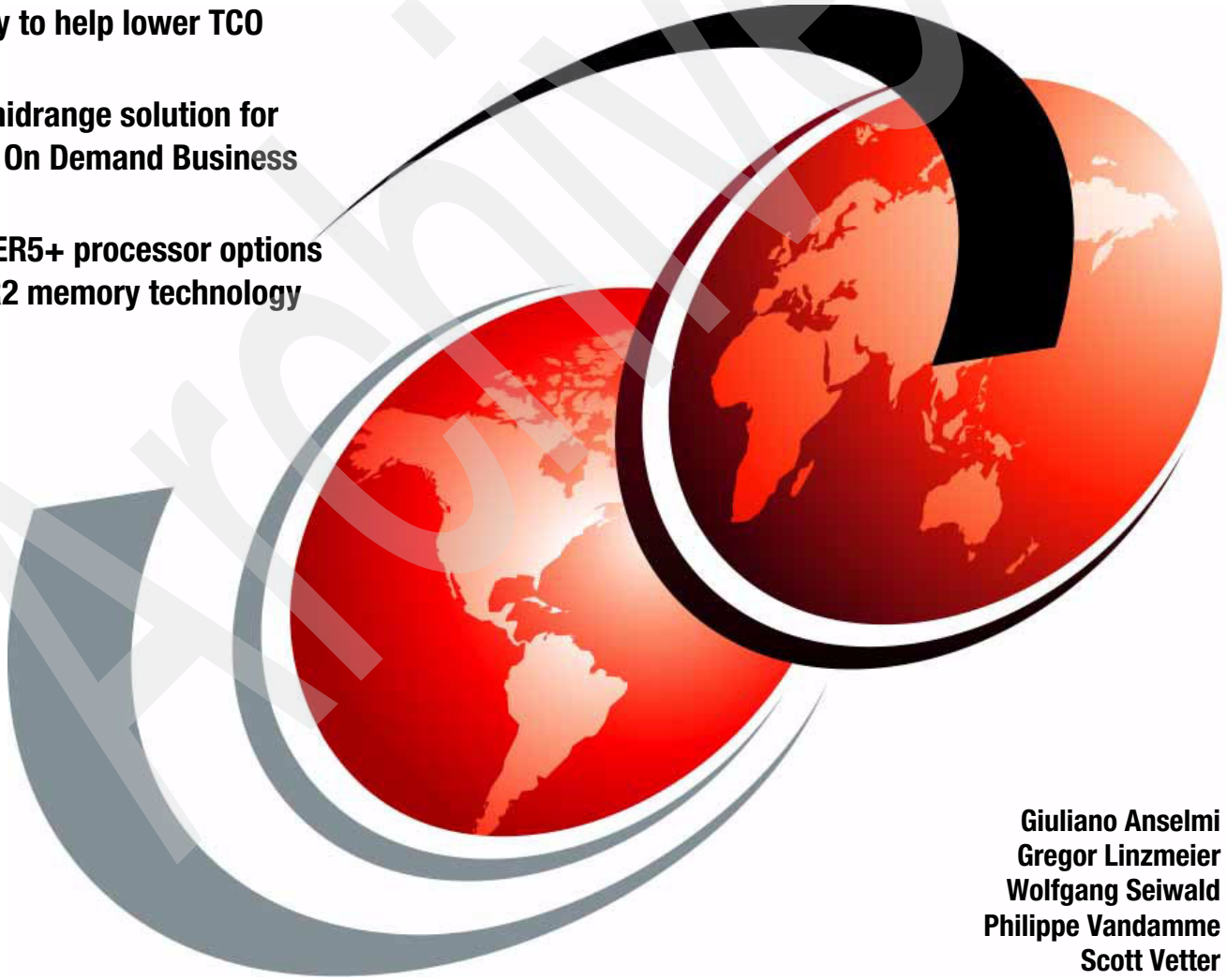
IBM

IBM System p5 570

Technical Overview and Introduction

- Finer system granulation using Micro-Partitioning technology to help lower TCO
- Modular midrange solution for managing On Demand Business
- New POWER5+ processor options using DDR2 memory technology



Giuliano Anselmi
Gregor Linzmeier
Wolfgang Seiwald
Philippe Vandamme
Scott Vetter

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM System p5 570 Technical Overview and Introduction

September 2006

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

Second Edition (September 2006)

This edition applies to the IBM System p5 570 (9117-570) and AIX 5L™ Version 5.3, product number 5765-G03.

Contents

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

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

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

1350™
AIX 5L™
AIX®
Chipkill™
DS4000™
DS6000™
DS8000™
Enterprise Storage Server®
eServer™
HACMP™
i5/OS®
IntelliStation®
IBM®
Micro-Partitioning™
OpenPower™
PowerPC®
POWER™
POWER Hypervisor™
POWER4™
POWER5™
POWER5+™
POWER6™
pSeries®
Redbooks™
Redbooks (logo) ™
RS/6000®
Service Director™
System p™
System p5™
System Storage™
TotalStorage®

The following terms are trademarks of other companies:

Internet Explorer, Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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

Preface

This IBM® Redpaper is a comprehensive guide covering the IBM System p5™ 570 UNIX® server. It introduces major hardware offerings and discusses their prominent functions.

Professionals wishing to acquire a better understanding of IBM System p5 products should read this Redpaper. The intended audience includes:

- Customers
- Sales and marketing professionals
- Technical support professionals
- IBM Business Partners
- Independent software vendors

This Redpaper expands the current set of IBM System p™ documentation by providing a desktop reference that offers a detailed technical description of the p5-570 system.

This Redpaper does not replace the latest marketing materials and tools. It is intended as an additional source of information that, together with existing sources, may be used to enhance your knowledge of IBM server solutions.

The team that wrote this Redpaper

This Redpaper was produced by a team of specialists from around the world working at the International Technical Support Organization, Austin Center.

Giuliano Anselmi is a certified pSeries® Presales Technical Support Specialist working in the Field Technical Sales Support group based in Rome, Italy. For seven years, he was an IBM eServer™ pSeries Systems Product Engineer, supporting Web Server Sales Organization in EMEA, IBM Sales, IBM Business Partners, Technical Support Organizations, and IBM Dublin eServer Manufacturing. Giuliano has worked for IBM for 14 years, devoting himself to RS/6000® and pSeries systems with his in-depth knowledge of the related hardware and solutions.

Gregor Linzmeier is an IBM Advisory IT Specialist for IBM System p workstation and entry servers as part of the Systems and Technology Group in Mainz, Germany supporting IBM sales, Business Partners, and clients with pre-sales consultation and implementation of client/server environments. He has worked for more than 15 years as an infrastructure specialist for RT, RS/6000, IBM IntelliStation® POWER™, and AIX® in large CATIA client/server projects.

Wolfgang Seiwald is an IBM Presales Technical Support Specialist working for the System Sales Organization in Salzburg, Austria. He holds a Diplomingenieur degree in Telematik from the Technical University of Graz. The main focus of his work for IBM in the past nine years has been in the areas of the IBM System p and the IBM AIX 5L™ operating system.

Philippe Vandamme is an IT Specialist working in pSeries Field Technical Support in Paris, France, EMEA West region. With 19 years of experience in semi-conductor fabrication and manufacturing and associated technologies, he is now in charge of IBM System p Pre-Sales Support. In his daily role, he supports and delivers training to the IBM and Business Partner Sales force.

The project that produced this Redpaper was managed by:

Scott Vetter
IBM U.S.

Thanks to the following people for their contributions to this project:

Jane Arbeitman, Ron Arroyo, John Banchy, Barb Hewitt, Jeanine Indest, Tenley Jackson, Andy McLaughlin, Thoi Nguyen, Jan Palmer, Charlie Reeves, Craig Shempert, Scott Smylie, Doug Szerdi, Joel Tendler, Ed Toutant, Bob Vidrick
IBM U.S.

Derrick Daines, Dave Williams
IBM UK

Volker Haug
IBM Germany

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners, or clients.

Your efforts will help increase product acceptance and client satisfaction. As a bonus, you'll develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

`ibm.com/redbooks/residencies.html`

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this Redpaper or other Redbooks™ in one of the following ways:

- Use the online **Contact us** review redbook form found at:

  `ibm.com/redbooks`

- Send your comments in an e-mail to:

  `redbook@us.ibm.com`

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

1

General description

The IBM System p5 570 rack-mount server is designed for greater application flexibility, with innovative technology, to capitalize on the e-business revolution at the midrange level for server environments. With POWER5+™ microprocessor technology, the p5-570 is a cost-effective, high-performance midrange UNIX server that includes Micro-Partitioning™ technology.

Dynamic logical partitioning is supported from the 2-core p5-570 to the 16-core p5-570 system, allowing up to 16 dedicated partitions. In addition, the optional Advanced POWER Virtualization hardware feature is designed to enable the p5-570 system to support up to 160 partitions on a 16-core system. The Micro-Partitioning technology is an advanced feature of the POWER5+ processor that enables multiple partitions to share a physical processor. The extended POWER Hypervisor™ controls dispatching the physical processors to each of the partitions using Micro-Partitioning technology. In addition to Micro-Partitioning technology, the Advanced POWER Virtualization feature enables sharing of network and storage adapters to satisfy the I/O requests of partitions that do not have a dedicated physical I/O adapter.

In combination with the extraordinary POWER5+ processor, the Micro-Partitioning technology is designed to increase system management efficiency and lowers operating expenses through the multiple use of single physical resources that are installed in the p5-570 system.

Simultaneous multi-threading, a standard feature of POWER5+ technology, enables two threads to be executed at the same time on a single processor. Simultaneous multi-threading is user-selectable with dedicated or processors from a shared pool for use by partitions using Micro-Partitioning technology.

The symmetric multiprocessor (SMP) p5-570 system features base 2-core, 4-core, 8-core, 12-core, and 16-core, 64-bit, copper-based, SOI-based POWER5+ microprocessors running at 1.9 GHz and 2.2 GHz with 36 MB off-chip Level 3 cache and DDR2 memory configurations. The system is based on a concept of system building blocks. The p5-570 building blocks are facilitated by the use of Processor interconnect and system SP Flex cables that enable as many as four 4-core p5-570 building blocks to be connected to achieve a true 16-core SMP combined system. Additional processor configurations are possible with the installation of Capacity on Demand (CoD) features. Main memory starting at 2 GB can be expanded to 128 GB in a single drawer (512 GB per 2.2 GHz system), based on the available

DIMMs, for higher performance and exploitation of 64-bit addressing, to meet the demands of enterprise computing, such as large database applications.

One p5-570 building block includes six hot-plug PCI-X[1] slots with Enhanced Error Handling (EEH) and an enhanced blind-swap mechanism, two Ultra320 SCSI controllers, one 10/100/1000 Mbps integrated dual-port Ethernet controller, two system ports, two USB 2.0 ports, two HMC ports, two remote RIO-2 ports, and two System Power Control Network (SPCN) ports.

The p5-570 includes two 3-pack front-accessible, hot-swap-capable disk bays. The six disk bays of one IBM System p5 570 building block can accommodate up to 1.8 TB of disk storage using the 300 GB Ultra320 SCSI disk drives. Two additional media bays are used to accept optional slim-line media devices, such as DVD-ROM or DVD-RAM drives. The p5-570 also has I/O expansion capability using the RIO-2 bus, which allows attachment of the 7311 Model D11 and 7311 Model D20 I/O drawers.

Additional reliability and availability features include redundant hot-plug cooling fans and redundant power supplies. Along with these hot-plug components, the p5-570 is designed to provide an extensive set of reliability, availability, and serviceability (RAS) features that include improved fault isolation, recovery from errors without stopping the system, avoidance of recurring failures, and predictive failure analysis.

[1] PCI stands for Perhiperal Component Interconnect, and the X stands for extended performances.

1.1 System specifications

Table 1-1 lists the general system specifications of a single p5-570 drawer.

