IBM ^ pSeries 690

Configuring for Performance

Harry M. Mathis, John D. McCalpin, Men-Chow Chiang, Frank P. O'Connell, Pat Buckland
IBM Server Group

IBM *October, 2001*

Configuring pSeries 690 for Performance

The IBM ^ pSeries™ 690 server introduces IBM's new POWER4 chip technology to the worlds of technical computing and commercial server workloads. In addition to the speed advantages provided by the significantly faster POWER4 processor, the pSeries 690 takes advantage of a new system architecture, and a faster system bus, memory subsystem and input/output (I/O) subsystem. Appropriately configuring the pSeries 690 server to take advantage of its speed and architecture will enable the customer to maximize the benefits of owning this leading-edge machine. This paper describes how the pSeries 690 system should be configured to provide optimum performance for different customer workloads.

The pSeries 690 Server

The pSeries 690 datacenter server, the latest and most powerful UNIX® server IBM has ever offered, brings more than just higher processor frequencies—it brings a complete new system architecture.

The pSeries 690 is the first in a family of IBM servers implementing the POWER4 system architecture, which includes the microprocessor, its associated Central Electronic Complex (CEC), memory subsystem and input-output (I/O) subsystem.

Initially the pSeries 690 is available with three versions of the POWER4 Multi-Chip Module (MCM), the basic building block for the pSeries 690 :

- an 8-way MCM feature running at 1.1 GHz
- an 8-way MCM Turbo feature running at 1.3 GHz
- a 4-way MCM HPC (High Performance Computing) feature running at 1.3 GHz.

The systems using the 8-way MCM feature are optimized for commercial workloads, such as those found in database, Web server and transaction processing systems, as well as for many types of technical workloads, such as those found in engineering and scientific environments. The systems using the 4-way MCM feature are optimized for the more data-intensive technical workloads that have larger memory bandwidth requirements per process. They use POWER4 chips that contain one rather than two microprocessors, hence each MCM consists of four POWER4 chips with a total of four microprocessors. These MCMs have the full complement of L2 and L3 caches and therefore provide twice the cache per processor as the 8-way MCM feature. This paper will assist information systems decision makers in planning and configuring the pSeries 690 so it will provide optimum performance for the expected workloads.

pSeries 690 System Architecture

The following provides a brief overview of the pSeries 690 system architecture. A detailed discussion of the POWER4 chip, its development, its architecture and its operation, is provided by the IBM Server Group White Paper, "***POWER4 System Microarchitecture***". A comprehensive explanation of POWER4 performance, including FORTRAN code examples and performance measurements, is provided in the IBM International Technical Support Organization (ITSO) Red Book, ***"POWER4 Introduction and Tuning Guide, Technical and Scientific Computing"***, to be published 12/2001.

The pSeries 690 system architecture is implemented through the POWER4 Central Electronic Complex (CEC). Logically, the CEC consists of the microprocessors, pervasive functions and the storage subsystem. Physically, the CEC consists of the microprocessor chip, the Level 3 (L3) cache chip and the memory controller chip, which controls main memory.

At the heart of the CEC is the POWER4 chip, which contains: either one or two microprocessors; the L2 cache running at the same speed as the microprocessors; the microprocessor interface unit, which is the interface for each microprocessor to the rest of the system; the directory and cache controller for the L3 cache; the fabric bus controller, which is at the heart of the system's interconnection design; and a GX bus controller that enables I/O devices to connect to the CEC.

The second component of the POWER4 CEC is the L3 cache, comprised of two 16MB eDRAM chips mounted on a separate module. Each POWER4 chip controls an L3 cache, connected between the POWER4 chip and the memory controller chip.

