networking/ncp

The online copy of this magazine also contains a form on which you may conveniently submit comments, suggestions, and subscription requests.

NCP and 3745/46 Today is published annually and distributed by subscription free of charge worldwide. Subscriptions are available to individuals and quantity dropshipment to IBM Branch Offices and IBM Business Partners on request. Back issues are also available by order number through your IBM Sales Representative or IBM Branch Office and available for review on the Internet available at:
ibm.com/networking/ncp

NCP and 3745/46 Today is produced by Hobart V. Love III, manager of NCP Development and Support, IBM Storage Networking Division. Project coordination, editorial review, graphic design, production and Web site development services are provided by IBM's Design and Information Development (D&ID) organization, RTP, NC.

Visit D&ID at:
ibm.com/networking/dandid/

Mailing address:
NCP and 3745/46 Today
IBM Corporation
P.O. Box 12195
Research Triangle Park, NC 27709-2195

Internet address for this edition:
ibm.com/networking/ncp

IBM
NCP and 3745/46 Today
A Publication of the IBM Storage Networking Division
Summer 2001
Storage Networking
More than an SNA Anagram
3746 Extended Functions 6
iSCSI: Furthering IBM Storage Networking Leadership
Leveraging SNA while Transforming to e-business
IBM Continues Drive to Storage Leadership
by Linda Sanford
NTuneMON CD Inside

Managing Editor: **Leyland King**
Technical Editors: **G. David Heath** and **Ralph Nissler**
Contributor and Technical Advisor: **Chuck Williams**
Graphic Designers: **Ann Roberson** and **Nicole Zeisz**
Illustrator: **Adam Brill**

NTuneMON Presentation and Limited-Time Trial Software Offer

The CD enclosed in this magazine contains:

1. A brief presentation about NTuneMON™ features

2. 90-day free trial NTuneMON software
The free trial software offer expires on November 30, 2001.

3. NTuneMON User's Guide

To run the presentation, enjoy the trial software, or review the user's guide, please follow instructions printed on the CD. If you need assistance with the free trial software please call 919 254-5240.

IBM Corporation
Department TYCA
PO Box 12195
Research Triangle Park, NC 27709
U.S.A.

Printed in the United States of America
6-01
All Rights Reserved

♻
Printed on recycled paper

For Position Only

G325-3426-11

IBM

NCP and 3745/46 Today

A Publication of the IBM Storage Networking Division

Summer 2001

Storage Networking

More than an SNA Anagram

3746 Extended Functions 6

iSCSI: Furthering IBM Storage Networking Leadership

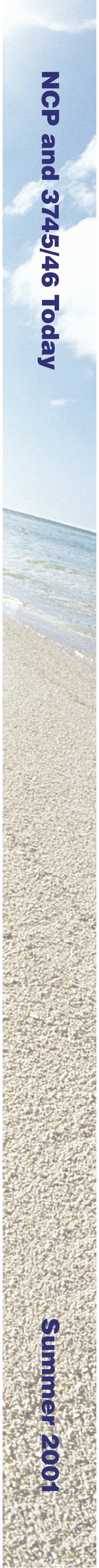
Leveraging SNA while Transforming to e-business

IBM Continues Drive to Storage Leadership
by Linda Sanford

NTuneMON CD Inside

Printer:
Please center text on spine.
We are providing a larger image than needed for the spine b/c we don't know how large the spine will be.

Please don't allow spine image to overlay the front image or spill onto the back.



NCP and 3745/46 Today

Summer 2001