# RS/6000 7025 Model F80

## Technical Overview and Introduction

May 9, 2000

Stephen Lutz

Shyam Manohar


Scott Vetter

International Technical Support Organization

Austin, TX

IBM

# IBM RS/6000 7025 Model F80 Server

IBM achieved leadership performance in the high-end SMP UNIX server marketplace through the balanced design of the 24-way RS/6000 7017 Model S80, which ranked top in the industry standard TPC-C benchmark at 135,815.

On May 9, 2000, IBM introduced the RS/6000 7025-F80, 7026-H80, and 7026-M80 to enhance the mid-range server lineup with products using many of the design elements that led to the success of the high-end Model S80.

### Overview

While the IBM RS/6000 Models F80 and H80 are the systems that provide a growth path for existing installations of Model F50s and H70s, the Model M80 is designed to provide leadership performance among the mid-range 8-way systems. The target for performance improvement over the existing mid-range Model F50 is over three times with the Model F80 and H80, and over five times with the Model M80 server.

In addition to providing CPU and I/O expandability, the Model F80 combined with the latest storage technology provides the maximum internal storage capability available among the current line of RS/6000 mid-range servers.

This paper discusses, in detail, the processor, memory, I/O, expandability, reliability, and other technical aspects related to the Model F80.

## IBM RS/6000 Model F80 Description

The IBM RS/6000 Model F80 is a member of the 64-bit family of symmetric multiprocessing (SMP) servers from IBM. The Model F80 is a 64-bit deskside system, which can be configured as a 1-, 2-, 4-, or 6-way SMP with up to 16 GB of real memory.

The Model F80 offers flexibility regarding the number of CPUs, memory DIMMs, PCI adapters, and disk drives desired for a specific application or usage.

### Physical Package

The Model F80 is packaged in a rugged black deskside steel chassis. The Model F80 server includes a modular hot-swap disk subsystem that allows fast, easy addition and replacement of drives. It has a maximum internal storage capacity of 254.8 GB (218.4 GB hot-swappable) and the possibility to double this as new storage technologies are made available. The Model F80 has a flexible I/O subsystem including ten 64-bit hot-plug PCI slots. It is shipped with the internal adapters and devices installed and configured; software can also be preinstalled if desired.

The Model F80 is designed to operate in a typical office environment with standard AC power at 100-127 volts or 200-240 volts. The Model F80 systems are built with two standard internal hot-swappable power supplies that combine to provide ample power for any configuration. An optional third internal hot-swappable power supply with two hot-pluggable fans can be ordered to provide redundant power and cooling, allowing the system to continue running in the event of a failed fan or power supply. In the event of one of the fans failing, the

other fans increase their speed to provide sufficient cooling. The hot-pluggable fans are performance monitored by the SPCN (see “System Power Control Network (SPCN)” on page 15). The hot-swappable power supplies and fans can be replaced concurrently if the optional redundant power is installed.

The dimensions of the Model F80 are 483 mm W x 728 mm D x 610 mm H (19.0” W x 28.7” D x 24.0” H). The weight is from 70 kg (115 lbs) to 95 kg (209 lbs) depending on the configuration. Unlike the Model F50, the Model F80 does not provide rollers.

The Model F80 is designed for customer setup of the machine and for subsequent addition of most features.

Figure 1 shows front/side view of a Model F80 showing the location of the major features. A discussion of these features follows.

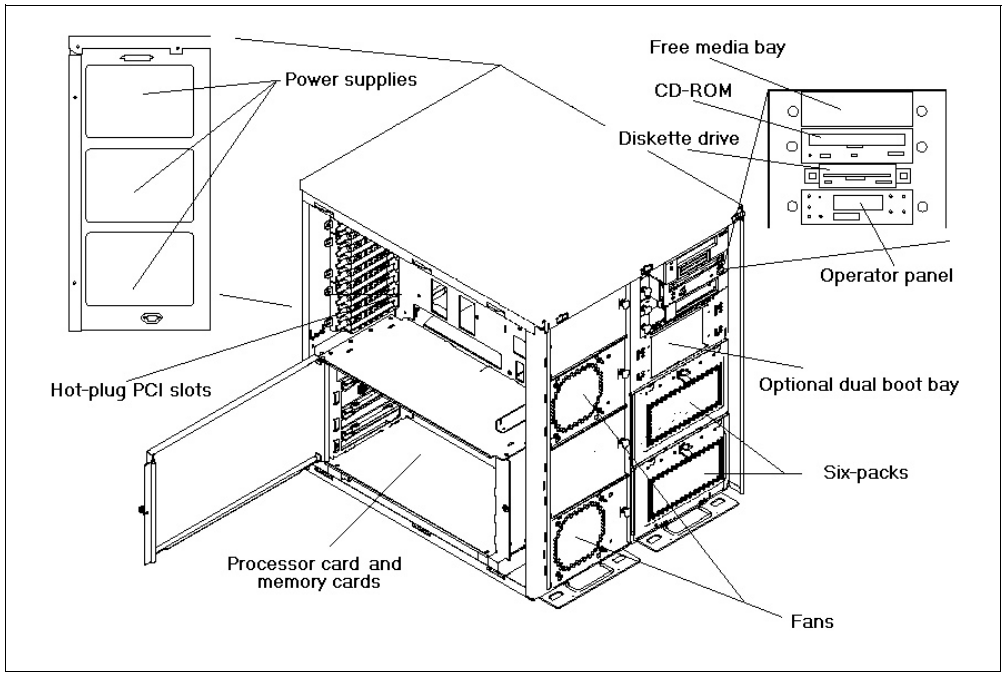



*Figure 1. Model F80 Overview*

On the bottom of the chassis are three slots on the backplane used for the processor card (top slot) and the two possible memory cards (middle and bottom slots). A *dummy* memory card is installed in all shipped units for each unused memory slot for safety and proper machine cooling.

The following ports are provided by Model F80 as shown in Figure 2:

- One Ultra2 SCSI port for external attachment use (mini 68-pin VHDCI[1] connector)

  The industry standard VHDCI 68-pin connector on the backside of the Model F80 allows attachment of various LVD and SE external subsystems. A 0.3 meter converter cable, VHDCI to P, mini-68 pin to 68-pin, (# 2118) can be used with older external SE subsystems to allow a connection to the VHDCI connector.

[1] Very High Density Cable Interconnect (VHDCI)

- 10/100 Mb/s Ethernet port (RJ-45 connector)
- Four serial ports (max. 230 KB/s, 9-pin D-shell)
- One parallel port (bi-directional)
- Test port
- Keyboard and mouse port

The test port is for diagnostics and is normally covered with a metal plate. It uses the same connector as the parallel port. To avoid confusion, this port should remain covered.