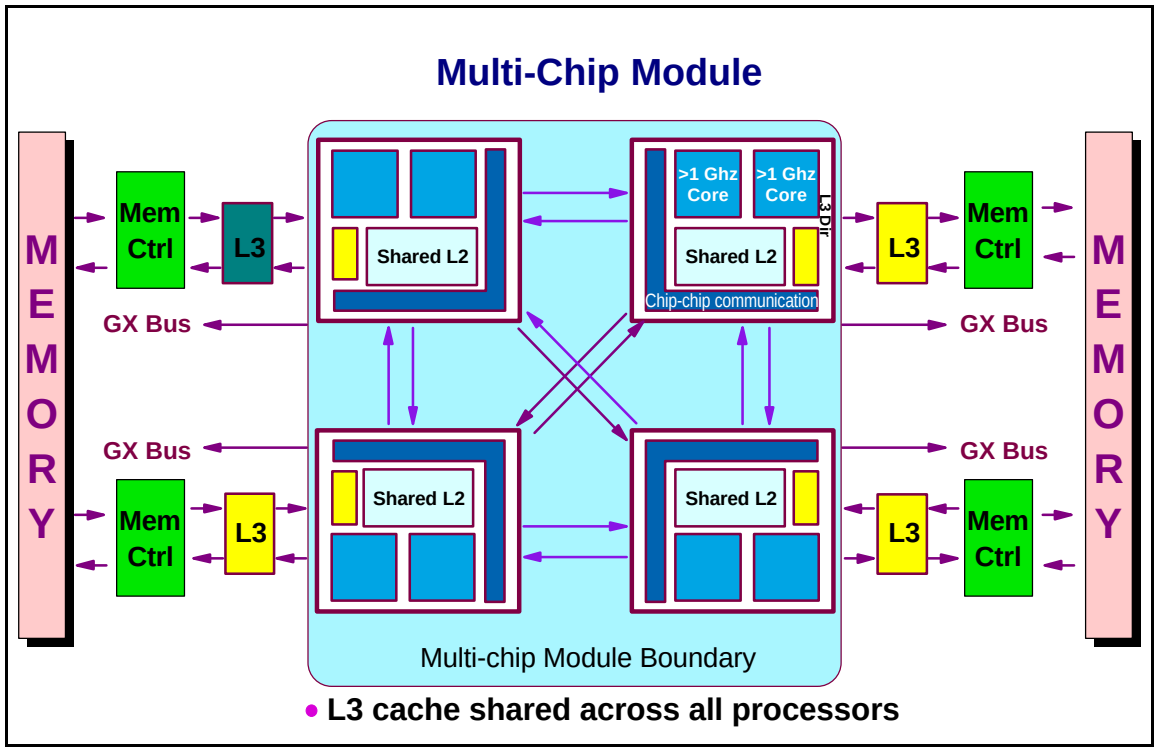

Figure 1. The Multi-Chip Module

The third component of the POWER4 CEC is the memory controller chip. It is connected to the L3 cache on one side and to synchronous memory interface (SMI) chips on the other to