Table 1-1 p5-570 specifications

Description	Range
Operating temperature	5 to 35 degrees C (41 to 95 F)
Relative humidity	8% to 80%
Maximum wet bulb	23 degrees C (73 F) (operating)
Noise level	6.2 to 7.1 bels (operating 4-core configurations)
Operating voltage	200 to 240 V AC 50/60 Hz
Maximum power consumption	1,300 watts (maximum)
Maximum power source loading	1.37 kVA (maximum configuration)
Maximum thermal output	4,437 BTU[a]/hr (maximum configuration)

a. British Termal Unit (BTU)

1.2 Physical package

One p5-570 drawer is packaged in a 4U[2] rack-mounted enclosure, and it is available only in the rack-mounted form factor. The following sections discuss the major physical attributes that are found on the p5-570 building block, as shown in Table 1-2.

Table 1-2 Physical packaging of the p5-570

Dimension	One p5-570 building block
Height	174.1 mm (6.85 in.)
Width	483 mm (19.0 in.)
Depth	790 mm (31.1 in.)
Weight	63.6 kg (140 lb.)

Using the p5-570 building block, an installed system can be made of one to four building blocks. To help ensure the installation and serviceability in non-IBM, industry-standard racks, review the vendor's installation planning information for any product-specific installation requirements. The processor and SP Flex cables present an additional planning requirement.

[2] One Electronic Industries Association Unit (1U) is 44.45 mm (1.75 in.).

Figure 1-1 shows some views of the p5-570.

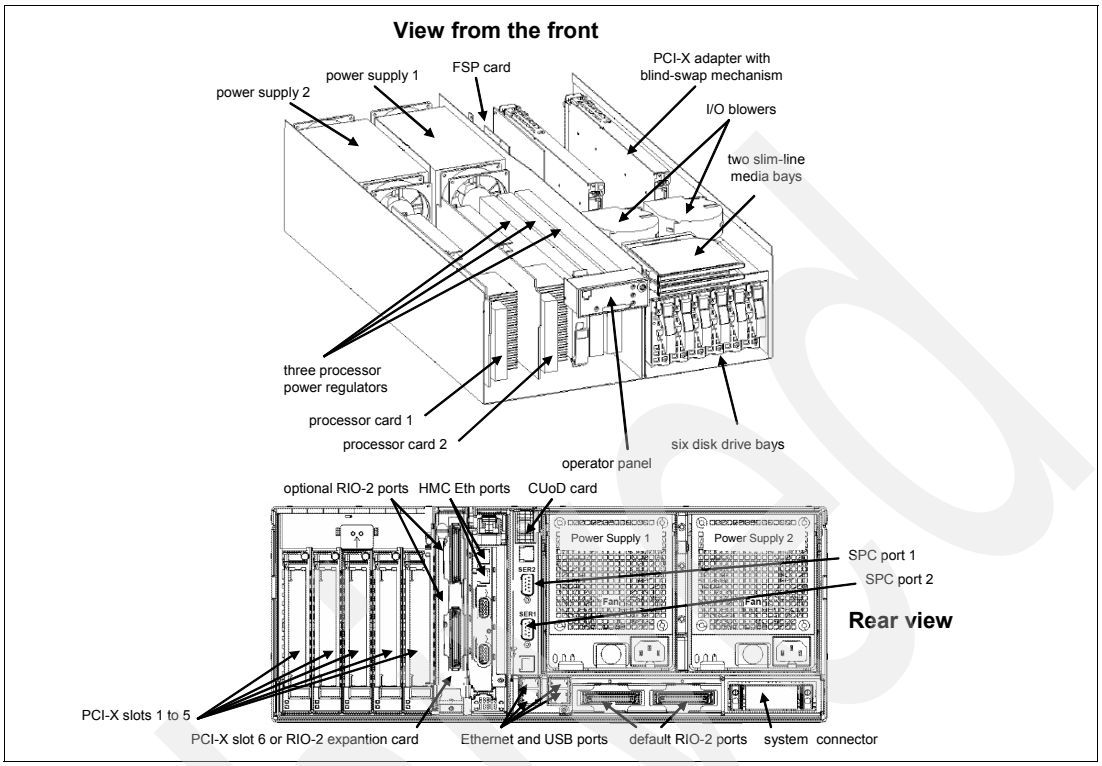


Figure 1-1 Views of the p5-570

1.3 Minimum and optional features

The p5-570 full configuration system is made of four p5-570 building blocks. It features:

- Up to eight processor books using the POWER5+ chip, for a total of 16 processors
- From 2 GB to 256 GB of total system memory capacity using 533 MHz DDR2 DIMM technology or 32 GB to 512 GB using 400 MHz DDR2 DIMM technology in a four-drawer system
- 24 SCSI disk drives for an internal storage capacity of 7.2 TB using 300 GB drives
- 24 PCI-X slots
- Eight slim-line media bays for optional optical storage devices

The combined system (made of more than one p5-570 building block) requires the proper Processor interconnect cable and the system SP Flex cable. (See 2.2.1, “Processor drawer interconnect cables” on page 22, and 2.4.2, “SP bus” on page 26.)

The p5-570 building block includes the service processor (SP), which is described in 2.15.4, “Service processor” on page 62, and the following native ports:

- Two 10/100/1000 Ethernet ports
- Two system ports
- Two USB 2.0 ports (external USB diskette drive 1.44 (FC 2591) can be used.)
- Two HMC ports
- Two remote I/O (RIO-2) ports

- Two SPCN ports

In addition, the p5-570 building block features two internal Ultra320 SCSI controllers, redundant hot-swap power supply and redundant hot-swap cooling fans, and redundant processor power regulators (FC 7768).

There is a CUoD card that is part of the hardware configuration. This card stores VPD and the processor information required for the management of CUoD features. Since the p5-570 can have processors in up to four physical building blocks, the card can be replaced or updated by the IBM service representative to reflect hardware configuration changes.

Note: In a p5-570 combined system made of more than one building block, only the two HMC ports and the two system ports in the building block with the service processor are available for use.

The system supports 32-bit and 64-bit applications, and it requires specific levels of the operating system. (See 2.14, “Operating system support” on page 54.)

1.3.1 Processor card features

Each p5-570 building block can contain 2-core processor cards with state-of-the-art, 64-bit, copper-based, POWER5+ microprocessors running at 1.9 GHz or 2.2 GHz. All card features are available only as Capacity Upgrade on Demand (CUoD). The initial order of the p5-570 system must contain the feature code related to the desired processor card, plus it must contain the processor activation feature code. The feature can be:

- *CUoD* (shipped from manufacturing as activated, or for later activation of available non-activated processors)
- *Reserve CoD* activation of prepaid processor
- *On/Off CoD* activation to use the On/Off CoD capabilities

Table 1-3 and Table 1-4 contain the feature codes for processor cards at the time of writing.

Table 1-3 Processor card feature codes

Processor card FC	Description
7782	Two processors, 0 activated, 1.9 GHz, eight DDR2 DIMM sockets
8338	Two processors, 0 activated, 2.2 GHz, eight DDR2 DIMM sockets

Table 1-4 Processor activation feature codes

Processor card FC	Description	
One processor activation	**For FC 7782**	**For FC 8338**
CUoD (permanent)	FC 7665	FC 7618
Reserve CoD	FC 7666 (30 days prepaid)	FC 7738 (30 days prepaid)
On/Off CoD (1-day billing)	FC 7718	FC 7624

Each processor card features one POWER5+ chip, with two processor cores that share 1.9 MB of L2 cache, 36 MB of L3 cache, eight slots for memory DIMMs using DDR2 technology, and requires a minimum of 2 GB memory. (See “1.3.2, “Memory features” on page 6.”)

1.3.2 Memory features

The processor cards that are used in the p5-570 system offer eight sockets for memory DIMMs. The total memory capacity requires four p5-570 building blocks and eight processor cards. Table 1-5 shows the memory feature codes that are available at the time of writing. The p5-570 system supports CUoD options for memory.

Memory FC 4498 and FC 4499 allows you to balance greater memory capacity for memory throughput.

Table 1-5 Memory feature codes

Feature code	Description
7892	2048 MB (4 x 512 MB) DIMMs, 276-pin, 533 MHz DDR2 SDRAM
7893	4096 MB (4 x1024 MB) DIMMs, 276-pin, 533 MHz DDR2 SDRAM
7894	8192 MB (4 x 2048 MB) DIMMs, 276-pin, 533 MHz DDR2 SDRAM
4497	16 GB (4 x 4096 MB) DIMMs, 276-pin, 533 MHz DDR2 SDRAM
4498	32 GB (4 x 8192 MB) DIMMs, 276-pin, 400 MHz DDR2 SDRAM
4499	32 GB (4 x 8192 MB) DIMMs, 276-pin, 400 MHz DDR2 SDRAM
7663	1 GB memory activation
4495	4/8 GB (4 x 2048 MB) CUoD DIMMs, 276-pin, 533 MHz DDR2 SDRAM
4496	8/16 GB (4 x 4096 MB) CUoD DIMMs, 276-pin, 533 MHz DDR2 SDRAM

We recommend that each processor card have an equal amount of memory installed. Balancing memory across the installed processor cards enables memory accesses to be distributed evenly over system components to provide optimal performance. Memory speed mixing is not supported.

FC 4498 and FC 4499 can only be used with processor FC 8338, the 2.2 GHz processor feature, and may be intermixed with each other on the same processor card. 400 MHz and 533 MHz memory cannot be mixed on the same processor card. Another processor card in the same CEC can have 533 MHz memory installed, however the entire system will then operate at the slower memory speed if FC 4498 or FC 4499 exist in the CEC.

1.3.3 Disk and media features

Each p5-570 building block features six disk drive bays and two slim-line media device bays. In a full configuration with four connected p5-570 building blocks, the combined system supports up to 24 disk bays; therefore, the maximum internal storage capacity is 7.2 TB (using the disk drive features available at the time of writing). The minimum configuration requires at least one 36.4 GB disk drive. Table 1-6 shows the disk drive feature codes that each bay can contain.

Table 1-6 Disk drive feature code description

Feature code	Description
3277	36.4 GB 15K RPM Ultra320 SCSI disk drive assembly
3274	73.4 GB 10K RPM Ultra320 SCSI disk drive assembly
3278	73.4 GB 15K RPM Ultra320 SCSI disk drive assembly

Feature code	Description
3275	146.8 GB 10K RPM Ultra320 SCSI disk drive assembly
3279	146.8 GB 15K RPM Ultra320 SCSI disk drive assembly
3578	300 GB 10K RPM Ultra320 SCSI disk drive assembly

In a full configuration, with four connected p5-570 building blocks, the combined system supports up to eight slim-line media device bays. To support two slim-line devices in each p5-570 building block, the optional media enclosure and backplane (FC 7869) is required.

Any combination of the following DVD-ROM and DVD-RAM drives can be installed:

- DVD-RAM drive, FC 5751
- DVD-ROM drive, FC 2640

A logical partition running a supported release of the Linux® operating system requires the media enclosure and backplane feature code, and a DVD-ROM drive or DVD-RAM drive.

1.3.4 USB diskette drive

The external USB 1.44 MB diskette drive for p5-570 systems (FC 2591) is available. This super-slim-line and lightweight USB attached diskette drive takes its power requirements from the USB port. A USB cable is provided. The drive can be attached to the integrated USB ports or to a USB adapter (FC 2738). A maximum of one USB diskette drive is supported per integrated controller/adapter. The same controller can share a USB mouse and keyboard.

1.3.5 I/O drawers

The p5-570 has six internal blind swap PCI-X slots: five are long slots and one is a short slot. The short PCI-X slot may also be used for the Remote I/O expansion card (FC 1800). If more PCI-X slots are needed, such as to extend the number of LPARs and partitions using Micro-Partitioning technology, up to 20 7311 Model D11 or 7311 Model D20 I/O drawers can be attached.

7311 Model D11 I/O drawer

The 7311 Model D11 I/O drawer features six long PCI-X slots. Only the blind-swap cassettes are supported:

- FC 7862, for full-sized PCI cards
- FC 7861, for short-sized PCI cards
- FC 7863, for double-wide PCI cards

Two 7311 Model D11 I/O drawers fit side-by-side in the 4U enclosure (FC 7311) mounted in a 19-inch rack, such as the IBM 7014-T00 or 7014-T42.