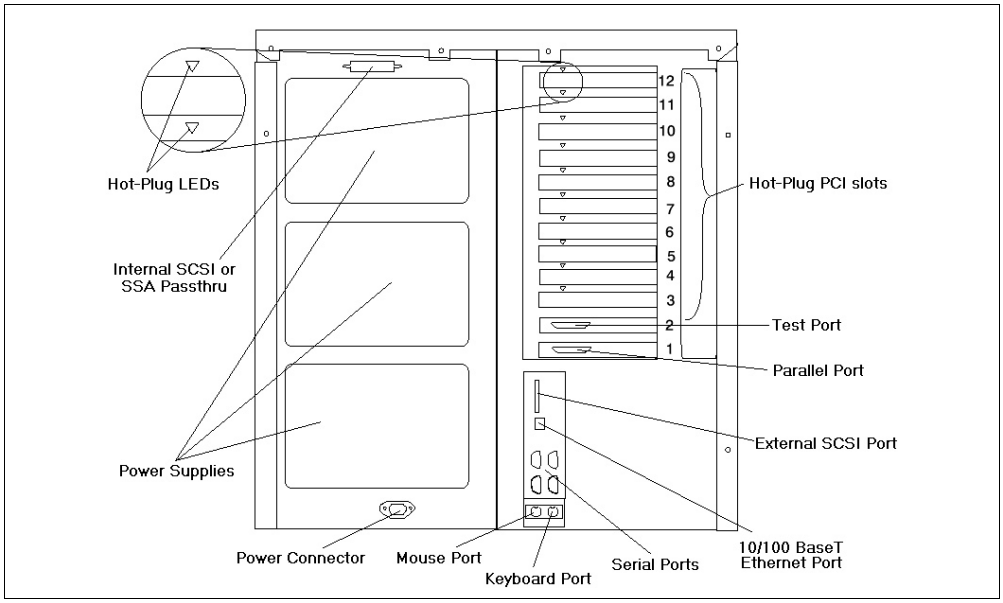



*Figure 2. Model F80 Rear View*

## Internal Storage

The system comes preconfigured with a CD-ROM and a diskette drive and one free media bay for customer expansion, such as a tape device. Any devices in the media bays are connected to the internal F/W SCSI controller (no additional cable is required).

The Model F80 features two six-packs providing 12 hot-swap disk bays and an additional boot disk bay for two additional disks (not hot-swappable). The two six-packs can be either equipped with a SCSI backplane (# 6553) or SSA backplane (# 6554). SCSI and SSA six-packs can be mixed.

The SCSI backplane supports one inch and 1.6 inch drives. If the older 1.6 inch drives are used, these occupy two adjacent bays. The new SCSI carrier has one of two interposers between the drive and backplane, depending on whether the drive is 68-pin or 80-pin SCSI. All SCSI drives sold new with a Model F80 will be 80-pin drives.

SCSI RAID is supported for the *under-the-cover* disks, but requires the addition of a SCSI RAID adapter. To cable a second internal SCSI RAID six-pack, a cable assembly is attached to an external SCSI port on the SCSI RAID adapter, run

through a passthru of the rear bulkhead on the power supply side of the backpanel, and attached to the SCSI backplane.

SSA disks require the addition of an SSA adapter. To cable internal SSA, a cable assembly is attached to two external ports on the SSA adapter. This runs through a passthru of the rear bulkhead on the power supply side of the backpanel and is attached to both ends of the SSA backplane. When using both SSA six-packs, the cable runs to one end of each of the two backplanes with a short SSA cable in between the two backplanes. These configurations provide a loop, which is part of the SSA architecture. The SSA six-pack requires dummy jumper cards in vacant bays to maintain the SSA loop, and so can only support one inch drives. Booting from SSA disks attached to an Advanced SerialRAID adapter (# 6225) is supported from the six-pack or external SSA disks provided that the disks are arranged in a non-RAID configuration.

The optional dual boot bay holds a two-pack located between the operator panel and the top six-pack of the cabinet (D13, D14 in Figure 3). They are SCSI drives attached to the same carriers as are used for the six-packs. They plug into a backplane in the two-pack that does not support hot-swap. The backplane is designed in such a way that the two disks can either be part of the same SCSI bus or attached to different SCSI busses to support mirroring of the boot image. The disks should have 80-pin connectors so that a flex interposer between the disks and the backplane is unnecessary. These two disks do not impact the two six-packs, so the six-packs can be used, for example, in a RAID configuration. If the dual boot bay option is not installed, one of the six-packs will not be able to provide RAID, because booting from RAID disks is not supported.

Figure 3 shows the internal devices and bays of a Model F80.

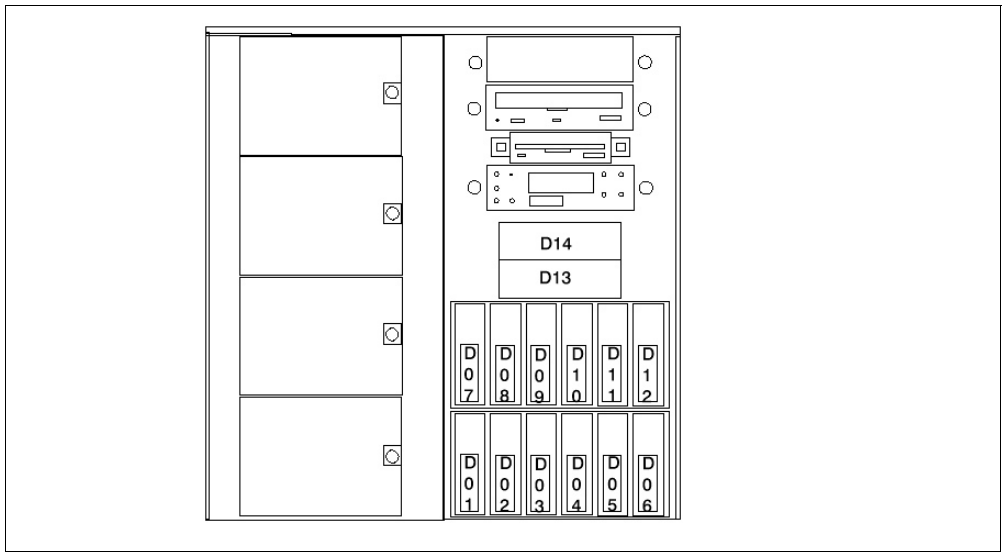



*Figure 3. Internal Devices and Bays of the Model F80*

## Operator Panel

The Model F80 cabinet incorporates the operator panel and indicators for the system. The panel consists of the following features (see Figure 4):

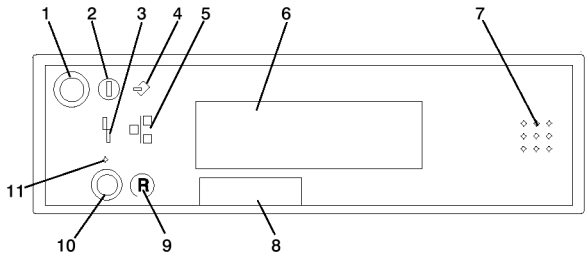
- Power On/Off Button
- Power-on LED (Green)
- System Attention LED (Yellow)