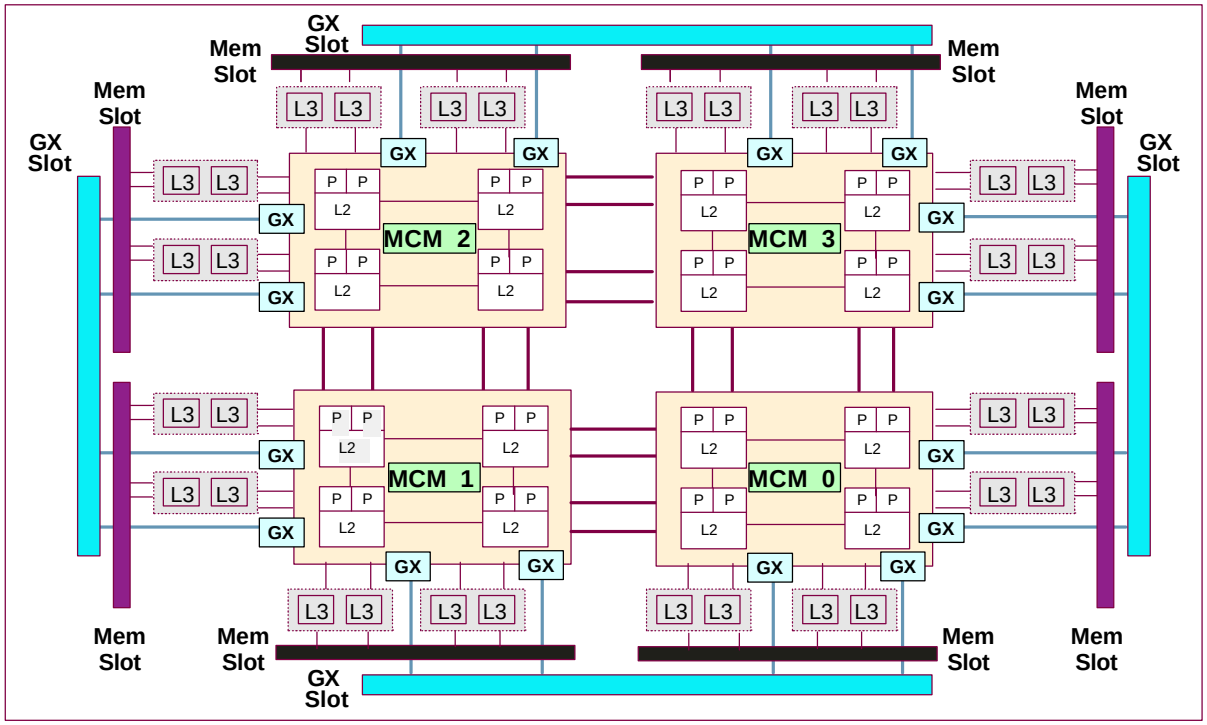

Figure 2. Four 8-way MCM Features Assembled into a 32-way pSeries 690

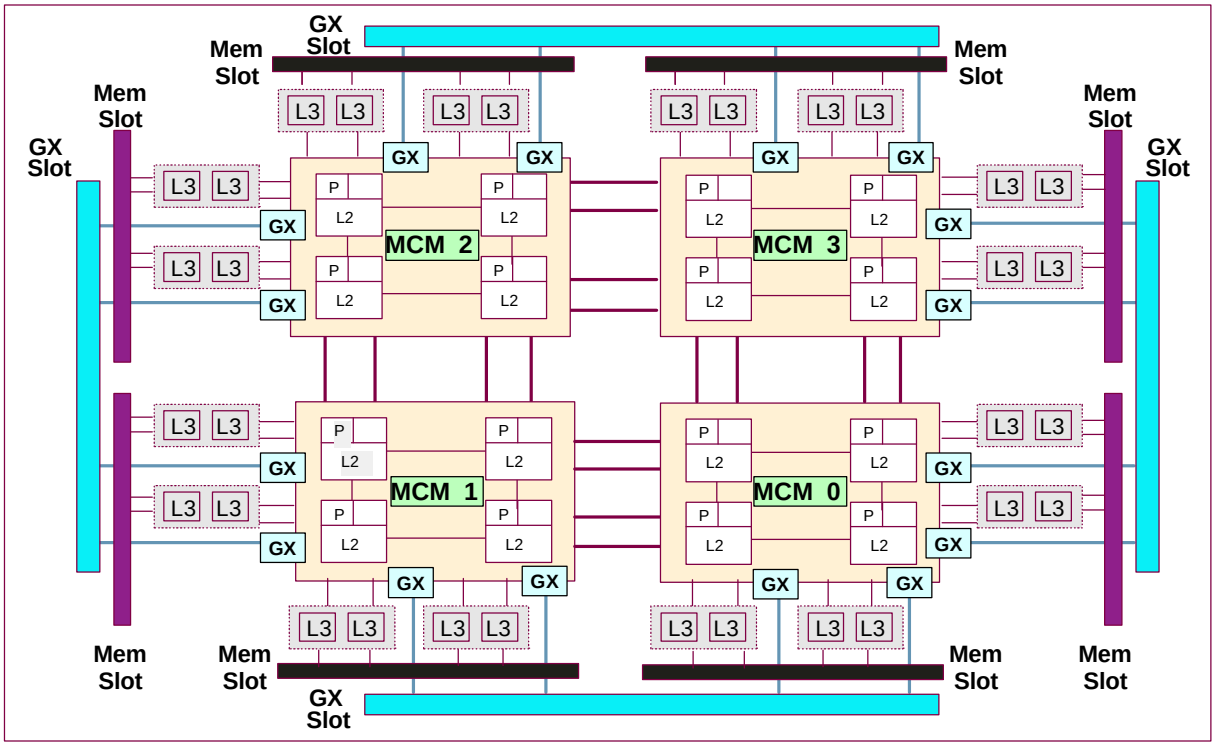

Figure 3. Four 4-way MCM Features Assembled into a 16-way pSeries 690 HPC

control main memory. Each memory controller chip can have one or two memory data ports and can support up to 16GB of memory. There is a separate memory controller for each POWER4 chip. Two memory controllers are packaged on each memory card, and a maximum of two memory cards can be attached to each MCM. In all system configurations, all memory and all I/O is transparently accessible to all processors.

The basic building block for pSeries 690 systems, the MCM, is shown in Figure 1. Each MCM contains four interconnected POWER4 chips, each with its own off-chip L3 cache.

Configurations

The pSeries 690 can contain up to 4 MCMs. Each MCM comprises either a 4-way or an 8-way symmetric multiprocessing (SMP) unit, depending on whether one or two microprocessors are present on each of the POWER4 chips. Table 1 lists the four available configurations with the 8-way MCM feature, and Table 2 lists the two available configurations with the 4-way MCM feature. Figure 2 shows a detailed layout of a 32-way pSeries 690 server configuration.

Figure 3 shows a detailed layout of a 16-way pSeries 690 HPC configuration. Note that this system has the same amount of cache, memory slots, memory bandwidth, and Remote I/O (RIO) ports as the 32-way system built from 8-way MCMs, but is configured with only half the number of processors. For workloads that have larger working sets per process or thread, or that have high bandwidth requirements, capable of saturating the memory subsystem when using fewer than eight processes or threads per MCM, this configuration is likely to provide superior price/performance.