The 7311 Model D11 I/O drawer offers a modular growth path for the p5-570 systems with increasing I/O requirements. A fully configured p5-570 supports 20 attached 7311 Model D11 I/O drawers. The combined system supports up to 144 PCI-X adapters. (In a full configuration, Remote I/O expansion cards are required.)

The I/O drawer has the following attributes:

- 4U rack-mount enclosure (FC 7311) that can hold one or two D11 drawers
- Six PCI-X slots: 3.3 V, keyed, 133 MHz blind-swap hot-plug

- Default redundant hot-plug power and cooling devices
- Two RIO-2 and two SPCN ports

7311 Model D11 I/O drawer physical package

Because the 7311 Model D11 I/O drawer must be mounted into the rack enclosure (FC 7311), these are the physical characteristics of one I/O drawer or two I/O drawers side-by-side:

- One 7311 Model D11 I/O drawer
  - Width: 223 mm (8.8 in.)
  - Depth: 711 mm (28.0 in.)
  - Height: 175 mm (6.9 in.)
  - Weight: 19.6 kg (43 lb.)
- Two I/O drawers in a 7311 rack-mounted enclosure have the following characteristics:
  - Width: 445 mm (17.5 in.)
  - Depth: 711 mm (28.0 in.)
  - Height: 175 mm (6.9 in.)
  - Weight: 39.1 kg (86 lb.)

7311 Model D20 I/O drawer

The 7311 Model D20 I/O drawer is a 4U full-size drawer, which must be mounted in a rack. It features seven hot-pluggable PCI-X slots and, optionally up to 12 hot-swappable disks arranged in two 6-packs. Redundant concurrently maintainable power and cooling is an optional feature (FC 6268). The 7311 Model D20 I/O drawer offers a modular growth path for the p5-570 systems with increasing I/O requirements. When a p5-570 is fully configured with 20 attached 7311 Model D20 drawers, the combined system supports up to 164 PCI-X adapters (in full configuration, a Remote I/O expansion card must be present, and 264 hot-swappable disks, for a total internal storage capacity of 38.7 TB using the 146.8 GB drive.

PCI-X and PCI cards are inserted into the slot from the top of the I/O drawer. The installed adapters are protected by plastic separators, which are designed to prevent grounding and damage when adding or removing adapters.

The drawer has the following attributes:

- 4U rack mount enclosure assembly
- Seven PCI-X slots: 3.3 V, keyed, 133 MHz hot-plug
- Two 6-pack hot-swappable SCSI devices
- Optional redundant hot-plug power
- Two RIO-2 and two SPCN ports

Note: The 7311 Model D20 I/O drawer initial order, or an existing 7311 Model D20 I/O drawer that is migrated from another pSeries system, must have the RIO-2 ports available (FC 6417).

7311 Model D20 I/O drawer physical package

The I/O drawer has the following physical characteristics:

- Width: 482 mm (19.0 in.)
- Depth: 610 mm (24.0 in.)

- Height: 178 mm (7.0 in.)
- Weight: 45.9 kg (101 lb.)

Figure 1-2 shows the different views of the 7311-D20 I/O drawer.

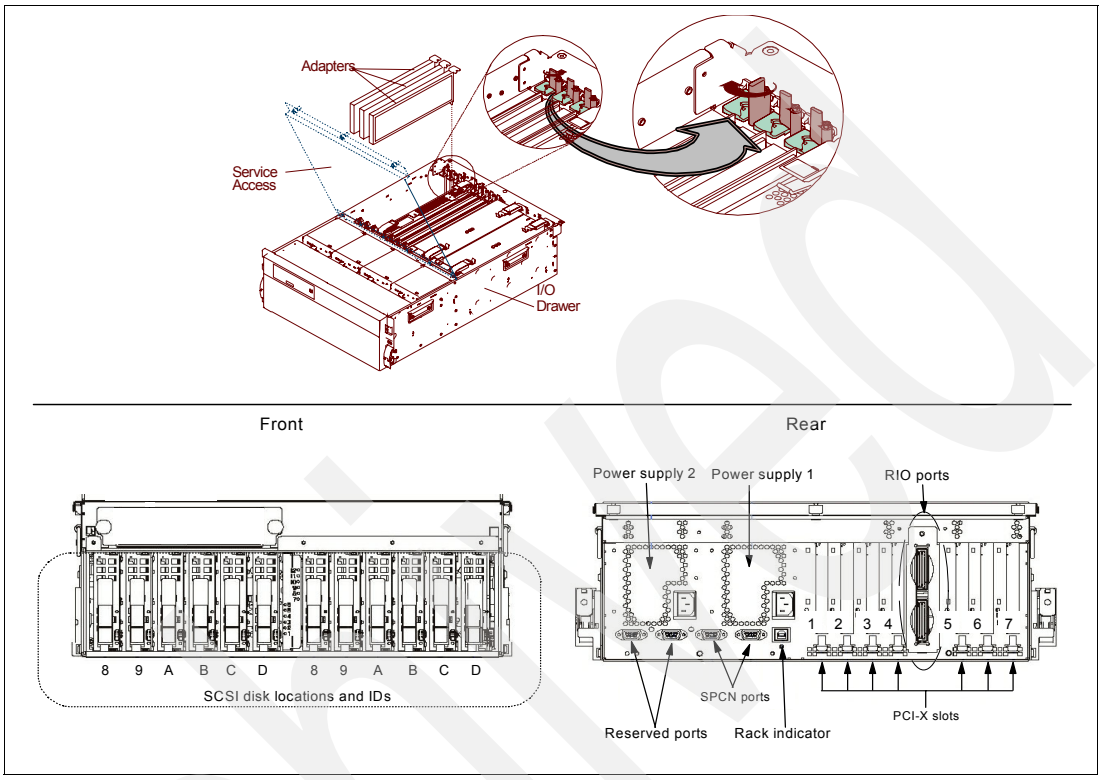


Figure 1-2 7311-D20 I/O drawer views

Note: The 7311 Model D10, and the 7311 Model D11, or the 7311 Model D20 I/O drawers are designed to be installed by an IBM service representative.

I/O drawers and usable PCI slots

The different I/O drawer model types can be intermixed on a single p5-570 server and within the same RIO-2 loop. Depending on the p5-570 system configuration, the maximum number of I/O drawers supported is different. Table 1-7 summarizes the maximum number of I/O drawers supported and the total number of PCI-X slots available when expansion consists of a single drawer type.

Table 1-7 Maximum number of I/O drawers supported and total number of PCI slots

p5-570 drawer/processors	Max number of I/O drawers	Total number of PCI-X slots	
		D11	D20
1 drawer / 2-core	4	30	34
1 drawer / 4-core	8	54	62
2 drawers / 8-core	12	84	96
3 drawers / 12-core	16	114	130

p5-570 drawer/processors	Max number of I/O drawers	Total number of PCI-X slots	
		D11	D20
4 drawers / 16-core	20	144[a]	164

[a] One slot is reserved for the Remote I/O expansion card.

1.3.6 Hardware Management Console models

The Hardware Management Console (HMC) provides a set of functions that are necessary to manage the p5-570 system when LPAR, Capacity on Demand without reboot, inventory and microcode management, and remote power control functions are needed. These functions include the handling of the partition profiles that define the processor, memory, and I/O resources that are allocated to an individual partition.

Table 1-8 lists the last model available at the time of writing for the desktop and rack-mounted HMC options for POWER5+ processor-based systems. Existing HMC models can be also used when running the correct level of software.

Table 1-8 Last HMC models available

Type-model	Description
7310-C04	IBM 7310 Model C04 desktop Hardware Management Console
7310-C05	IBM 7310 Model C05 deskside Hardware Management Console
7310-CR3	IBM 7310 Model CR3 rack-mount Hardware Management Console

Systems require Ethernet connectivity between the HMC and one of the Ethernet ports of a service processor. Ensure that sufficient Ethernet adapters are available to enable public and private networks if you need both.

The 7310 Model C04 is a desktop model and the 7310 Model C05 is a deskside model with only one native 10/100/1000 Ethernet port, but can be extended with other two additional dual-port 10/100/1000 Gb Ethernet adapters.

The 7310 Model CR3 is a 1U, 19-inch rack mountable drawer that has two native Ethernet ports and can be extended with one additional two-port 10/100/1000 Gb Ethernet adapter.

In HMC managed installations with very high demand for high availability, two HMCs are recommended. The p5-570 service processor allows the connection of two HMCs, so there are no additional features needed for a p5-570 to support a dual HMC environment. The HMCs provide a locking mechanism so that only one HMC at a time has write access to the service processor.

When an HMC is connected to the p5-570, the integrated system ports are disabled. If you need serial connections, for example, non-Ethernet HACMP™ heartbeat, you need to provide an additional asynchronous adapter (see 2.6.8, “Asynchronous PCI-X adapters” on page 32).

Note: It is not possible to connect POWER4™ and POWER5™ or POWER5+ processor-based systems simultaneously to the same HMC, but it is possible to connect POWER5 and POWER5+ processor-based systems together to the same HMC.

1.4 System racks

The IBM 7014 Model T00 and T42 are 19-inch racks for general use with IBM System p5, IBM eServer p5, pSeries, and OpenPower™ rack-mount servers. The racks provide increased capacity, greater flexibility, and improved floor space utilization.

If a System p5 server is to be installed in a non-IBM rack or cabinet, you must ensure that the rack conforms to the EIA[3] standard EIA-310-D (see 1.4.6, “OEM rack” on page 16).

Note: It is the client’s responsibility to ensure that the installation of the drawer in the preferred rack or cabinet results in a configuration that is stable, serviceable, safe, and compatible with the drawer requirements for power, cooling, cable management, weight, and rail security.

1.4.1 IBM 7014 Model T00 rack

The 1.8-meter (71-in.) Model T00 is compatible with past and present IBM System p systems. The T00 rack has the following features:

- 36 EIA units (36U) of usable space.
- Optional removable side panels.
- Optional highly perforated front door.
- Optional side-to-side mounting hardware for joining multiple racks.
- Standard business black or optional white color in OEM format.
- Increased power distribution and weight capacity.
- Optional reinforced (ruggedized) rack feature (FC 6080) provides added earthquake protection with modular rear brace, concrete floor bolt-down hardware, and bolt-in steel front filler panels.
- Support for both ac and dc configurations.
- The dc rack height is increased to 1926 mm (75.8 in.) if a power distribution panel is fixed to the top of the rack.
- Up to four power distribution units (PDUs) can be mounted in the PDU bays (see Figure 1-3 on page 13), but others can fit inside the rack. See 1.4.3, “The ac power distribution unit and rack content” on page 12.
- An optional rack status beacon (FC 4690). This beacon is designed to be placed on top of a rack and cabled to servers, such as a p5-570 and other components inside the rack. Servers can be programmed to illuminate the beacon in response to a detected problem or changes in the system status.
- A rack status beacon junction box (FC 4693) should be used to connect multiple servers to the beacon. This feature provides six input connectors and one output connector for the rack. To connect the servers or other components to the junction box or the junction box to the rack, status beacon cables (FC 4691) are necessary. Multiple junction boxes can be linked together in a series using daisy chain cables (FC 4692).
- Weights:
  - T00 base empty rack: 244 kg (535 pounds)
  - T00 full rack: 816 kg (1795 pounds)

[3] Electronic Industries Alliance (EIA). Accredited by American National Standards Institute (ANSI), EIA provides a forum for industry to develop standards and publications throughout the electronics and high-tech industries.