  This LED indicates to a user that there is an attention condition on the system.
- System Activity LEDs (Green)

  These LEDs report the status of the integrated SCSI and Ethernet ports.
- Operator Panel Display

  The display has two lines of sixteen characters each. The display shows reference codes from the service processor, the SPCN, and the operating system. These codes can be either informational codes or error codes. Informational codes will have the System Attention LED off; error codes will have the System Attention LED on.
- Service Processor Reset Button

  The service processor reset button is a single execution button used to reset the service processor and bring the system back into standby mode. Access to this button is restricted by only having access to this button through a *pinhole* in the operator panel cover. This button is for service use only.
- Speaker (beeper)




| | | | |
|---|---|---|---|
| 1 | Power On/Off Button | 7 | Speaker |
| 2 | Power-on LED (Green) | 8 | Serial Number Plate |
| 3 | System Attention LED (Yellow) | 9 | Reset Icon |
| 4 | SCSI Activity LED (Green) | 10 | Reset Button |
| 5 | LAN Activity LED (Green) | 11 | For Service Use Only |
| 6 | Operator Panel Display | | |

*Figure 4. Operator Panel*

## System Architecture and Technical Overview

Figure 5 shows the system schematic of the Model F80. In this section the different physical components in the schematic and the SMP processor configurations are discussed.

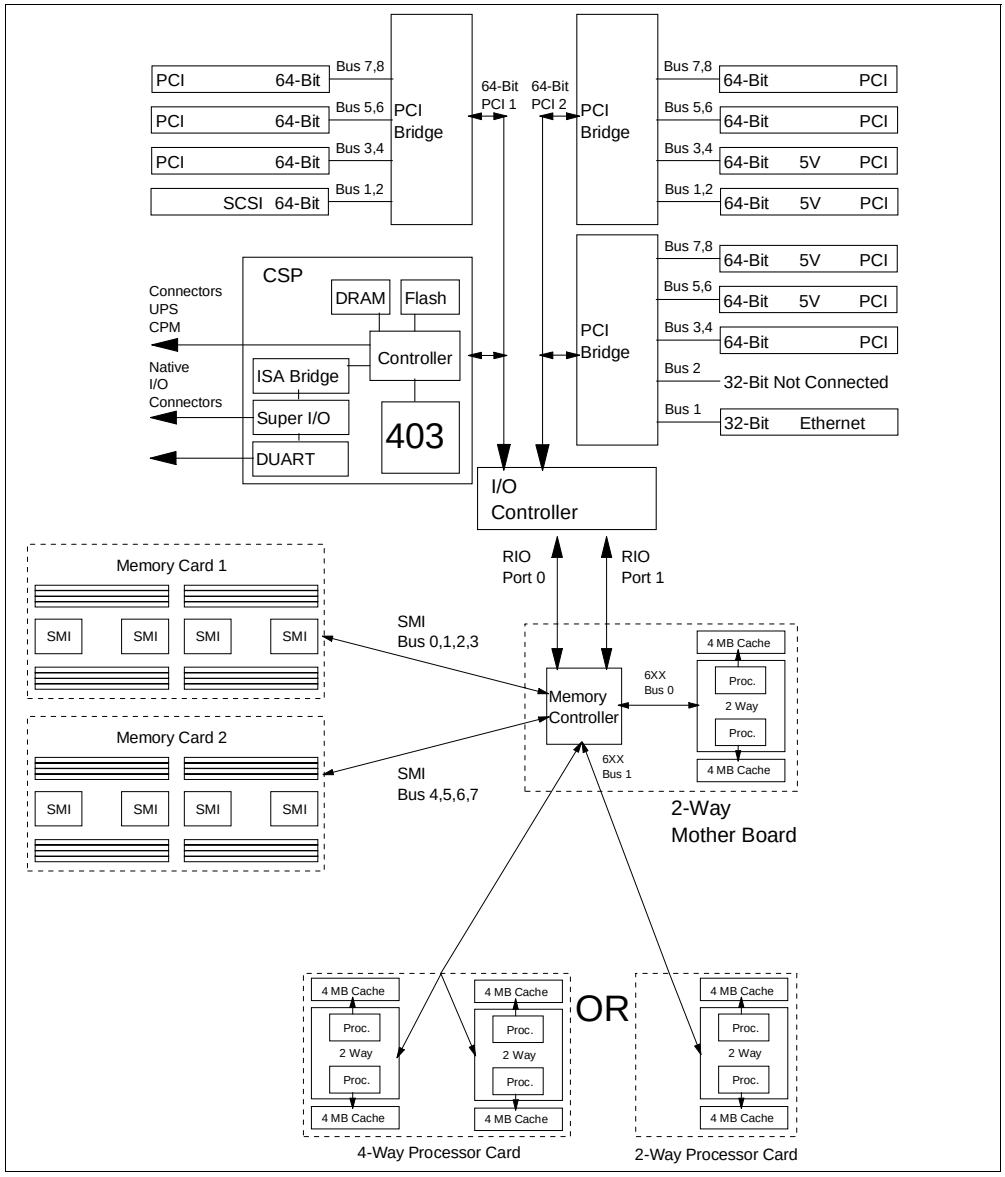



*Figure 5. RS/6000 Model F80 System Schematic for 2-, 4-, or 6-Way SMP*

## CPU Architecture

The key components in the CPU include the processor, the processor packaging, memory controller, memory subsystem, and the I/O interface.

### RS64 III RISC Processor

The RS64 III processor card used in the Model F80 has the following attributes:

- 450 MHz or 500 MHz operating frequency

- 128 KB on-chip L1 instruction cache with parity and refetch
- 128 KB on-chip L1 data cache with ECC
- On-chip L2 cache directory
- 4 MB of off-chip L2 cache using ECC double data rate (DDR) SRAM per processor for 2-, 4-, and 6-way SMPs and 2 MB of off-chip L2 cache using ECC Single Data Rate (SDR) SRAM for a 1-way system.
- PowerPC 6xx bus architecture, 16-byte wide bus interface

The RS64 III processor is available in two operating frequencies, 450 MHz and 500 MHz. The frequency is accomplished by leveraging IBMs copper technology (CMOS 7S) along with an innovative design of timing-critical paths.

The copper technology and an improved manufacturing process allow the chip to operate at 1.8V. The lower operating voltage coupled with the smaller circuit dimensions result in reduced wattage in the RS64 III and allow additional function to be placed on the chip.