Processors	MCMs	L3	Memory Slots Usable	RIO Ports
8	1	4	2	4
16	2	8	4	8
24	3	12	6	12
32	4	16	8	16

Table 1: 8-Way MCM Feature pSeries 690 Configurations

Processors	MCMs	L3	Memory Slots Usable	RIO Ports
8	2	8	4	8
16	4	16	8	16

Table 2: 4-way MCM Feature pSeries 690 HPC Configurations

Configuring the Memory Subsystem

Memory Sizing for the pSeries 690

The optimal amount of memory for a particular system depends upon many factors, not the least being the requirements of the key applications targeted for the system. It is important to note, however, that the size and number of memory cards in the system will determine the maximum system bandwidth that the system can deliver.

Memory Balancing

To maximize memory performance on pSeries 690 systems, memory interleaving is employed. If an MCM has two memory cards of the same size installed, memory is interleaved in a round-robin fashion across the four memory controllers with 512-byte granularity. Referring to Figure 1, assuming that the first 512-byte block of memory is in the memory card on the left of the MCM and is accessed by the L3 in the lower left corner of the MCM, the second 512-byte block is accessed by the L3 in the upper left corner, the third 512-byte is accessed from the memory card on the right side of the MCM by the L3 in the upper right corner, and the fourth 512-byte block is accessed by the L3 in the lower right corner. This continues throughout the memory range afforded by the two cards, and each L3 handles only one-fourth of the memory addresses for that MCM.

If an MCM has one memory card attached (for example, in Figure 1, if the memory card on the left is present but the memory card on the right is not present), then the memory is interleaved only on the single memory card by the two L3's on the left side of the MCM. The command queues associated with these two L3's must then process twice the traffic they would in the first case for the same application, and this reduces the bandwidth available for the MCM.

If an MCM has two memory cards of different sizes attached, then the two cards are treated independently with each card being two-way interleaved. For example, if an 8GB and a 32GB card have been installed in the MCM's memory slots, eighty percent of the data for an application will be handled by two of the L3's and twenty percent of it will be handled by the other two L3's, and this will reduce the effective bandwidth as a result of uneven use of the L3 command queues.

Memory Card Size	Ports	SMI's / Port	DIMM's / SMI
4GB	1	4	1
8GB	1	4	2
16GB	2	4	2
32GB	2	4	2 (Stacked)

Table 3: Memory Card Characteristics

Memory card size can also impact system performance. As shown in Table 3, 4GB and 8GB memory cards have 1 port between each memory controller and memory, whereas 16GB and 32GB cards have 2 ports between each of the memory controllers and memory and accordingly support a greater bandwidth.

Since the pSeries 690 supports two memory cards per MCM, a maximum of eight cards can be plugged into a fully populated system of four MCMs. With card sizes of 4GB, 8GB, 16GB and 32GB memory cards available, the customer's options range from a system containing 8GB to one containing 256GB. For performance reasons, the minimum amount of memory is restricted to 8GB. The preferred configuration is two 4GB memory cards rather than one 8GB memory card. For the data center applications that the pSeries 690 is designed to support, the recommended configuration for an 8-way system is at least 16GB. Table 4 provides suggested memory card placement for various pSeries 690 systems. (*Note that at present 4GB and 16GB memory cards cannot be intermixed with each other or with 8GB and 32GB memory cards.*)

Model/N-way	Memory	MCM 1	MCM 2	MCM 3	MCM 4
pSeries 690 8-way	**8GB**	**4 & 4**			
	16GB	**8 & 8**			
	32GB	**16 & 16**			
	64GB	**32 & 32**			
pSeries 690 16-way	**All 8-Way +**				
	16GB	**4 & 4**	**4 & 4**		
	32GB	**8 & 8**	**8 & 8**		
	64GB	**16 & 16**	**16 & 16**		
	80GB	**32 & 32**	**8 & 8**		
	128GB	**32 & 32**	**32 & 32**		
pSeries 690 24-way	**All 16-way +**				
	24GB	**4 & 4**	**4 & 4**	**4 & 4**	
	48GB	**8 & 8**	**8 & 8**	**8 & 8**	
	96GB	**8 & 8**	**8 & 8**	**32 & 32**	
	96GB	**16 & 16**	**16 & 16**	**16 &16**	
	144GB	**8 & 8**	**32 & 32**	**32 & 32**	
	192GB	**32 & 32**	**32 & 32**	**32 & 32**	
pSeries 690 32-way	**All 24-way +**				
	32GB	**4 & 4**	**4 & 4**	**4 & 4**	**4 & 4**
	64GB	**8 & 8**	**8 & 8**	**8 & 8**	**8 & 8**
	112GB	**8 & 8**	**8 & 8**	**8 & 8**	**32 & 32**
	160GB	**8 & 8**	**8 & 8**	**32 & 32**	**32 & 32**
	208GB	**8 & 8**	**32 & 32**	**32 & 32**	**32 & 32**
	256GB	**32 & 32**	**32 & 32**	**32 & 32**	**32 & 32**
pSeries 690 HPC 8-way	**Same as pSeries 690 16-way**				
pSeries 690 HPC 16-way	**Same as pSeries 690 32-way**				

Table 4. Memory Card Placement for pSeries 690 and pSeries 690 HPC Servers

If the memory requirement is for an amount of memory not shown in Table 4, then the system should be configured by using matched pairs of memory cards having sufficient memory to

meet or exceed the requirement. Ideally, for overall data center support of a wide variety of applications, all available memory slots should be filled, with pairs of the same size memory cards used in all cases. Maximum sustainable memory bandwidth for an MCM is achieved when memory is equal to or greater than 32GB per MCM (2 x 16GB cards).