1.4.2 IBM 7014 Model T42 rack

The 2.0-meter (79.3-inch) Model T42 addresses the client requirement for a tall enclosure to house the maximum amount of equipment in the smallest possible floor space. The features that differ in the Model T42 rack from the Model T00 include:

- 42 EIA units (42U) of usable space (6U of additional space).
- The Model T42 supports ac only.
- Weights:
  - T42 base empty rack: 261 kg (575 pounds)
  - T42 full rack: 930 kg (2045 pounds)

Optional Rear Door Heat eXchanger (FC 6858)

Improved cooling from the Rear Door Heat eXchanger enables clients to more densely populate individual racks, freeing valuable floor space without the need to purchase additional air conditioning units. The Rear Door Heat eXchanger features:

- Water-cooled heat exchanger door designed to dissipate heat generated from the back of computer systems before it enters the room.
- An easy-to-mount rear door design that attaches to client-supplied water, using industry standard fittings and couplings.
- Up to 15 KW (approximately 50,000 BTUs/hr) of heat removed from air exiting the back of a fully populated rack.
- One year limited warranty.

Physical specifications:

- Approximate height: 1945.5 mm (76.6 in.)
- Approximate width: 635.8 mm (25.03 in.)
- Approximate depth: 141.0 mm (5.55 in.)
- Approximate weight: 31.9 kg (70.0 lb.)

Client responsibilities:

- Secondary water loop (to building chilled water)
- Pump solution (for secondary loop)
- Delivery solution (hoses and piping)
- Connections: Standard 3/4-inch internal threads

1.4.3 The ac power distribution unit and rack content

For rack models T00 and T42, 12-outlet PDUs (FC 9188 and FC 7188) are available.

Four PDUs can be mounted vertically in the T00 and T42 racks. See Figure 1-3 on page 13 for the placement of the four vertically mounted PDUs. In the rear of the rack, two additional PDUs can be installed horizontally in the T00 rack and three in the T42 rack. The four vertical mounting locations will be filled first in the T00 and T42 racks. Mounting PDUs horizontally consumes 1U per PDU and reduces the space available for other racked components. When mounting PDUs horizontally, we recommend that you use fillers in the EIA units occupied by these PDUs to facilitate proper air-flow and ventilation in the rack.

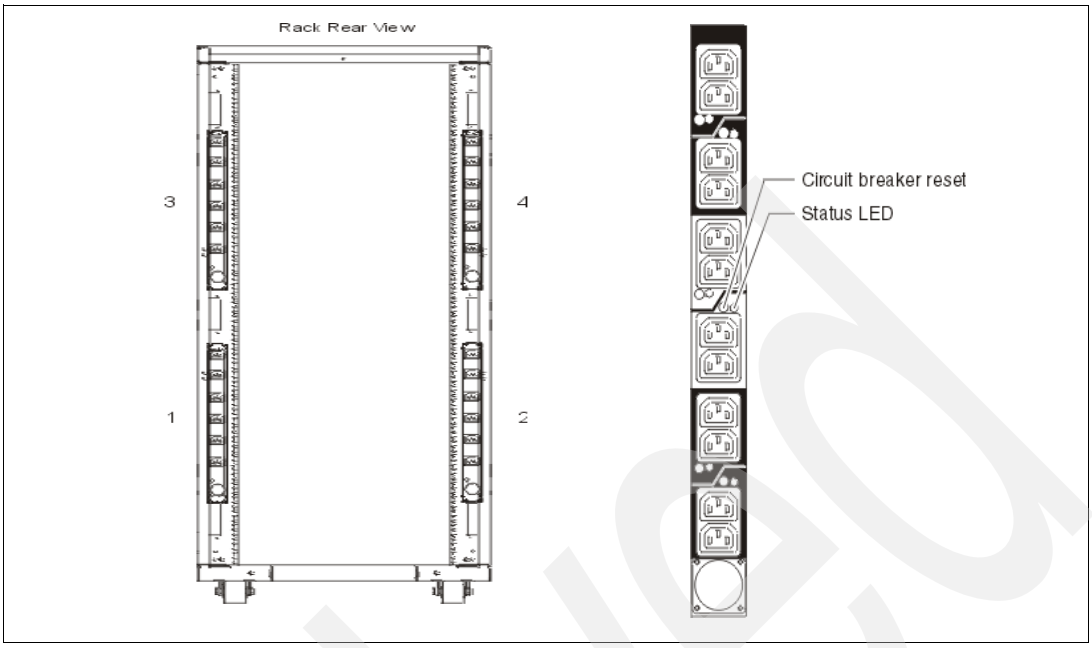


Figure 1-3 PDU placement and PDU view

For detailed power cord requirements and power cord feature codes, see *IBM System p5, eServer p5 and i5, and OpenPower Planning*, SA38-0508. For an online copy, see the IBM Systems Hardware Information Center. You can find it at:

`http://publib.boulder.ibm.com/eserver/`

Note: Ensure that the appropriate power cord feature is configured to support the power being supplied.

The Base/Side Mount Universal PDU (FC 9188) and the optional, additional, Universal PDU (FC 7188) support a wide range of country requirements and electrical power specifications. The PDU receives power through a UTG0247 power line connector. Each PDU requires one PDU-to-wall power cord. Nine power cord features are available for different countries and applications by varying the PDU-to-wall power cord, which must be ordered separately. Each power cord provides the unique design characteristics for the specific power requirements. To match new power requirements and save previous investments, these power cords can be requested with an initial order of the rack or with a later upgrade of the rack features.

The PDU has 12 client-usable IEC 320-C13 outlets. There are six groups of two outlets fed by six circuit breakers. Each outlet is rated up to 10 amps, but each group of two outlets is fed from one 15 amp circuit breaker.

Note: Based on the power cord that is used, the PDU can supply from 4.8 kVA to 19.2 kVA. The total kilovolt ampere (kVA) of all the drawers plugged into the PDU must not exceed the power cord limitation.

The Universal PDUs are compatible with previous models.

Note: Each p5-570 or a system drawer to be mounted in the rack requires two power cords, which are not included in the base order. For maximum availability, we highly recommend connecting power cords from the same p5-570 or system drawer to two separate PDUs. These PDUs being connected to are two independent client power sources.

1.4.4 Rack-mounting rules for p5-570

The primary rules that should be followed when mounting the p5-570 into a rack are:

- The p5-570 is designed to be placed at any location in the rack. For rack stability, it is advisable to start filling a rack from the bottom.

  For p5-570 configurations with multiple building blocks, all drawers must be installed together in the same rack, in a contiguous space within the rack.
- Any remaining space in the rack can be used to install other systems or peripherals, provided that the maximum permissible weight of the rack is not exceeded and the installation rules for these devices are followed.
- Before placing a p5-570Q into the service position, it is essential that the rack manufacturer's safety instructions have been followed regarding rack stability.
- Special consideration must be taken to avoid a flange on the top of the rack to clear the front bezel.
- When a p5-570 system is installed in an IBM 7014-T00 or 7014-T42 rack that has no front door, a Thin Profile Front Trim Kit must be ordered for the rack. The front bezel on a model 570 drawer is too wide to be used with the previously announced rack trim kits (FC 6107 and FC 6081). The required trim kit for the 7014-T00 rack is FC 6246. The required trim kit for the 7014-T42 rack is FC 6247.
- The IBM 7014-T42 rack is constructed with a small flange at the bottom of EIA location 37. This requires special placement rules when a model 570 system is installed near the top of a 7014-T42 rack to avoid interference with the front bezel. No system drawer can be installed in EIA positions 34, 35, or 36. A two-drawer system cannot be installed above position 29. (The position number refers to the bottom of the lowest drawer.)

1.4.5 Useful rack additions

This section highlights some solutions available to provide a single point of management for environments composed of multiple System p5-570 servers or other IBM System p systems.

IBM 7212 Model 102 IBM TotalStorage storage device enclosure

The IBM 7212 Model 102 is designed to provide efficient and convenient storage expansion capabilities for selected System p servers. The IBM 7212 Model 102 is a 1U rack-mountable option to be installed in a standard 19-inch rack using an optional rack-mount hardware feature kit. The 7212 Model 102 has two bays that can accommodate any of the following storage drive features:

- A Digital Data Storage (DDS) Gen 5 DAT72 Tape Drive provides a physical storage capacity of 36 GB (72 GB with 2:1 compression) per data cartridge.
- A VXA-2 Tape Drive provides a media capacity of up to 80 GB (160 GB with 2:1 compression) physical data storage capacity per cartridge.
- A Digital Data Storage (DDS-4) tape drive provides 20 GB native data capacity per tape cartridge and a native physical data transfer rate of up to 3 MBps that uses a 2:1 compression so that a single tape cartridge can store up to 40 GB of data.

- A DVD-ROM drive is a 5 1/4-inch, half-high device. It can read 640 MB CD-ROM and 4.7 GB DVD-RAM media. It can be used for alternate IPL[4] (IBM-distributed CD-ROM media only) and program distribution.
- A DVD-RAM drive with up to 2.7 MBps throughput. Using 3:1 compression, a single disk can store up to 28 GB of data. Supported DVD disk native capacities on a single DVD-RAM disk are as follows: up to 2.6 GB, 4.7 GB, 5.2 GB, and 9.4 GB.

Flat panel display options

The IBM 7316-TF3 Flat Panel Console Kit can be installed in the system rack. This 1U console uses a 15-inch thin film transistor (TFT) LCD with a viewable area of 304.1 mm x 228.1 mm and a 1024 x 768 pels[5] resolution. The 7316-TF3 Flat Panel Console Kit has the following attributes:

- Flat panel color monitor
- Rack tray for keyboard, monitor, and optional VGA switch with mounting brackets
- IBM Travel Keyboard mounts in the rack keyboard tray (Integrated Trackpoint and UltraNav)

IBM PS/2 Travel Keyboards are supported on the 7316-TF3 for use in configurations where only PS/2 keyboard ports are available.

The IBM 7316-TF3 Flat Panel Console Kit provides an option for the USB Travel Keyboards with UltraNav. The keyboard enables the 7316-TF3 to be connected to systems that do not have PS/2 keyboard ports. The USB Travel Keyboard can be directly attached to an available integrated USB port or a supported USB adapter (FC 2738) on System p5 servers or 7310-CR3 and 7315-CR3 HMCs.

The Netbay LCM (Keyboard/Video/Mouse) Switch (FC 4202) provides users single-point access and control of up to 64 servers from a single console. The Netbay LCM Switch has a maximum video resolution of 1600 x 280 and mounts in a 1U drawer behind the 7316-TF3 monitor. A minimum of one LCM feature (FC 4268) or USB feature (FC 4269) is required with a Netbay LCM Switch (FC 4202). Each feature can support up to four systems. When connecting to a p5-570, FC 4269 provides connection to the server USB ports. Only the PS/2 keyboard is supported when attaching the 7316-TF3 to the LCM Switch.

When selecting the LCM Switch, consider the following information:

- The KVM Conversion Option (KCO) cable (FC 4268) is used with systems with PS/2 style keyboard, display, and mouse ports.
- The USB cable (FC 4269) is used with systems with USB keyboard or mouse ports.
- The switch offers four ports for server connections. Each port in the switch can connect a maximum of 16 systems:
  - One KCO cable (FC 4268) or USB cable (FC 4269) is required for every four systems supported on the switch.
  - A maximum of 16 KCO cables or USB cables per port can be used with the Netbay LCM Switch to connect up to 64 servers.

[4] Initial program load
[5] Picture elements

Note: A server microcode update might be required on installed systems for boot-time System Management Services (SMS) menu support of the USB keyboards. The update might also be required for the LCM switch on the 7316-TF3 console (FC 4202). For microcode updates, see the following URL:

`http://techsupport.services.ibm.com/server/mdownload`

We recommend that you have the 7316-TF3 installed between EIA 20 to 25 of the rack for ease of use. The 7316-TF3 or any other graphics monitor requires the POWER GXT135P graphics accelerator (FC 1980) to be installed in the server, or some other graphics accelerator, if supported.

Hardware Management Console 7310 Model CR3

The 7310 Model CR3 Hardware Management Console (HMC) is a 1U, 19-inch rack-mountable drawer supported in the 7014 racks. For additional HMC specifications, see 2.13, “Hardware Management Console” on page 50.