The size of the level one (L1) instruction and data caches is 128 KB each. Innovative custom circuit design techniques were used to maintain the one cycle load-to-use latency for the L1 data cache. The level two (L2) cache directory was integrated into the RS64 III chip, reducing off-chip accesses which impact performance.

IBM used double data rate (DDR) SRAM technology for the L2 cache in the RS64 III processor. DDR technology provides two transfers of data on the 16-byte wide L2 data bus every SRAM clock cycle. The DDR SRAM technology also reduced L2 access latency as measured by nanoseconds.

### Processor Boards

The processor boards used in the Model F80 for 1-, 2-, 4-, and 6-way SMP configurations come in the form of a single book and are described as follows:

#### Single Processor

As shown in Figure 6, a single processor board consists of a single RS64 III processor operating at 450 MHz, on-board memory slots, and a memory controller in a single book. Upgrades to additional processors require changing of the processor book. However, the single processor board is a cost-reduced package.

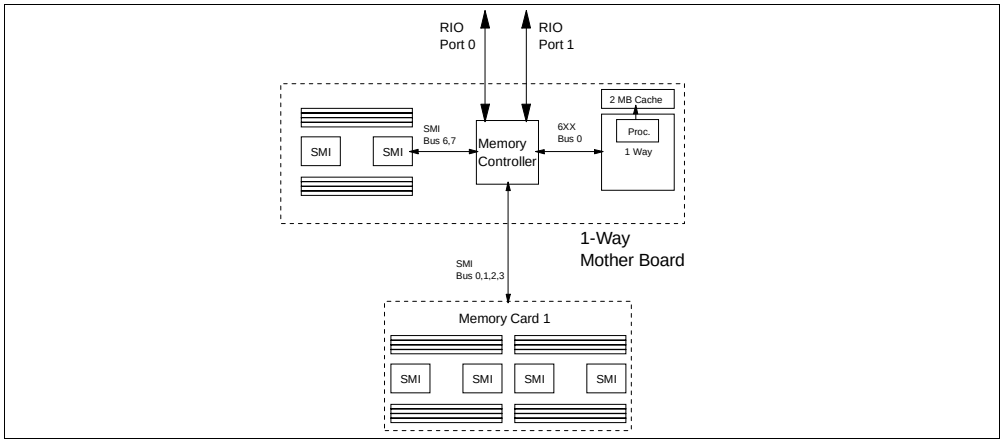



*Figure 6. RS/6000 Model F80 System Schematic for 1-Way Processor*

### 2- and 4- Way SMP

A 2-way SMP configuration is provided by a processor board consisting of a pair of RS64 III processors operating at 450 MHz and a memory controller. Expansion to 4-way SMP is provided by interfacing an additional processor board consisting of a pair of RS64 III processors operating at 450 MHz. However, the upgrade from 2-way to 4-way SMP is offered as a book swap, since the addition of the processor card on the processor book is too delicate for field handling.

### 6-Way SMP

A 6-way SMP configuration uses two processor boards which are interfaced to each other. One processor board consists of a pair of RS64 III processors operating at 500 MHz and a memory controller. The other processor board consists of four RS64 III processors operating at 500 MHz. Upgrades from a 2- or 4-way SMP to a 6-way SMP are offered as a book swap.

### Memory Controller

A single custom chip provides the function of the memory controller and the I/O hub in the Model F80. The controller chip provides interfaces to processors, memory, and the I/O subsystem.

The RS64 III processors on the processor boards are connected to the memory controller through the PowerPC 6xx bus. The controller chip is a part of the first processor board. The memory controller provides a single 6xx bus interface in a single processor configuration. For 2-way SMP configurations, the controller provides a 6xx bus interface to the pair of RS64 III processors present in the same board. The memory controller provides another 6xx bus interface for CPU expansion using an additional processor board. The 4- and 6-way SMP configurations consists of a total of two processor boards which uses the two 6xx bus interfaces provided by the memory controller installed together in a book.

In the Model F80, the 6xx bus is a 16-byte wide bus and the operating clock rate of the bus depends upon the processor clock speed. The 6xx bus operates at a clock rate of 150 MHz, for a processor clock speed of 450 MHz. And for a processor clock speed of 500 MHz, the 6xx bus operates at a clock rate of 125 MHz.

### Memory Subsystem

The memory controller provides two memory bus interfaces and provides the reliability functions of ECC as well as memory scrubbing. Memory scrubbing provides a built-in hardware function that is designed to perform continuous background reads of data from memory, checking for correctable errors. The memory configuration for a single processor configuration and 2-, 4-, or 6-way configurations is explained as follows:

- In a single processor configuration, the on-board memory, consisting of eight DIMM slots, is interfaced to one of the two memory interfaces in the controller. The other interface is used by a separate riser memory card. The riser memory card provides 16 DIMM slots. While the DIMM slots in the on-board memory are populated in pairs, the slots in the riser memory card are populated in quads. The minimum configuration requires a pair of DIMMs in the on-board memory. Once the on-board memory slots are filled and more memory capacity is desired, the DIMMs are moved to the riser memory card and the next increment is made as a quad. The single processor configuration

can provide a maximum memory of 8 GB by using the riser memory card and populating each of the 16 slots using 512 MB DIMMs.

- In 2-, 4-, or 6-way SMP configurations, the memory is provided using two separate riser memory cards, each with 16 DIMM slots and populated with DIMMs in quads. The two riser cards are interfaced to the two memory interfaces in the memory controller. The minimum 2-way SMP configuration requires a single riser memory card populated with a quad of DIMMs. The second riser memory card with minimum of a quad of DIMMs can be configured only after the 16 DIMM slots in the first riser memory card are fully populated. 2-, 4-, or 6-way processor configuration can support up to 16 GB maximum memory by using the two memory riser cards fully populated with 512 MB DIMMs.

The Model F80 uses 200-pin 10ns SDRAM DIMMs. DIMMs of equal sizes must be used, while populating in pairs or quads. DIMM size used in one pair or quad can, however, coexist with a different DIMM size used in another pair or quad.

In the Model F80, the bus interface from each riser memory card to the memory controller is 8-bytes wide and operates at clock rate double that of the PowerPC 6xx bus. No additional memory bandwidth can be achieved by splitting memory between cards.

### Bus Bandwidth

The following are the theoretical maximum bandwidths as applicable for a 6-way 500 MHz SMP configuration:

- Total memory bandwidth: 2 GB/s
- Total processor bandwidth: 2 GB/s
- Total I/O bandwidth: 1 GB/s (500 MB/s bi-directional)

The following are the theoretical maximum bandwidths applicable for 2-, or 4-way 450 MHz SMP configurations:

- Total memory bandwidth: 2.4 GB/s
- Total processor bandwidth: 2.4 GB/s
- Total I/O bandwidth: 1 GB/s (500 MB/s bi-directional)

### I/O Hub Function

The memory controller also functions as the I/O hub. The controller provides two RIO (remote I/O) ports. The two RIO ports are attached to an I/O host bridge chip. Each RIO port has two uni-directional 1-byte wide links. All the I/O transfers take place using one primary RIO port, which operates at 500 MHz (500 MB/s bi-directional or an aggregate of 1 GB/s). The controller uses the other RIO port, which operates at 250 MHz (250 MB/s bi-directional or aggregate of 500 MB/s), as a fail-over to the primary RIO port. In contrast to Model H80 or Model M80, these RIO connections in the Model F80 are not visible outside the cabinet, but the function is the same.