Workload Considerations

Technical and commercial workloads tend to have different access patterns, and this creates slightly different requirements for the placement of memory cards in pSeries 690 . The main memory access patterns for technical workloads often are sequential in nature, which prefetching built into the hardware and a high level of memory interleaving can exploit to achieve maximum memory bandwidth. As mentioned previously, for a given amount of main memory, a higher level of memory interleaving is achieved with a larger number of smaller memory cards than with a smaller number of larger memory cards. For example, a pSeries 690 with eight 16GB cards will offer higher bandwidth, hence better technical performance, than a system with four 32GB memory cards.

The main memory access patterns of commercial workloads tend to be more random in nature. Thus the strategy of spreading the access to more memory cards or increasing the interleaving factor with symmetric placement of memory cards has little effect on performance. The major factor impacting commercial workloads is memory size: commercial performance usually increases with the amount of the main memory.

Since technical workloads are more sensitive to memory card size and placement than commercial workloads are, the configuration effort for a pSeries 690 system that will be supporting both technical and commercial workloads should concentrate primarily on configuring for the technical workloads. In systems with mixed technical and commercial workloads, the need for bandwidth in the technical applications needs to be traded off against memory upgrade flexibility, which is often a consideration for systems used for commercial applications.

Input/Output Subystem

Topology

The pSeries 690 and pSeries 690 HPC have a leading-edge I/O subsystem that complements the POWER4 CEC. A schematic for the pSeries 690 and pSeries 690 HPC, including the I/O subsystem (as it will appear on 04/02) is shown in Figure 4. The minimum I/O configuration for the pSeries 690/pSeries 690 HPC includes a base I/O book, a single 1EIA media drawer and one 7040-61D 24 inch 4EIA I/O drawer. The base I/O book, which contains the service processor, supports up to two 7040-61D I/O drawers.

The maximum I/O configuration on 12/01 is six 7040-61D I/O drawers, and this capability will expand to eight 7040-61D I/O drawers on 04/02. This configuration requires the base I/O book and extra I/O books. The second I/O book supports up to four 7040-61D I/O drawers. The third I/O book supports an additional two 7040-61D I/O drawers. Note that two MCMs must be present before the second I/O book can be installed, and that with two MCMs present, only half of the second book's RIO ports are active. Three MCMs must be present before six

7040-61D I/O drawers can be supported. After 04/02 the full complement of eight 7040-61D I/O drawers can be supported on 16-way pSeries 690 HPC and 32-way pSerie s690 .