1.4.6 OEM rack

The p5-570 can be installed in a suitable OEM rack, provided that the rack conforms to the EIA-310-D standard for 19-inch racks. This standard is published by the Electrical Industries Alliance, and a summary of this standard is available in the publication *IBM System p5, eServer p5 and i5, and OpenPower Planning*, SA38-0508.

The key points mentioned in this documentation are as follows:

- The front rack opening must be 451 mm wide + 0.75 mm (17.75 in. + 0.03 in.), and the rail-mounting holes must be 465 mm + 0.8 mm (18.3 in. + 0.03 in.) apart on center (horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges). See Figure 1-4 for a top view showing the specification dimensions.

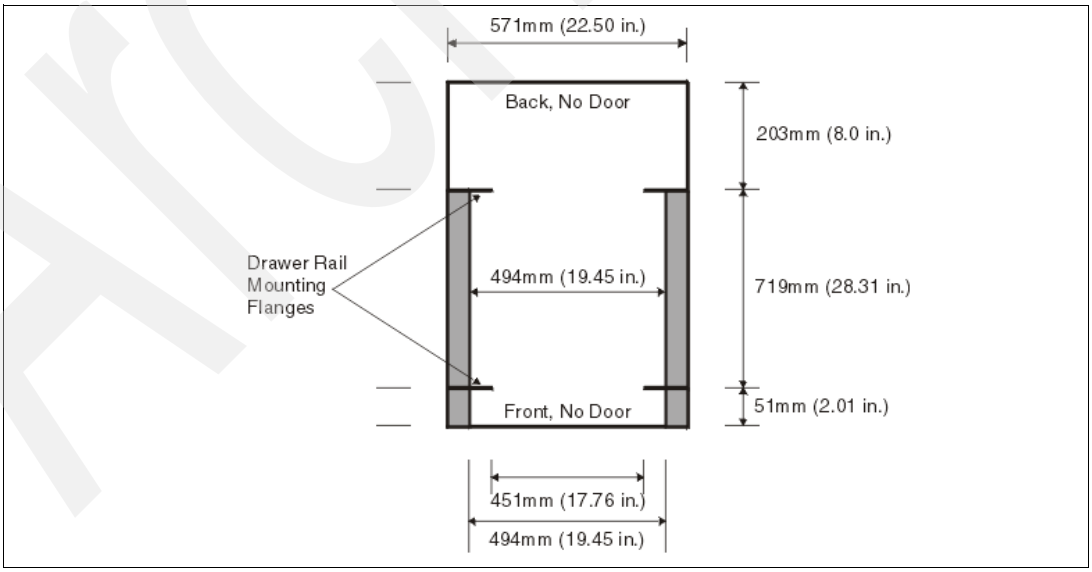


Figure 1-4 Top view of non-IBM rack specification dimensions

- The vertical distance between the mounting holes must consist of sets of three holes spaced (from bottom to top) 15.9 mm (0.625 in.), 15.9 mm (0.625 in.), and 12.67 mm (0.5 in.) on center, making each three-hole set of vertical hole spacing 44.45 mm (1.75 in.)

apart on center. Rail-mounting holes must be 7.1 mm + 0.1 mm (0.28 in. + 0.004 in.) in diameter. See Figure 1-5 and Figure 1-6 for the top and bottom front specification dimensions.

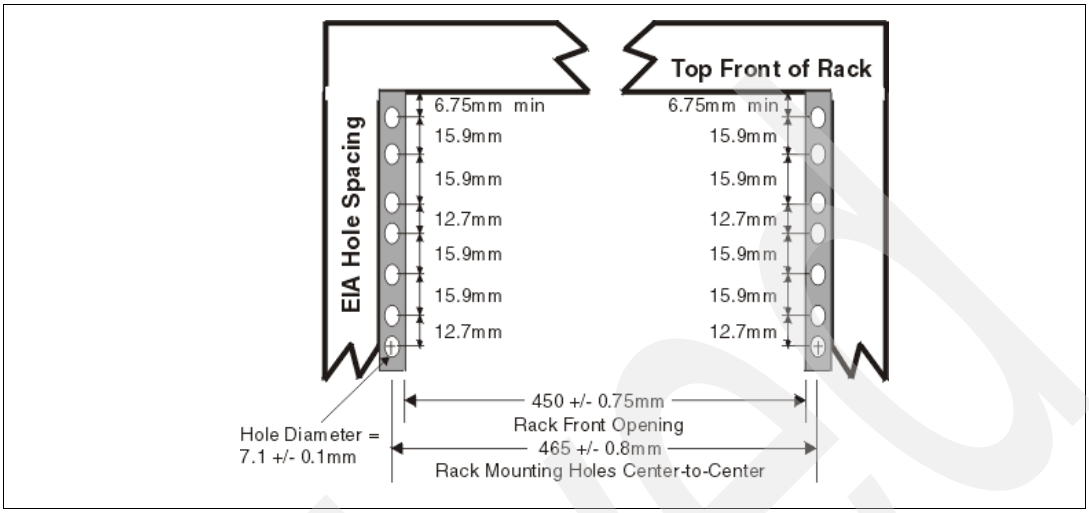


Figure 1-5 Rack specification dimensions, top front view

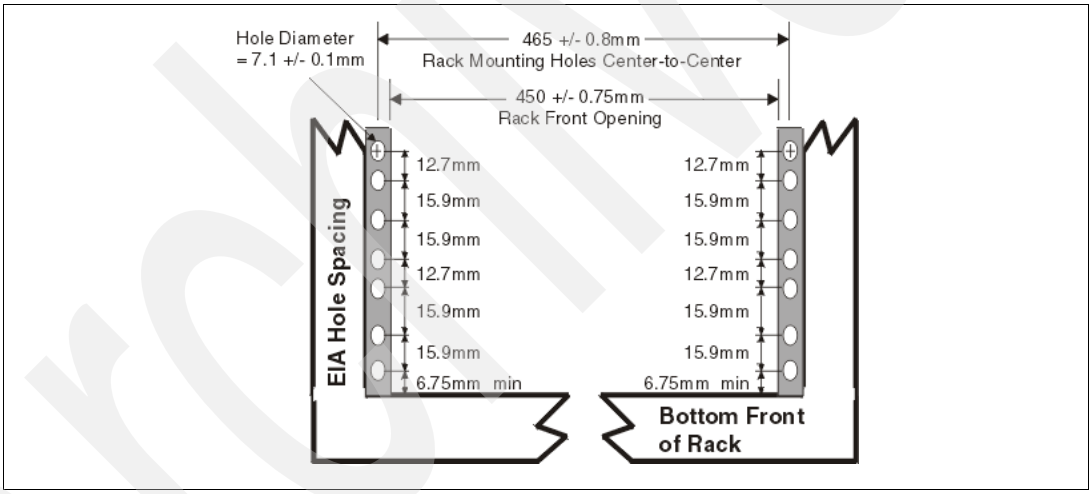


Figure 1-6 Rack specification dimensions, bottom front view

- It might be necessary to supply additional hardware, such as fasteners, for use in some manufacturer's racks.
- The system rack or cabinet must be capable of supporting an average load of 15.9 kg (35 lb.) of product weight per EIA unit.
- The system rack or cabinet must be compatible with drawer mounting rails, including a secure and snug fit of the rail-mounting pins and screws into the rack or cabinet rail support hole.

Note: The OEM rack must only support ac-powered drawers. We strongly recommend that you use a power distribution unit (PDU) that meets the same specifications as the PDUs to supply rack power. Rack or cabinet power distribution devices must meet the drawer power requirements, as well as the requirements of any additional products that will be connected to the same power distribution device.

1.5 Statement of Direction

IBM is committed to enhancing their client's investments in the IBM System p product line. Based on this commitment, IBM plans to provide an upgrade path from the current p5-570 server to the IBM next generation POWER6™ processor-based enterprise servers.

All statements regarding IBMs future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Any reliance on these Statements of Direction is at the relying party's sole risk and will not create liability or obligation for IBM.

2

Architecture and technical overview

This chapter discusses the overall system architecture represented by Figure 2-1, with its major components described in the following sections. The bandwidths that are provided throughout the section are theoretical maximums used for reference. You should always obtain real-world performance measurements using production workloads.

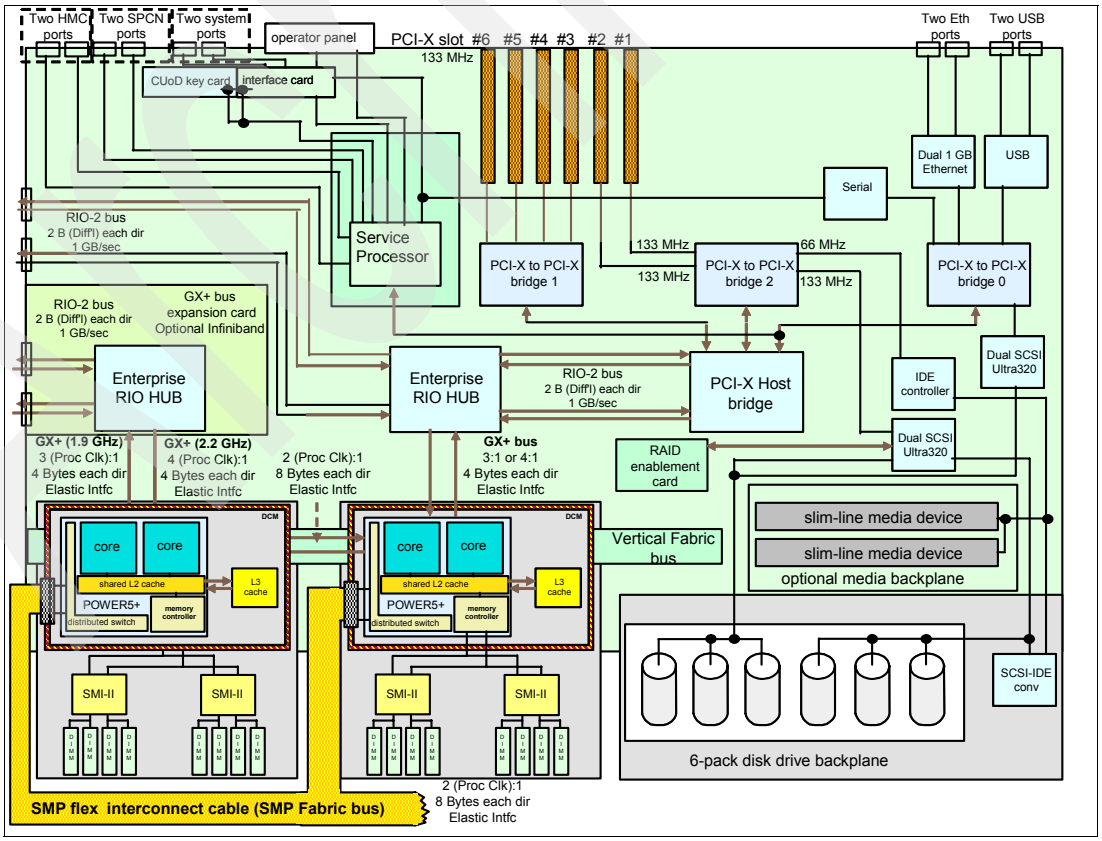


Figure 2-1 p5-570 logic data flow

2.1 The POWER5+ processor

The POWER5+ processor capitalizes on all the enhancements brought by the POWER5 chip. For a detailed description of the POWER5 chip, refer to*IBM @server p5 550 Technical Overview and Introduction*, REDP-9113.

Figure 2-2 shows a high-level view of the POWER5+ processor.