## Internal I/O Architecture

As already discussed, the system includes one I/O host bridge chip managing all the I/O between the I/O adapters and the memory controller using RIO connections. On the other side, the I/O host bridge provides two primary PCI busses, operating at 66 MHz and 64-bit wide.

The service processor and a PCI-to-PCI bridge chip are connected to the first primary PCI bus. The PCI-to-PCI bridge provides three 64-bit hot-plug PCI slots and the onboard dual SCSI adapter (F/W SCSI internal, Ultra2 SCSI external). A PCI-to-ISA bridge chip is connected to the service processor providing an ISA bus. The ISA bus is used by the National Super I/O chip providing the floppy drive controller, two of the four serial ports, keyboard and mouse ports, and the parallel printer interface. A 16552 DUART chip is also connected to the service processor providing the other two serial ports.

The second bus is connected to another two PCI bridges, which provide another seven 64-bit hot-plug PCI slots. The onboard 10/100 Mb/s Ethernet adapter is connected to this chip.

Each slot represents a separate PCI bus, which simplifies the hot-plug functionality. Figure 7 shows the design of the I/O architecture.

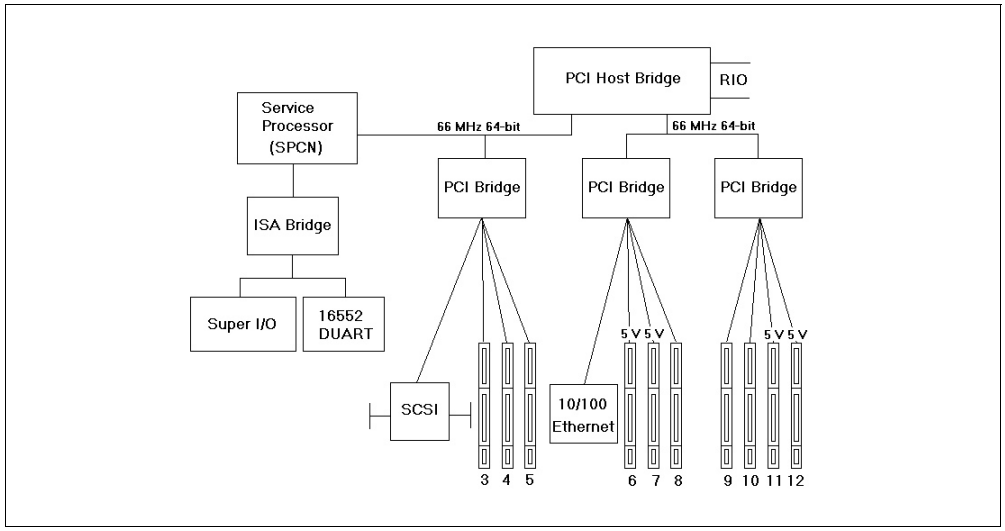



*Figure 7. Internal Architecture of Model F80*

### PCI Slots

All PCI slots are PCI 2.2 compliant and are hot-plug enabled, which allows most PCI adapters to be removed, added, or replaced without powering down the system. This function enhances system availability and serviceability.

Six 64-bit slots operate at 3.3V signaling at 66 MHz, in contrast to the four 64-bit slots which operate at 5V signaling at 50 MHz (see Figure 7). When adding adapters to the system, it is important which signaling the adapter works: 3.3V, 5V, or universal, which means the adapter works at both voltages. That is, for example, the reason why a PCI 3-Channel Ultra2 SCSI RAID Adapter (# 2494) can be placed only in slots 6, 7, 11, or 12. Refer to the *PCI Adapter Placement Reference Guide*, SA38-0538 for further information.

### Hot-Plug PCI Adapters

The function of hot-pluggable PCI adapters is to provide concurrent additions or removals of PCI adapters when the system is running. This function is explained in the following paragraphs.

In the chassis, the installed adapters inside the slots are protected by plastic separators, designed to prevent grounding and damage when adding or removing adapters. The hot-plug LEDs outside the chassis indicate if a adapter can be plugged in or removed from the system. These LEDs are also visible inside the chassis. Inside, the light from the LED is routed to the top of the plastic separators, using light pipes, which makes it very easy to locate the right slot. The hot-plug PCI adapters are secured with retainer clips on top of the slots; therefore, you do not need a screwdriver to add or remove a card and there is no screw to drop inside the chassis causing damage to the system.

The function of hot-plug is not only provided by the PCI slot, but also by the function of the adapter. Most adapters are hot-pluggable, but some are not. Be aware that some adapters must not be removed when the system is running, for example, the adapter with the operating system disks connected to it, or the adapter that provides the system console. Refer to the *PCI Adapter Placement Reference Guide*, SA38-0538 for further information.

To manage hot-plug PCI adapters, it is important to turn off slot power before adding, removing, or replacing the adapter, which is done by the operating system. There are three possibilities for managing hot-plug PCI slots in AIX:

- Command line:
  - `lsslot` - List slots and their characteristics
  - `drslot` - Dynamically reconfigures slots
- SMIT
- WebSM

When working with the commands and tools mentioned above, the hot-plug LEDs (see Figure 2) change their state. Table 1 shows the possible states of the hot-plug LEDs.

*Table 1. Hot-Plug LED Indications*

| LED Indication | PCI Slot Status | Definition |
|---|---|---|
| Off | Off | Slot power is off. It is safe to remove or replace adapters. |
| On (not flashing) | On | Slot power is on. Do not remove or replace adapters. |
| Flashing slowly (one flash per second) | Identify | Indicates the slot has been identified by the software; do not remove or replace adapters at this time. |
| Flashing fast (six to eight flashes per second) | Action | Indicates the slot is ready for adding, removing, or replacing of adapters. |

To add a hot-plug PCI adapter use the `drslot` command to set the slot first into the Identify state (LED flashes slowly) to verity the right slot was selected. After pressing **Enter,** the LED changes its state to the Action state (LED flashes fast).

Then add the adapter to the system. When finished, press **Enter** again to turn on slot power. The hot-plug LED will change its state to On. Now the adapter is integrated into the system and can be configured using AIX `cfgmgr` (Configuration manager).

Removal of adapters requires deconfiguration in AIX first. The adapter must be in a defined state or removed from the ODM.