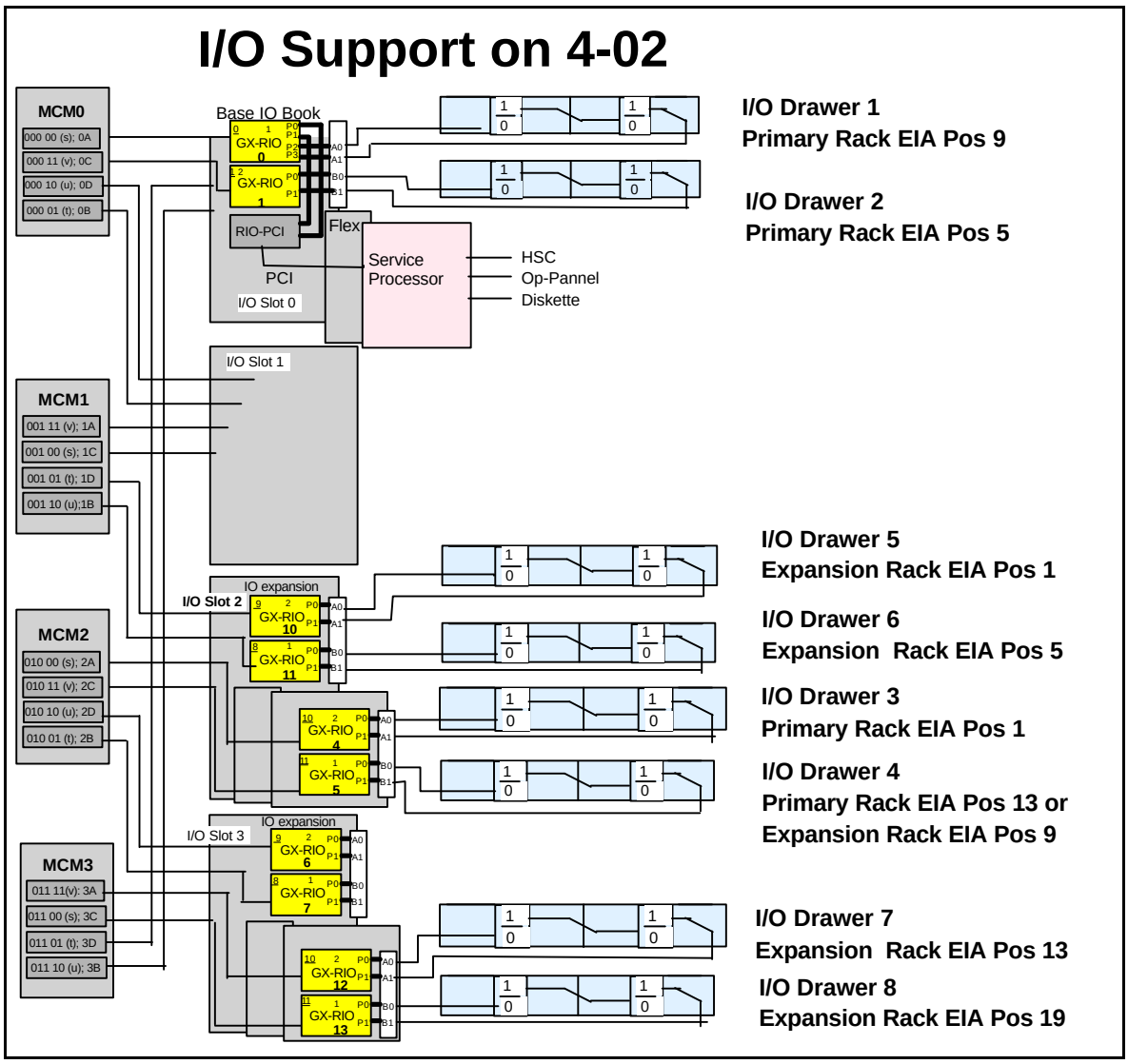

Figure 4. pseries 690 / pSeries 690 HPC I/O Subsystem

The media drawer contains an operator panel, a floppy disk drive and up to four SCSI attached media devices. The operator panel and the diskette drive are cabled to the service processor. The SCSI media is split into two groups of two devices. The media is powered from the first 7040-61D I/O drawer and the SCSI devices are cabled to SCSI adapters in PCI slots in the first I/O drawer. The supported media devices are CD-ROM, DVD-RAM and 4 mm Tape.

The 7040-61D I/O drawer, represented within Figure 5, is a 4 EIA drawer that contains support for 20 full length PCI adapters and 16 hard drives. It contains two PCI I/O planars that have ten 64-bit PCI slots, and two integrated Ultra3 SCSI controllers each. The first seven of the PCI slots have 3.3V PCI bus signaling and will support operating at 66 MHz or 33 MHz depending on the adapter. The last three PCI slots have 5V PCI bus signaling and operate at 33 MHz. Seven of the PCI slots and one of the Ultra3 SCSI controllers for each planar are

connected to a 64-bit PCI Host Bus (PHB), while the other three of the PCI slots and the other Ultra3 SCSI controller are connected to a 32-bit PHB. The PHBs in turn are connected to the RIO-to-PCI bridge of the planar. Each I/O planar has two RIO ports. One RIO port of each I/O planar is connected to a RIO port on an I/O book in the CEC. The other RIO port is connected to the other I/O planar. Normally each I/O planar uses its direct connection to the CEC for I/O, but if there is a RIO failure an I/O planar can route data through the connection to the other I/O planar and share the remaining RIO cable. Each integrated Ultra3 SCSI adapter is internally hard wired to a 4-slot DASD back plane in the front of the I/O drawer so that there are four groups of four hard drives.