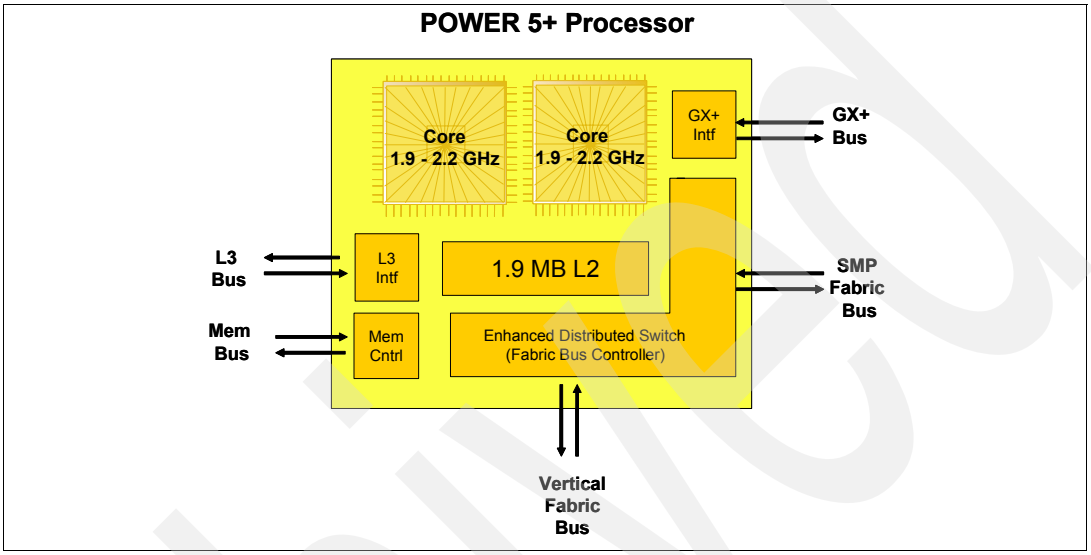


Figure 2-2 POWER5+ processor

The CMOS10S technology in the POWER5+ processor uses a 90 nm fabrication process, which enables:

- Performance gains through faster clock rates
- Chip size reduction (243 mm compared with 389 mm)

The POWER5+ processor is 37% smaller than the POWER5 chip. It consumes less power and requires less sophisticated cooling. Thus, you can use the POWER5+ processor in servers where previously you could only use low frequency chips due to cooling restrictions.

The POWER5+ design provides the following additional enhancements:

- New pages sizes in ERAT and TLB. Two new pages sizes (64 KB and 16 GB) recently added in PowerPC® architecture.
- New segment size in SLB. One new segment size (1 TB) recently added in PowerPC architecture.
- The TLB size has been doubled in the POWER5+ over the POWER5 processor. The TLB in POWER5+ has 2048 entries.
- Floating-point round to integer instructions. New instructions (frfin, frfiz, frfip, frfim) have been added to round floating-point numbers integers with the following rounding modes: nearest, zero, integer plus, integer minus.
- Improved floating-point performance.
- Lock performance enhancement.
- Enhanced SLB read.

- True Little-Endian mode. Support for the True Little-Endian mode as defined in the PowerPC architecture.
- Double the SMP support. Changes have been made in the fabric, L2 and L3 controller, memory controller, GX+ controller, and chip RAS to provide support for the QCM (Quad-Core Module) that allows the SMP system sizes to be double than that is available in POWER5 DCM-based servers. However current POWER5+ implementations support only a single address loop.
- Several enhancements have been made in the memory controller for improved performance.

Enhanced redundancy in L1 Dcache, L2 cache and L3 directory. Independent control of the L2 cache and the L3 directory for redundancy to allow split-repair action has been added. More wordline redundancy has been added in the L1 Dcache. In addition, Array Built-In Self Test (ABIST) column repair for the L2 cache and the L3 directory has been added.

2.2 Processor cards

In the p5-570 system, the POWER5+ processor has been packaged with the L3 cache chip into a cost-effective Dual Chip Module (DCM) package. The storage structure for the POWER5+ processor chip is a distributed memory architecture that provides high-memory bandwidth. Each processor can address all memory and sees a single shared memory resource. As such, a single DCM and its associated L3 cache and memory are packaged on a single processor card. Access to memory behind another processor is accomplished through the fabric buses. The p5-570 supports up to two processor cards (each card is a 2-core) in any building block. Each processor card has a single DCM containing a POWER5+ processor chip and a 36 MB L3 module. I/O connects to the Central Electronic Complex (CEC) subsystem using the GX+ bus. Each DCM provides a single GX+ bus for a total system capability of two GX+ buses. The GX+ bus provides an interface to a single device, such as the RIO-2 buses.

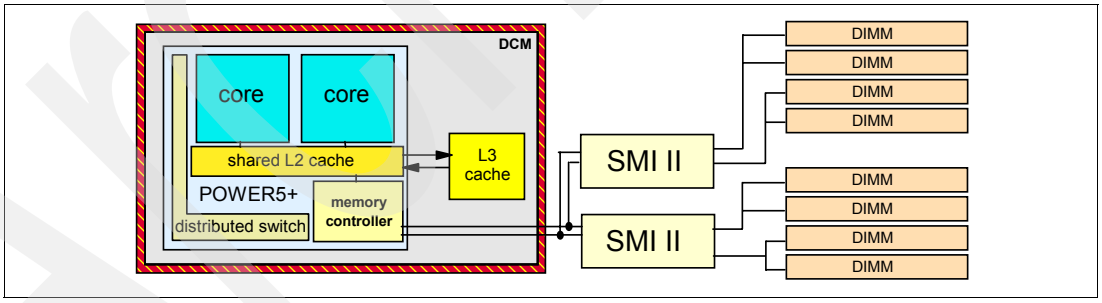


Figure 2-3 p5-570 1.9 GHz DCM diagram

The processor card also contains LEDs for each FRU[1] on the CPU card, including the CPU card itself. Figure 2-4 shows a processor card layout view.

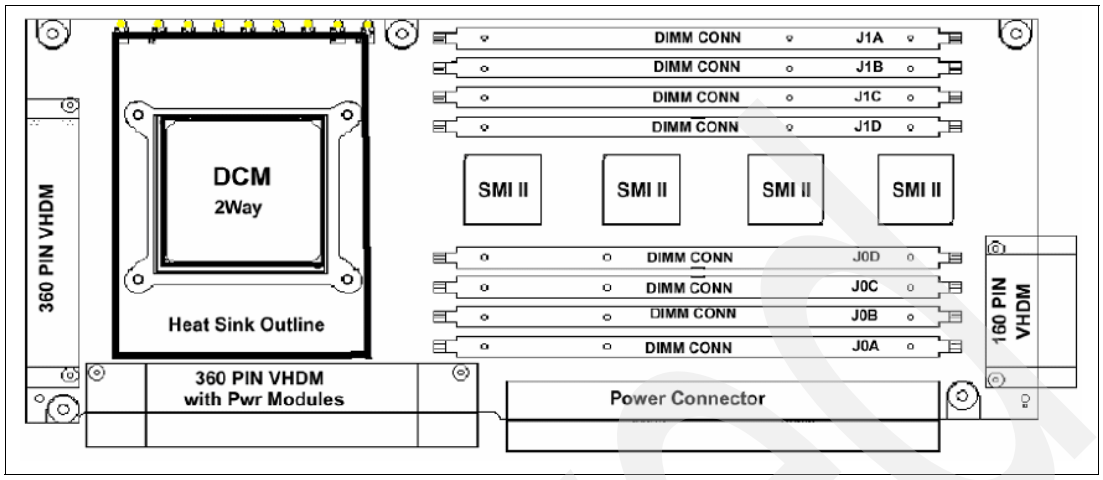


Figure 2-4 2.2 GHz Processor card with four SMI DDR2 memory socket layout

There are two system backplanes in the p5-570 system. A GX+ bus planar, which docks vertically into the system planar, is always present in the system. The processor cards dock directly into this backplane from the front. A horizontal backplane exists below the CPU cards that is co-planar with the I/O backplane. This backplane routes the vertical fabric bus between the processor cards. This backplane is also used for power distribution from the CPU regulators that are housed next to the processor cards. (See Figure 1-1 on page 4.)

2.2.1 Processor drawer interconnect cables

In combined systems that are made of more than one p5-570 building block, the connection between processor cards in different building blocks is provided with a processor drawer interconnect cable. Different processor drawer interconnect cables are required for the different numbers of p5-570 building blocks that a combined system can be made of, as shown in Figure 2-5 on page 23.

Because of the redundancy and fault recovery built-in to the system interconnects, a drawer failure does not represent a system failure. Once a problem is isolated and repaired, a system reboot may be required to reestablish full bus speed, if the failure was specific to the interconnects.

The SMP fabric bus that connects the processors of separate p5-570 building blocks is routed on the interconnect cable that is routed external to the building blocks. The flexible cable attaches directly to the processor cards, at the front of the p5-570 building block, and is routed behind the front covers (bezels) of the p5-570 building blocks. There is an optimized cable for each drawer configuration. Figure 2-5 on page 23 illustrates the logical fabric bus connections between the drawers, and shows the additional space required left of the bezels for rack installation.

[1] Field replacement unit

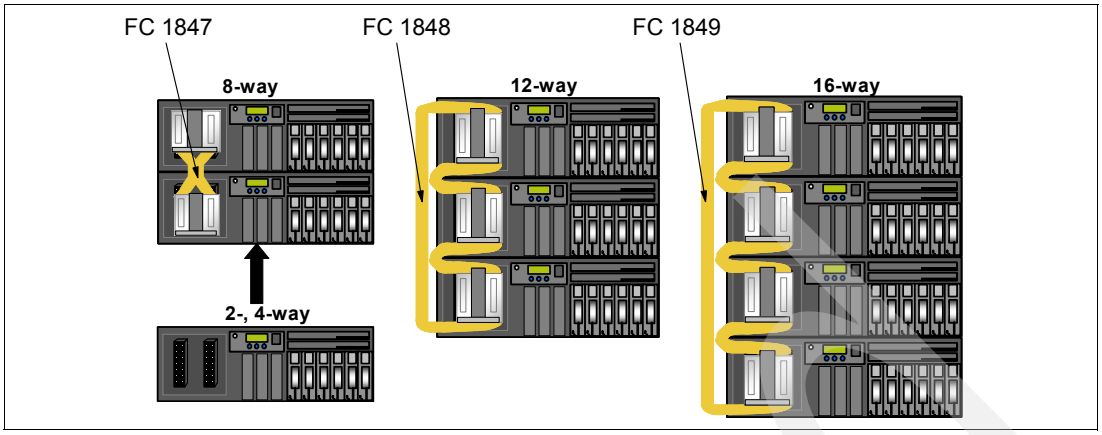


Figure 2-5 Logical p5-570 building blocks connection

2.2.2 Processor clock rate

The p5-570 system features base 2-core, 4-core, 8-core, 12-core, and 16-core configurations with the POWER5+ processor running at 1.9 GHz and 2.2 GHz. The processor card running at 1.9 GHz is available with 533 MHz DDR2 memory technology, and the 2.2 GHz card will accept either 533 MHz or 400 MHz memory.

Note: Any system made of more than one processor card must have all processor cards running at the same speed.

To verify the processor characteristics on a running system, use one of the following commands:

- **lsattr -El proc*X***

  Where *X* is the number of the processor, for example, proc0 is the first processor in the system. The output from the command is similar to the following output (*False*, as used in this output, signifies that the value cannot be changed through an AIX 5L command interface):

```
frequency   1900000000          Processor Speed       False
smt_enabled true                Processor SMT enabled False
smt_threads 2                   Processor SMT threads False
state       enable              Processor state       False
type        powerPC_POWER5      Processor type        False
```

- **pmcycles -m**

  The **pmcycles** command (available with AIX 5L) uses the performance monitor cycle counter and the processor real-time clock to measure the actual processor clock speed in MHz. The following output is from a 4-core p5-570 system running at 1.9 GHz with simultaneous multithreading enabled:

```
Cpu 0 runs at 1900 MHz
Cpu 1 runs at 1900 MHz
Cpu 2 runs at 1900 MHz
Cpu 3 runs at 1900 MHz
Cpu 4 runs at 1900 MHz
Cpu 5 runs at 1900 MHz
Cpu 6 runs at 1900 MHz
Cpu 7 runs at 1900 MHz
```

Note: The **pmcycles** command is part of the bos.pmapi fileset. Use the **lslpp -l bos.pmapi** command to determine if it is installed on your system.