Figure 8 shows an example of adding a hot-plug PCI adapter to a running system.

```
# lsslot -c pci
# Slot Description                       Device(s)
P1-I3  PCI 64 bit, 66 MHz, 3.3 volt slot  Empty
P1-I4  PCI 64 bit, 66 MHz, 3.3 volt slot  Empty
P1-I5  PCI 64 bit, 66 MHz, 3.3 volt slot  ent1
P1-I6  PCI 64 bit, 50 MHz, 5 volt slot    Empty
P1-I7  PCI 64 bit, 50 MHz, 5 volt slot    Empty
P1-I8  PCI 64 bit, 66 MHz, 3.3 volt slot  scsi2, scsi3
P1-I9  PCI 64 bit, 66 MHz, 3.3 volt slot  Empty
P1-I10 PCI 64 bit, 66 MHz, 3.3 volt slot  Empty
P1-I11 PCI 64 bit, 50 MHz, 5 volt slot    ssa0
P1-I12 PCI 64 bit, 50 MHz, 5 volt slot    Empty

# drslot -c pci -Ia -s P1-I6

The visual indicator for the specified PCI slot has
been set to the identify state. Press Enter to continue
or enter x to exit.

[Enter]

The visual indicator for the specified PCI slot has
been set to the action state. Insert the PCI card
into the identified slot, connect any devices to be
configured and press Enter to continue. Enter x to exit.

[Enter]
```

*Figure 8. Example: Adding a Hot-Plug PCI Adapter*

## Software Requirements

The Model F80 requires AIX 4.3.3 with the AIX 4330-03 recommended maintenance package (APAR IY09047), which is included on all pre-installed systems and on the 04/2000 Update CD that ships with AIX 4.3.3 as of April 2000.

In addition, there is APAR IY09814, which includes additional fixes that were not available before the 4330-03 package was shipped. In order to install the Model F80 from CD, you need an AIX 4.3.3 CD dated 04/2000 (LCD4-0286-05) or later, because the system will not boot from older AIX 4.3.3 CDs. You can also download the actual maintenance level from the Internet to install the machine using NIM[2]. The URL to obtain the maintenance level is:

`http://techsupport.services.ibm.com/rs6k/fixes.html`

[2] Network Installation Manager (NIM)

If you have problems downloading the latest maintenance level, ask your IBM Business Partner or IBM representative.

## Investment Protection and Expansion

The following sections discuss how configurations, upgrades, and design features help you lower your cost of ownership.

### High Availability

Reliability of the system is further hardened by using the HACMP clustering solution available across the entire range of RS/6000 servers. The HACMP solution exploits redundancy between server resources and provides application uptime. The Model F80 is available in a high-availability cluster solution package named the HA-F80. This solution consists of the following components:

- Two Model 7025-F80 Enterprise Servers
- AIX Version 4.3.3 operating system (unlimited user license), or later
- HACMP 4.3.1 cluster software, or later
- One 7133-T40 SSA disk subsystem with at least four disk drives
- All necessary redundant hardware and cables

This solution is sold at a price lower than the sum of its parts. Ask your IBM Business Partner or IBM representative for further information.

### Reliability, Availability, and Serviceability (RAS) Features

Some RAS features such as redundant power supplies or N+1 hot-plug fans are already discussed. Additional topics are covered in the following sections.

#### Error Recovery for Caches and Memory

The RS64 III processor L1 cache, the L2 cache, system busses, and the memory are protected by error correction code (ECC) logic. The ECC codes provide single bit error correction and double bit error detection for the L2 cache and the memory. All recovered error events are reported by an attention interrupt to the service processor, where they are monitored for threshold conditions.

The standard memory card has single error-correct and double-error detect ECC circuitry to correct single-bit memory failures. The double-bit detection helps maintain data integrity by detecting and reporting multiple errors beyond what the ECC circuitry can correct. In many cases (using DIMMs with 18 DRAM chips and when memory is configured in quads, for example), memory chips are organized such that the failure of any specific memory module only affects a single bit within an ECC word (bit scattering) thus allowing for error correction and continued operation in the presence of a complete chip failure (chip kill recovery).

Another function, named *memory scrubbing*, provides a built-in hardware function, which performs continuous background reads of data from memory, checking for correctable errors. Correctable errors are corrected and rewritten to memory, and a threshold counter is maintained that will signal the service processor with a special attention when the threshold is exceeded.

### Dynamic CPU Deallocation

The processors are continuously monitored for errors such as L2 cache ECC errors. When a predefined error threshold is met, an error log with warning severity and threshold exceeded status is returned to AIX. At the same time, the service processor marks the CPU for deconfiguration at the next boot. In the meantime, AIX will attempt to migrate all resources associated with that processor (tasks, interrupts, etc.) to another processor, and then stop the failing processor.

The capability of dynamic CPU deallocation is only active in systems with more than two processors, because device drivers and kernel extensions, which are common to multi-processor and uni-processor systems would change their mode to uni-processor mode with unpredictable results.

### Persistent CPU and Memory Deconfiguration

CPUs and memory modules with a failure history are marked *bad* to prevent them from being configured on subsequent boots. This history is kept in the VPD[3] records on the FRU[4], so the information moves physically with the FRU and is cleared when the FRU is replaced, and stays with the failed FRU when it is returned to IBM. A CPU or memory module is marked bad when:

- It fails BIST[5]/POST[6] testing during boot (as determined by the service processor).
- It causes a machine check or check stop during runtime and the failure can be isolated specifically to that CPU or memory module (as determined by the service processor).
- It reaches a threshold of recovered failures (for example, ECC correctable L2 cache errors, see the preceding) that result in a predictive call-out (as determined by service processor).

During CEC initialization, the service processor checks the VPD values and does not configure CPUs or memory that are marked bad, much in the same way that it would deconfigure them for BIST/POST failures.

### I/O Expansion (RIO) Recovery

The RIO interface supports packet retry on its interface, which means that it will automatically try to resend a packet if it gets no acknowledgment or a bad response until a time-out threshold is reached.

RIO also supports a closed loop topology configuration, which is required for RS/6000 products. RIO hubs will automatically attempt to reroute packets through the alternate RIO port if a successful transmission cannot be completed (for example, the retry threshold is exceeded) through the primary port. Therefore, no single link failure in the RIO loop will cause the system to go down, although the failure will be reported for deferred maintenance.

### PCI Bus Error Recovery

As described in the PCI slot section, every slot is connected through a PCI-to-PCI bridge chip to a primary PCI bus; thereby, each slot is logically and physically isolated onto its own individual PCI bus. This fact provides a special error

[3] Vital Product Data (VPD)
[4] Field Replaceable Unit (FRU)
[5] Built-in self-test (BIST)
[6] Power-on self-test (POST)

handling mode that allows the bridge chip to *freeze* access to an adapter when a PCI bus error occurs on the interface between that adapter and bridge chip. In this frozen mode, DMAs[7] are blocked, stores to that device address space are discarded, and loads result in a return value of all 1s. Device drivers can be programmed to look for these dummy responses on loads and can attempt recovery. The AIX support for this function is not available yet.