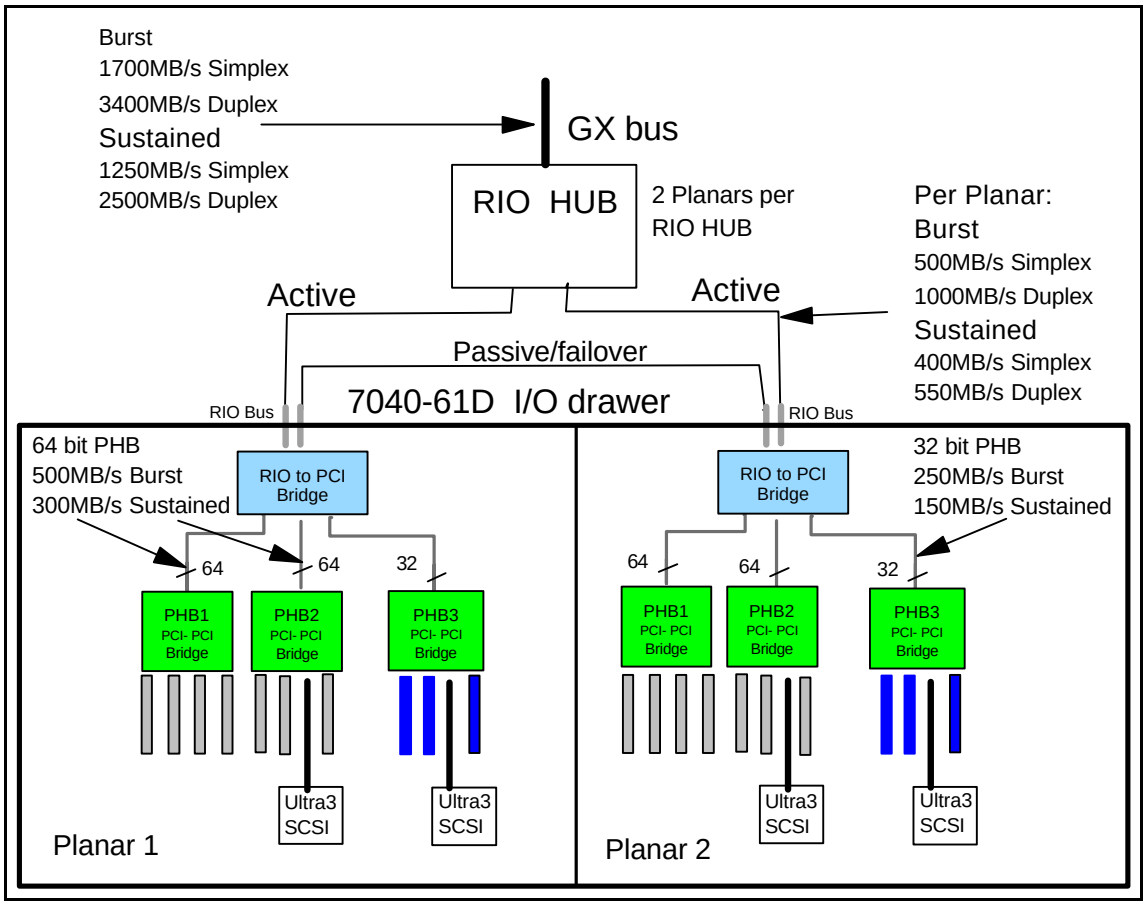

Figure 5. The I/O Subsystem and Bandwidths for each Stage

I/O Sizing

I/O sizing involves determining what workloads the pSeries 690 will be supporting and the expected I/O bandwidth for each. If it is a replacement for an existing system, then calculating the I/O bandwidth to be supported should be straightforward: the projected bandwidth will equal the existing bandwidth modified by whatever application growth is expected.

If the pSeries 690 will be supporting new workloads, then calculating the expected I/O bandwidth becomes more complicated. First the expected workloads need to be identified and characterized. The technical and commercial workload categories used in discussions on

memory sizing can be used with the caveat that, from an I/O standpoint, not all technical workloads behave alike, and not all commercial workloads behave alike.

In general, technical workloads read and write large bursts of time-sensitive, sequential data for short periods during execution, and this imposes high peak bandwidth requirements on the I/O system. Many technical workloads have been optimized for execution to enable staging of time-sensitive data from I/O devices so that it will be available when the processors need it, and this tends to spread out I/O and reduce peak I/O requirements.

Commercial workloads, with the exception of business intelligence applications that behave more like technical applications, tend to require low I/O bandwidth. These workloads, such as transaction processing and Web server applications, normally have large numbers of reads and writes spread randomly over the I/O devices. While the volume of disk accesses for this category are large, the amount of data transferred for each access is relatively small (4k bytes to 16k bytes). Often database systems cache table indexes in memory for use during processing to minimize table reads, and this also tends to keep I/O bandwidth requirements low. Large volumes of random disk accesses leads to the need for large numbers of disk arms. It is common for systems supporting database systems to need so many disk arms that there is unused space on the disks containing the database. In this case, it is usually advisable to purchase the smallest capacity, high-speed disk drives available and go for large numbers of disks.

Web server workloads tend to behave like transaction processing workloads, with low data rates and random reads and writes across I/O devices, although Web applications tend to emphasize communications I/O bandwidth over disk I/O bandwidth. The dominant direction of data flow for communications adapters in a Web serving environment is out of the system (communications “Put” or “Send”). The dominant direction of data flow for a disk adapter in a Web serving environment is from the disk into the system memory (disk reads). Mixing both of these adapters on the same RIO hub will take better advantage of the full duplex natures of the 7040-61D I/O drawers, RIO hubs and GX buses.