2.3 Memory subsystem

The p5-570 memory controller is internal to the POWER5+ processor. It interfaces to either two (1.9 GHz) or four (2.2 GHz) SMI-II buffer chips and eight pluggable DIMMs per processor card, as described in 2.2, “Processor cards” on page 21. The minimum memory for a p5-570 processor-based system is 2 GB on a 1.9 GHz or 2.2 GHz system using 533 MHz DDR2 DIMMs or 32 GB on a 2.2 GHz system using 400 MHz DDR2 DIMMs. The maximum installable memory is 256 GB (using 533 MHz memory DIMM technology) or 512 GB (using 400 MHz memory DIMM technology only available on the 2.2 GHz processor card. The p5-570 total memory depends on the number of available processor cards. Figure 2-6 shows memory slot availability.

2.3.1 Memory placement rules

The memory features that are available for the p5-570 at the time of writing are listed in 1.3.2, “Memory features” on page 6. Each memory feature consists of four DIMMs, or quad, and must be installed according to Figure 2-6. The first quad slots are J0A, J1A, J0C, and J1C. For the second quad, the slots are J0B, J1B, J0D, and J1D. The placement rules for the 2.2 GHz card is the same.

Note: A quad must consist of a single feature (that is, made of four identical DIMMs). Mixing DIMM capacities in a quad is not supported.

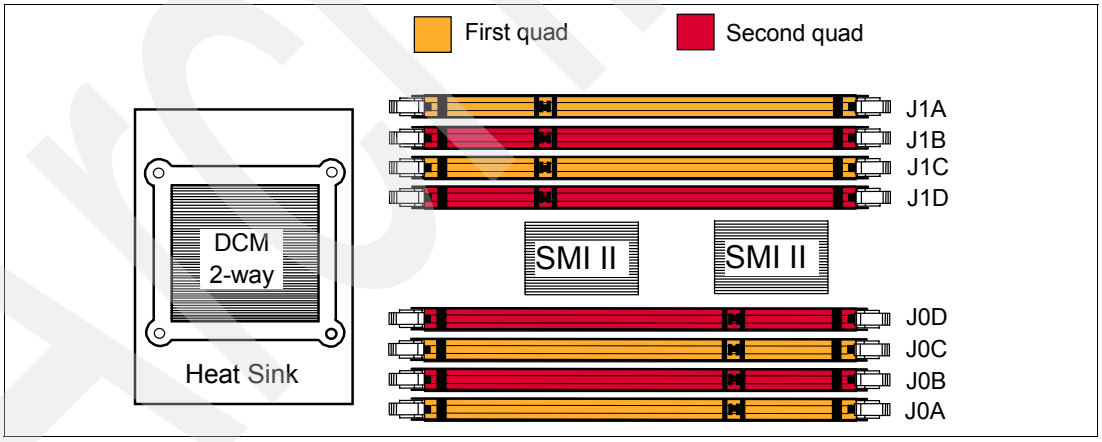


Figure 2-6 Memory placement for the p5-570 1.9 GHz DDR2 card

2.3.2 OEM memory

OEM memory is not supported or certified for use in IBM System p servers. If the p5-570 or server is populated with OEM memory, you could experience unexpected and unpredictable behavior, especially when the system is using Micro-Partitioning technology.

All IBM memory is identified by an IBM logo and a white label that is printed with a barcode and an alphanumeric string, as illustrated in Figure 2-7 on page 25.

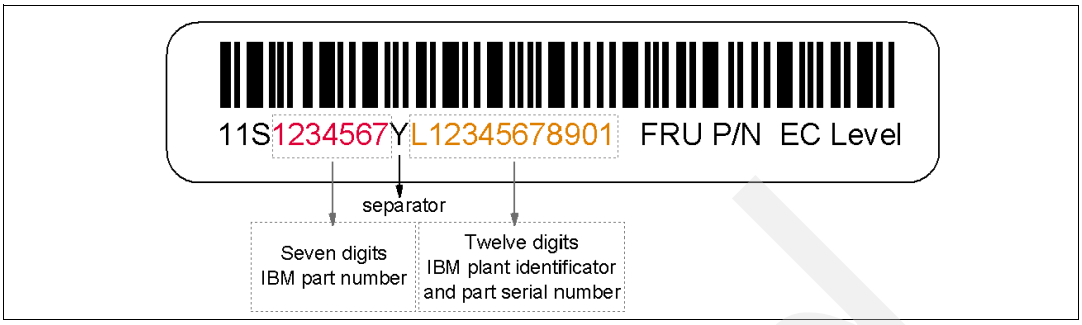


Figure 2-7 IBM memory certification label

2.3.3 Memory throughput

The memory subsystem throughput is based on the speed of the memory, not the speed of the processor. An elastic interface, contained in the POWER5+ processor, buffers reads and writes to and from memory and the processor. On 1.9 GHz processor cards, there are two SMIs, each with a single 8 byte read and 2 byte write DDR2 bus to the processor on each processor card. A DDR2 bus allows double reads or writes per clock cycle. If a 533 MHz memory feature is selected (which runs at 528 MHz), the throughput is (16 x 2 x 528) + (4 x 2 x 528) or 21120 MBps or 21.1 GBps per processor card. For a building block with two processor cards, this value is doubled.

2.2 GHz processor cards contain an additional set of two SMIs to manage the increased throughput. However, in this configuration, the paths are 4 bytes for read operations and 2 bytes for write. Therefore, the throughput is (4 + 2) * 4 * 2 * 528 = 25.34 GBps or 50.68 GBps for a 4-core node. These values are maximum theoretical throughputs for comparison purposes only.

Table 2-1 provides the theoretical throughputs values for different configurations.

Table 2-1 Theoretical throughput rates

Processor speed (GHz)	Processor type	Cores	Memory max. (GBps)	L2 to L3 (GBps)	GX+ RIO (GBps)
1.9	POWER5+	2-core	21.12	30.4	5
1.9	POWER5+	4-core	42.24	60.8	9.0
1.9	POWER5+	8-core	84.48	121.6	18.1
1.9	POWER5+	12-core	126.72	182.4	27.2
1.9	POWER5+	16-core	168.96	243.2	36.2
2.2	POWER5+	2-core	25.34	35.2	4.4
2.2	POWER5+	4-core	50.68	70.4	8.4
2.2	POWER5+	8-core	101.37	140.8	16.8
2.2	POWER5+	12-core	152.06	211.2	25.2
2.2	POWER5+	16-core	202.75	281.6	33.6

2.4 System buses

The following sections provide additional information related to the internal buses.

2.4.1 RIO-2 buses and GX+ card

Each DCM provides a GX+ bus that is used to connect to an I/O subsystem or Fabric Interface card. In a p5-570 drawer, there are two GX+ buses, one from each processor card. Each p5-570 has one GX slot with a single GX+ bus. The GX+ slot is not active unless the second processor card is installed. It is not required for CUoD processor cards to be activated in order for the associated GX+ bus to be active. All GX+ cards are hot-pluggable.

A choice of the following offer expansion opportunities:

- Remote I/O expansion card (FC 1800)
- Dual-port 4X Host Channel Adapter Infiniband card (FC 1810)

The p5-570 provides two external RIO-2 ports, which can operate up to 1 GHz. An add-in GX+ adapter card (Remote I/O expansion card, FC 1800) adds two more RIO-2 ports. When this card is installed, PCI adapter slot 6 must remain empty. The RIO-2 ports are used for I/O expansion to external I/O drawers. The supported I/O drawers are the 7311 Model D11 and 7311 Model D20. The Remote I/O expansion card must be installed starting with the first p5-570 building block.

2.4.2 SP bus

In addition to the processor drawer interconnect cable (described in 2.2.1, “Processor drawer interconnect cables” on page 22), the interconnection of multiple p5-570 building blocks requires the proper SP Flex cable to ensure the vital data communications between the building blocks. (See Figure 2-8.) The SP Flex cable contains the system interconnect signals, such as JTAG, I2C, clocks, and others.

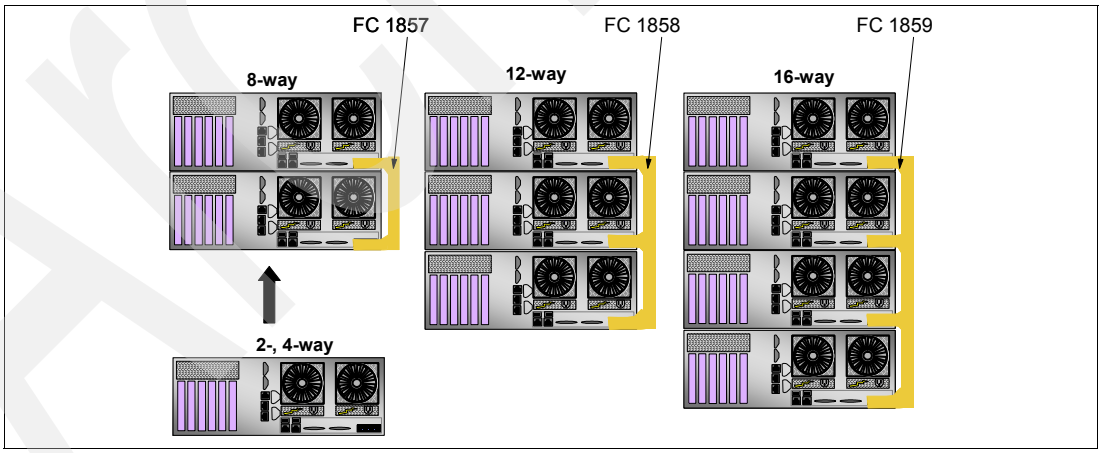


Figure 2-8 FSP Flex cables

2.5 Internal I/O subsystem

PCI-X, where the X stands for extended, is an enhanced PCI bus that delivers a bandwidth of up to 1 GBps, running a 64-bit bus at 133 MHz. PCI-X is backward-compatible, so the p5-570 can support existing 3.3 V PCI adapters.

The internal I/O subsystem resides on the system planar, and the SP is packaged on a separate service processor card. Each card is a separate FRU. An internal RIO-2 bus is imbedded in the system planar. The system planar contains both the Enterprise RIO-2 hub and the PCI-X Host bridge chip to connect to the integrated I/O that is packaged on the system planar. Two RIO-2 ports of the Enterprise hub chip are used for the integrated I/O and the remaining two ports are routed to external connectors.

The system planar provides six PCI-X slots and several integrated PCI devices that interface the three PCI-X to PCI-X bridges to the primary PCI-X buses on the PCI-X Host bridge chip.

PCI-X slot 6 can accept a short PCI-X or PCI card, and its space is shared with the Remote I/O expansion card, therefore if the Remote I/O expansion card is installed, this slot must remain empty. The remaining PCI-X slots are full-length cards. The dual 1 Gb Ethernet adapter is integrated on the system planar.

The PCI-X slots in the p5-570 system support hot-plug and Extended Error Handling (EEH). In the unlikely event of a problem, EEH-enabled adapters respond to a special data packet that is generated from the affected PCI-X slot hardware by calling system firmware, which examines the affected bus, allows the device driver to reset it, and continues without a system reboot.

2.6 64-bit and 32-bit adapters

IBM offers 64-bit adapter options for the p5-570, as well as 32-bit adapters. Higher-speed adapters use 64-bit slots because they can transfer 64 bits of data for each data transfer phase. Generally, 32-bit adapters can function in 64-bit PCI-X slots; however, some 64-bit adapters cannot be used in 32-bit slots. For a full list of the adapters that are supported on the p5-570 system, and for important information regarding adapter placement, see the *IBM Systems Hardware Information Center*. You can find it at:

http://publib.boulder.ibm.com/eserver/

The p5-570 internal PCI-X slots support a wide range of PCI-X I/O adapters to handle your I/O requirements.

2.6.1 LAN adapters

The dual port internal 10/100/1000 Mbps RJ-45 Ethernet controller integrated on the system planar can be used to connect to a local area network (LAN).