### System Power Control Network (SPCN)

SPCN consists of a set of power/environmental controllers, interconnected by a set of serial communication links. In Model F80 systems, the SPCN function is integrated into the service processor and provides the following functions:

- Powering all the system parts up or down, when requested.

  The SPCN hardware has connections to the VPD that is resident on each of the pluggable cards and the backplane. The VPD is located on each of the cards in the form of an $I^2C$ chip. This chip is accessed during initial power on sequence and the data contents are read by the service processor. Using this function, the service processor decides not to use components that are marked *bad.*

- Powering down all the system parts on critical power faults
- Monitors power, fans, and thermal conditions in the system for problem conditions, which result in an EPOW. EPOW stands for environmental and power off warnings and is a function to inform the service processor or the operating system early, about an event that happened in the hardware. There are different warnings; such as cooling warnings or power fail warnings which result in entries in the error log. If there is a serious error, such as the temperature reaches a specific limit, the system will be shutdown.
- Reporting power and environmental faults, as well as faults in the SPCN network itself, on operator panels and through the service processor
- Assigning and writing location information into various VPD elements in the system.

### Disk Redundancy (Mirroring, RAID, Dual Controllers)

RS/6000 and AIX provide a number of options for increasing the robustness of storage subsystems, all of which involve some level of redundancy of disks and/or adapters.

AIX disk mirroring provides the ability to define transparent double or triple redundancy of disk data by mapping disk write data to two or three physical disks. On disk reads, the request is issued to all disks in the mirror group, and the first error-free response is returned, which also has some performance benefits. If one of the disks fails, the data is still readable from the other disk(s).

There are also customer options for SCSI and SSA RAID controller adapters, which can provide the same protection with better performance and less redundancy overhead. Also available are storage subsystems that provide under-the-covers redundancy for high availability.

To provide protection against adapter failures, AIX also supports dual-controller options where the same disk subsystem can be accessed through both a primary adapter path and through a backup adapter path if the primary fails.

[7] Direct memory access (DMA)

### Hot Swap Disk and Service Aid

The hardware within the system is designed with the capability to remove and install disks without powering down the system.

An AIX Diagnostics Service Aids provides positive identification (a blinking LED) at the disk device as a visual aid for removal.

## Service Processor

The Model F80 has an integrated enhanced service processor. When the system is powered down, but still plugged into an active power source, the service processor and SPCN functions are still active under standby power. This function provides enhanced RAS by not requiring AIX to be operational for interfacing with a system administrator or service director for RS/6000. This means that all service processor menu functions (using the local, remote, or terminal concentrator console), as well as dial out capability, are available even if the system is powered down or unable to power up. The next sections talk about selected features of the enhanced service processor.

### Automatic Reboot

The system will automatically reboot (if the appropriate policy flags are set) in the following conditions:

- Power is restored after a power loss during normal system operation.
- Hardware checkstop failures.
- Machine check interrupt.
- Operating system hang (Surveillance failure).
- Operating system failure.

### Surveillance

The service processor, if enabled through service processor setup parameters, performs a surveillance of AIX through a heartbeat mechanism. If there is no heartbeat within the time-out period, the service processor does the following:

- Creates a system reset to allow an AIX dump to occur.
- Upon receiving a reboot request (either after the dump, or immediately if dump is not enabled), the service processor captures scan debug data for the system.
- Reboots the system.

### Dial-Out (Call Home), Dial-In

If enabled, the service processor can dial a preprogrammed telephone number to report errors. When enabled, it is also possible to access the service processor remotely through a modem connection. When the service processor is in standby mode, because the system is powered off, or an error occurred, the service processor monitors an incoming phone line to answer calls, prompts for a password, verifies the password, and remotely display the standby menu. The remote session can be mirrored on the local ASCII console if the server is so equipped and the user enables this function.

### Processor and Memory Boot Time Deconfiguration

As described previously, processors can be dynamically deconfigured by the system. It is also possible to deconfigure processors and also memory with menus of the service processor for benchmarking reasons. For further information, refer to the *RS/6000 Enterprise Server Model F80 Service Guide*, SA38-0568.

> **Note**
>
> If the memory is to be temporary deconfigured (for benchmarking or sizing, for example), it is also possible to use the AIX *rmss* command to simulate a specific amount of memory (only below the real memory limit).

### Fast Boot

This feature, set as the default, allows you to select the IPL type, mode, and speed for your boot capabilities using service processor menus. Selecting fast boot results in several diagnostic tests being skipped and a shorter memory test being run; therefore, the startup process is faster, but possible problems might not be discovered at startup.

### Service Processor Restart

The service processor design for the Model F80 includes the ability to reset the service processor. This enables the system firmware to force a hard reset of the service processor if it detects a loss of communication. Since this would typically occur while the system is already up and running, the service processor reset will be accomplished without impacting system operation.

### Boot to SMS Menu

The Boot Mode menu allows to select among other things to boot to SMS menu. This function provides booting into SMS menu without pressing a key. This function is useful, because it is not necessary to wait in front of the system and press the **F1** (graphic display) or **1** (ASCII terminal) at the right moment.

## System Upgrades

For owners of a 7025-F50, it is possible to upgrade to a Model F80 while keeping the original serial number.

Model F50 systems converted to Model F80 systems will require replacement of all system processors. The first Model F50 processor card can be replaced with one Model F80 2-way or greater processor card via feature conversion. This processor conversion is available at the time of initial model upgrade only. A maximum of one F50 processor card may be converted to one F80 processor card for each system being converted. The existing Model F50 processor being replaced is returned to IBM.

Memory DIMMs from the F50 can move to the Model F80. Keep in mind, if the DIMMs will be installed on the memory cards, that they have to be installed in quads. Therefore it might be necessary to order two additional DIMMs to complete the quad.

Most of the adapters can also move from Model F50 to Model F80. Concerning the graphics accelerators, only a GXT120P or GXT130P is supported in Model

F80. For further information about adapters, especially if the adapter is supported in Model F80, refer to the *PCI Adapter Placement Reference Guide*, SA38-0538.

Most disk drives can move into the Model F80, but the hot-plug carriers used in the Model F80 are new to RS/6000. Therefore, drive migration from the Model F50 requires removal of the drive from the old blue-handled carrier and assembly into the new Model F80 carrier. Migration of 68-pin drives will limit the performance of the bus to which they are attached to SE 40 MB/s, rather than the LVD 80 MB/s, which is possible for an entire complement of 80-pin Ultra2 drives. Since the Model F80 has fewer disk bays, migration of data to larger disks may be required during the upgrade. Refer to the *RS/6000 Systems Handbook 2000*, SG24-5120 (available June 2000), or ask your IBM Business Partner or IBM representative for further information.