Once each workload has been characterized, the next step is to identify the numbers and types of I/O devices that will be required to support each workload, and the amount of I/O bandwidth that will be handled by each device. Note that device peak bandwidth may be much higher than sustained bandwidth, and that peak loads rarely occur on all devices simultaneously. To minimize I/O latencies to the pSeries 690 processors and optimize the overall performance of the pSeries 690 , the adapters and I/O subsystems should be planned, both in numbers and through placement, to operate at 60% to 80% of their hardware capabilities.

Given the device bandwidths, it is relatively straightforward to determine the total I/O bandwidth required of the system. However, in addition to the maximum bandwidth that 7040-61D I/O drawers can support, there are adapter limits that must be considered. Once the total bandwidth requirements and the total numbers and types of adapters have been determined, the sizing process must match these with the bandwidth and numbers and types of adapters supportable by each 7040-61D I/O drawer. Table 5 lists some commonly used I/O

adapters and typical bandwidths that can be expected from each, along with the maximum numbers of each type that can be plugged into a 7040-61D I/O drawer's PCI slots.

	Bandwidth	64-bit PHBs (4)	32-bit PHBs (2)
SP Switch2 Adapters		**1 Per 64-bit PHB**	**0**
SP Switch2 FC 8937	Up to 200MB/s	4	0
High Performance PCI Adapters		**2 per 64-bit PHB**	**1 per 32-bit PHB**
2GB FC FC 6228	Up to 150MB/s	8 Total / Drawer	2 Total /Drawer
1GB Ethernet FCs 2969,2975	Up to 150MB/s	8 Total / Drawer	2 Total /Drawer
Dual Ultra-3 SCSI FC 6203	Up to 175MB/s	8 Total / Drawer	2 Total /Drawer
622 ATM FC 2946	Up to 100MB/s	8 Total / Drawer	2 Total /Drawer
SSA 40 FC 6230	Up to 90MB/s	3 on PHB1, 2 on PHB2 10 Total / Drawer	2 Total /Drawer

Table 5. Commonly Used Adapters, Bandwidths and Limits per 7040-61D Drawer

Table 6 lists the elements and the bandwidths they support at each stage of the I/O subsystem. The RIO hub stage, with its sustained duplex and simplex bandwidth rates of 1100MB/s and 800MB/s respectively provides the effective limit to the numbers of adapters that can be plugged into each 7040-61D I/O drawer. The pSeries 690 I/O architecture allows the bandwidth supported by the system to scale with the number of drawers attached. The total bandwidth required by a system will vary depending on the application.

Stage	Burst Duplex	Burst Simplex	Sustained Duplex	Sustained Simplex
64-bit PHB (each)		500MB/s		300MB/s
32-bit PHB (each)		250MB/s		150MB/s
PHB Total (6 PHBs)		2500MB/s		1500MB/s
RIO Hub	2000MB/s	1000MB/s	1100MB/s	800MB/s
GX Bus	3400MB/s	1700MB/s	2500MB/s	1250MB/s

Table 6. Bandwidth Capabilities by I/O Stage

Conclusions

The pSeries 690 and pSeries 690 HPC are at the leading-edge of technology, both in processor power and speed, and in I/O capabilities. There are configuration decisions that should be made that will enable the customer to get maximum performance from these systems. These actions are:

- Plan for sufficient memory to support the types of expected workloads
- Plan to pair memory cards of the same sizes in adjacent memory slots
- Plan for sufficient 7040-61D I/O drawers to support the I/O bandwidths expected and the numbers and types devices to be attached.

Special Notices

IBM Corporation
Marketing Communications
Server Group
Route 100
Somers, NY 10589

Produced in the United States of America
10-01

More details on IBM UNIX hardware, software and solutions may be found at: ibm.com/servers/eserver/pseries.

You can find notices, including applicable legal information, trademark attribution, and notes on benchmark and performance at ibm.com/rs6000/hardware/specnote.html

IBM, the IBM logo, the e-business logo, pSeries are registered trademarks or trademarks of the International Business Machines Corporation in the United States and/or other countries. The list of all IBM marks can be found at: http://iplswww.nas.ibm.com/wpts/trademarks/trademar.htm.

The [e(logo) server] brand consists of the established IBM e-business logo followed by the descriptive term “server”.

UNIX is a registered trademark in the United States and other countries, licensed exclusively through X/Open Company, Limited.

Other company, product and service names may be trademarks or service marks of others.

IBM may not offer the products, programs, services or features discussed herein in other countries, and the information may be subject to change without notice.

General availability may vary by geography.

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

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Any performance data contained in this document was determined in a controlled environment. Results obtained in other operating environments may vary significantly.