The old chassis of Model F50 is to be returned to IBM.

When upgrading from an Model F50 that does not run AIX 4.3.3, it is required to upgrade to AIX 4.3.3 or higher before you can upgrade the F50 system.

### External Storage Expandability

The storage expansion for the Model F80 is can be provided through several IBM storage options. The storage subsystems can be connected externally as stand-alone tower or from within a rack.

External disk storage capacity can also be provided by attaching the Model F80 to storage servers. Using differential Ultra SCSI, the Model F80 can be attached to the IBM Enterprise Storage Server. And by using the Fibre Channel Adapter, the Model F80 can be attached to the IBM Fibre Channel RAID Storage server or the IBM Enterprise Storage Server.

### SP Attachment

The Model F80 can neither be attached as a logical node, nor is it supported to be used as a Control Workstation (CWS) in an SP system.

## Reference

The following sections list additional materials available for further research.

### System Documentation

For more detailed information, refer to the following documents:

- *RS/6000 Enterprise Server Model F80 Installation Guide*, SA38-0569
- *RS/6000 Enterprise Server Model F80 User's Guide*, SA38-0567
- *RS/6000 Enterprise Server Model F80 Service Guide*, SA38-0568
- *PCI Adapter Placement Reference Guide*, SA38-0538

### Select IBM Redbooks

The following IBM Redbooks are related to the material discussed in this paper:

- *RS/6000 Systems Handbook 2000*, SG24-5120 (Available June 2000)
- *RS/6000 S-Series Enterprise Servers Handbook*, SG24-5113

- *IBM Enterprise Storage Server*, SG24-5465
- *Monitoring and Managing IBM SSA Disk Subsystems*, SG24-5251
- *AIX 4.3 Differences Guide*, SG24-2014
- *NIM: From A to Z in AIX 4.3*, SG24-5524
- *AIX Logical Volume Manager, from A to Z: Introduction and Concepts, SG24-5432*
- *Understanding IBM RS/6000 Performance and Sizing*, SG24-4810

### Select Internet Links

For more detailed information, see the following Web sites:

```
http://www.rs6000.ibm.com/
http://www.rs6000.ibm.com/hardware/enterprise/
http://www.rs6000.ibm.com/resource/hardware_docs/index.html
http://www.rs6000.ibm.com/cgi-bin/ds_form
http://www.rs6000.ibm.com/support/micro/
http://www.ibm.com/servers/aix/
http://www.chips.ibm.com/
http://www.research.ibm.com/topics/serious/chip/
http://www.storage.ibm.com/
http://www.hursley.ibm.com/~ssa/rs6k/
http://www.redbooks.ibm.com/
```

## Acknowledgements

Assistance creating this white paper came from the following individuals and was appreciated. Thank you!

| | |
|---|---|
| Amy, Larry | IBM Austin |
| Babnik-Gomiscek, Jana | IBM Slovenia |
| Beebe, Bill | IBM Austin |
| Enders, Rich | IBM Austin |
| Frey, Brad | IBM Austin |
| Haug, Volker | IBM Germany |
| Henderson, Daniel | IBM Austin |
| Lee, Dennis | IBM Austin |
| McCord, Scott | IBM Austin |
| Ruth, Dave | IBM Austin |
| Scherrer, Carolyn | IBM Austin |
| Shempert, Craig | IBM Austin |
| Stys, Mike | IBM Austin |
| Toutant, Ed | IBM Austin |

## Biographies

**Stephen Lutz** is an IT Specialist in the technical support for RS/6000 and NUMA-Q, part of the Web Server Sales Organization in Stuttgart, Germany. He holds a degree in Computer Science from the Fachhochschule Karlsruhe - University of Technology and is an IBM Certified Advanced Technical Expert. Stephen is a member of the High-End Technology Focus Group, supporting IBM sales, Business Partners, and customers with pre-sales consultation and implementation of client/server environments.

**Shyam Manohar** is a Marketing Manager from IBM, India with the Web Server Group in ESG. Shyam is one of the leading top contributors from India for the IBM-ASEAN region. With his valuable experience in the RISC/UNIX marketplace, Shyam holds unique expertise in product positioning, especially in competitive scenarios. He is part of a focus team engaged in large high-end server opportunities.

## Special Notices

IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. The use of this information or the implementation of any techniques described herein is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. Customers attempting to adapt these techniques to their own environments do so at their own risk.

IBM is not responsible for printing errors in this publication that result in pricing or information inaccuracies.

The information contained in this document represents the current views of IBM on the issues discussed as of the date of publication. IBM cannot guarantee the accuracy of any information presented after the date of publication.

All prices shown are IBMs suggested list prices; dealer prices may vary.

IBM products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Information provided in this document and information contained on IBMs past and present Year 2000 Internet Web site pages regarding products and services offered by IBM and its subsidiaries are "Year 2000 Readiness Disclosures" under the Year 2000 Information and Readiness Disclosure Act of 1998, a U.S statute enacted on October 19, 1998. IBMs Year 2000 Internet Web site pages have been and will continue to be our primary mechanism for communicating year 2000 information. Please see the "legal" icon on IBMs Year 2000 Web site (www.ibm.com/year2000) for further information regarding this statute and its applicability to IBM.

Any performance data contained in this document was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

The following terms are registered trademarks of International Business Machines Corporation in the United States and/or other countries: ADSTAR, AIX, AIX/6000, AIXwindows, AS/400, C Set++, CICS, CICS/6000, DB2, ESCON, IBM, Information Warehouse, Intellistation, LANStreamer, LoadLeveler, Magstar, MediaStreamer, Micro Channel, MQSeries, Net.Data, Netfinity, OS/2, OS/400, OS/390, Parallel Sysplex, POWERparallel, RS/6000, S/390, Service Director, System/390, ThinkPad, TURBOWAYS, VisualAge. The following terms are trademarks of International Business Machines Corporation in the United States and/or other countries: AIX PVMe, AS/400e, DB2 OLAP Server, DB2 Universal Database, DEEP BLUE, e-business (logo), eNetwork, GigaProcessor, HACMP/6000, Intelligent Miner, Network Station, POWER2 Architecture, PowerPC 604, SmoothStart, SP, Videocharger, Visualization Data Explorer, WebSphere. A full list of U.S. trademarks owned by IBM may be found at http://iplswww.nas.ibm.com/wpts/trademarks/trademar.htm.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks or registered trademarks of Microsoft Corporation in the United States, other

countries, or both. UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group. Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both. Lotus and Lotus Notes are trademarks or registered trademarks of Lotus Development Corporation. Tivoli, TME, TME 10, and TME 10 Global Enterprise Manager are trademarks of Tivoli Systems, Inc. Other company, product, and service names may be trademarks or service marks of others.