Table 2-2 lists additional LAN adapters that were available at the time of writing, when building an initial order of a p5-570 system. IBM supports an installation with NIM using Ethernet and token-ring adapters (CHRP[2] is the platform type).

Table 2-2 Available LAN adapters

Feature code	Adapter description	Slot	Size
5700	Gigabit Ethernet (Fiber)	64	short
5701	10/100/1000 Ethernet (UTP)	64	short
5706	2-port 10/100/1000 Ethernet (UTP)	64	short
5707	2-port Gigabit Ethernet - SX (Fiber)	64	short

[2] CHRP stands for Common Hardware Reference Platform, a specification for PowerPC-based systems that can run multiple operating systems.

Feature code	Adapter description	Slot	Size
5721	10 Gigabit Ethernet-SR PCI-X (Fiber)	64	short
5722	10 Gigabit Ethernet-LR PCI-X (Fiber)	64	short
5740	4-port 10/100/1000 Gigabit Ethernet PCI-X	64	short

2.6.2 Graphic accelerators

The p5-570 supports up to two enhanced POWER GXT135P (FC 2849) 2D graphic accelerators. The POWER GXT135P is a low-priced 2D graphics accelerator for pSeries and p5 servers. It can be configured to operate in either 8-bit or 24-bit color modes, running at 60 Hz to 85 Hz. This adapter supports both analog and digital monitors. The adapter requires one short 32-bit or 64-bit PCI-X slot.

2.6.3 SCSI adapters

To connect to external SCSI devices, the adapters listed in Table 2-3 are available, at the time of writing, to be used in p5-570 system.

Note: Previous SCSI adapters are also supported to be used in the p5-550Q, but cannot be part of an initial order configuration. Customers that would like to connect existing external SCSI devices can contact their IBM service representative.

Table 2-3 Available SCSI adapters

Feature code	Adapter description	Slot	Size
5736	Ultra320 SCSI PCI-X	64	short
5734	Ultra320 SCSI RAID PCI-X	64	long

2.6.4 Integrated RAID options

Every p5-570 building block system is delivered with a disk drive cage that supports up to six disk drive units, offering both internal RAID and non-RAID solutions. When internal RAID solution is not required, at least one 36.4 GB 15K disk drive (FC 3277) is required.

The internal RAID solution requires at least three 36.4 GB 15K disk drives (FC 3277) and the SCSI RAID Enablement Card (FC 5728). Other supported disk drives may be ordered in place of FC 3277. When the SCSI RAID Enablement Card is installed in the system, it re-sequences the two SCSI controllers that support the six disk drive bays, transforming the system from two logical 3-packs of disk drives to one physical 6-pack of disk drives. It does this by disabling one of the integrated SCSI adapters. This also adds the requirement that all six disks be assigned to a single LPAR when the card is present.

The RAID implementation requires a minimum of three disk drives to form a RAID array, so when an order comes in place with FC 5728, at least three disk drives must be in the order list.

Note: Because the p5-570 building block has six disk drive bays, customers performing upgrades must plan accordingly to ensure the correct handling of their RAID arrays.

The p5-570 system supports external RAID solutions, and this requires an additional PCI-X adapter (such as the FC 5737) and external disk drives enclosure.

RAID Capacity limitation: There are limits to the amount of disk drive capacity allowed in a single RAID array. Using the 32-bit AIX 5L kernel, there is capacity limitation of 1 TB per RAID array. Using the 64 bit kernel, there is a capacity limitation of 2 TB per RAID array. For RAID adapter and RAID enablement cards, this limitation is enforced by AIX 5L when RAID arrays are created using the PCI-X SCSI Disk Array Manager.

At the time of writing, AIX supports a maximum of 1 TB for SCSI RAID. Check with your IBM representative for current limits.

2.6.5 iSCSI

iSCSI is an open, standards-based approach by which SCSI information is encapsulated using the TCP/IP protocol to allow its transport over IP networks. It allows transfer of data between storage and servers in block I/O formats (defined by iSCSI protocol) and thus enables the creation of IP SANs. With iSCSI, an existing network can transfer SCSI commands and data with full location independence and define the rules and processes to accomplish the communication. The iSCSI protocol is defined in iSCSI IETF draft-20.

For more information about this standard, see:

http://tools.ietf.org/html/rfc3720

Although iSCSI can be, by design, supported over any physical media that supports TCP/IP as a transport, today's implementations are only on Gigabit Ethernet. At the physical and link level layers, systems that support iSCSI can be directly connected to standard Gigabit Ethernet switches and IP routers. iSCSI also enables the access to block-level storage that resides on Fibre Channel SANs over an IP network using iSCSI-to-Fibre Channel gateways, such as storage routers and switches.

The iSCSI protocol is implemented on top of the physical and data-link layers and presents the operating system with the standard SCSI Access Method command set. It supports SCSI-3 commands and reliable delivery over IP networks. The iSCSI protocol runs on the host initiator and the receiving target device. It can either be optimized in hardware for better performance on an iSCSI host bus adapter (such as FC 1986 and FC 1987 supported in IBM System p5 servers) or run in software over a standard Gigabit Ethernet network interface card. IBM in System p5 systems support iSCSI in the following two modes:

Hardware Using iSCSI adapters (see “IBM iSCSI adapters” on page 30).

Software Supported on standard Gigabit adapters, additional software (see “IBM iSCSI software Host Support Kit” on page 30) must be installed. The main processor is utilized for processing related to the iSCSI protocol.

Initial iSCSI implementations are targeted for small to medium-sized businesses and departments or branch offices of larger enterprises that have not deployed Fibre Channel SANs. iSCSI is an affordable way to create IP SANs from a number of local or remote storage devices. If there is Fibre Channel present, which it is typically in a data center, it can be accessed by the iSCSI SANs (and vice versa) using iSCSI-to-Fibre Channel storage routers and switches.

iSCSI solutions always involve the following software and hardware components:

Initiators These are the device drivers and adapters that are located on the client. They encapsulate SCSI commands and route them over the IP network to the target device.

Targets The target software receives the encapsulated SCSI commands over the IP network. The software can also provide configuration support and storage-management support. The underlying target hardware can be a storage appliance that contains embedded storage; it can also be a gateway or bridge product that contains no internal storage of its own.

IBM iSCSI adapters

New iSCSI adapters in IBM System p5 systems offer the advantage of increased bandwidth through the hardware support of the iSCSI protocol. The 1 Gigabit iSCSI TOE PCI-X adapters support hardware encapsulation of SCSI commands and data into TCP and transport it over the Ethernet using IP packets. The adapter operates as an iSCSI TCP/IP Offload Engine. This offload function eliminates host protocol processing and reduces CPU interrupts. The adapter uses Small form factor LC type fiber optic connector or copper RJ45 connector.

Table 2-4 lists the iSCSI adapters that can be ordered.

Table 2-4 Available iSCSI adapters

Feature code	Description	Slot	Size	Max
5713	Gigabit iSCSI TOE PCI-X on copper media adapter	64	short	3
5714	Gigabit iSCSI TOE PCI-X on optical media adapter	64	short	3

IBM iSCSI software Host Support Kit

The iSCSI protocol can also be used over standard Gigabit Ethernet adapters. To utilize this approach, download the appropriate iSCSI Host Support Kit for your operating system from the IBM NAS support Web site at:

`http://www.ibm.com/storage/support/nas/`

The iSCSI Host Support Kit on AIX 5L and Linux acts as a software iSCSI initiator and allows access to iSCSI target storage devices using standard Gigabit Ethernet network adapters. To ensure the best performance, enable TCP Large Send, TCP send and receive flow control, and Jumbo Frame for the Gigabit Ethernet Adapter and the iSCSI target. Tune network options and interface parameters for maximum iSCSI I/O throughput in the operating system.

IBM System Storage N series

The combination of System p5 and IBM System Storage™ N Series as the first of a new generation of iSCSI enabled storage products provides an end-to-end set of solutions. Currently the System Storage N series features three models: N3700, N5200, and N5500 with:

- Support for entry-level and midrange customers that require Network Attached Storage (NAS) or Internet Small Computer System Interface (iSCSI) functionality
- Support for Network File System (NFS), Common Internet File System (CIFS), and iSCSI protocols
- Data ONTAP software (at no charge), with plenty of additional functions such as data movement, consistent snapshots, and NDMP server protocol, some available through optional licensed functions

- Enhanced reliability with optional clustered (2-node) failover support.

2.6.6 Fibre Channel adapters

The PCI-X adapters are 64-bit, short- form factor with an LC-type external fibre connector that provides single or dual initiator capability over an optical fiber link or loop.

Table 2-5 lists the Fibre Channel adapters that can be ordered.

Table 2-5 Available Fibre Channel adapters

Feature code	Description	Slot	Size
5758	1-port 4 Gbps PCI-X Fibre Channel	64	short
5759	2-port 4 Gbps PCI-X Fibre Channel	64	short

The Fibre Channel PCI-X adapters can be used to attach devices either directly or with the supported Fibre Channel Switches. If you are attaching a device or switch with an SC-type fibre connector, an LC-SC 50 Micron Fiber Converter Cable (FC 2456) or an LC-SC 62.5 Micron Fiber Converter Cable (FC 2459) is also required.

2.6.7 InfiniBand Host Channel adapters

The p5-570 supports the PCI-X Dual-port 4x InfiniBand Host Channel Adapter (FC 1810), which allows the attachment of the Topspin Server Switch models 120 and 270. You use the 4x IB cables to connect to the Topspin Server Switches.

The PCI-X Dual-port 4x InfiniBand Host Channel Adapter is an Extra-high Bandwidth (EHB) PCI-X adapter. Only one EHB adapter can be connected per PCI-X Host Bridge (PHB). Optimal performance is not likely if EHB and other high performance adapters are installed together in one system.

Topspin Server Switch models 120 and 270

Switches are the fundamental components of an InfiniBand fabric. An IBM System p5 server proposal might also include the Topspin Server Switch model 120 and 270 in an initial system order.

The Topspin 120 and 270 Server Switch are a programmable switching platform that consists of a switched multiterabit interconnect and an intelligent control architecture. The high-bandwidth, low-latency interconnection is extremely adaptable. As a result, an outstanding level of application scaling, rapid deployment, and resource consolidation can be achieved.

See the following link for more Topspin Server Switch information:

http://www.topspin.com/solutions/index.htm

2.6.8 Asynchronous PCI-X adapters

The asynchronous PCI-X adapters provide a connection of asynchronous EIA-232 or RS-422 devices. In the case of a cluster configuration or high-availability configuration, if the plan is to connect the IBM System p5 servers using a serial connection, the use of the two default system ports is not supported, but one of the features in Table 2-6 is supported.

Table 2-6 Asynchronous PCI-X adapters

Feature code	Description
2943	8-Port Asynchronous Adapter EIA-232/RS-422
5723[a]	2-Port Asynchronous IEA-232 PCI Adapter (9-pin)

a. In many cases, the 5723 async adapter is configured to supply a backup HACMP heartbeat. In these cases, a serial cable (FC 3927 or FC 3928) must be also configured. Both of these serial cables and the 5723 adapter have 9-pin connectors.

2.6.9 Additional support for owned PCI-X adapters

The lists of the major PCI-X adapters that can be configured in a system when an initial configuration order is going to be built are described in 2.6.1, “LAN adapters” on page 27 to 2.6.1, “LAN adapters” on page 27, but the list of all the supported PCI-X adapters, with the related support for additional external devices, is more extended.

Clients that would like to use their own PCI-X adapters can contact the IBM service representative to verify if they are supported.

2.6.10 System ports

The system ports S1 and S2, at the rear of the system, are only available if the system is not managed using a Hardware Management Console (HMC). In this case, the S1 and S2 ports provide limited support of serial consoles or modems.

If an HMC is connected, a *virtual serial console* is provided by the HMC (logical device vsa0 under AIX) and also a modem can be connected to the HMC. The S1 and S2 ports are not usable in this case.

If serial port function is needed, optional PCI adapters are available, see 2.6.8, “Asynchronous PCI-X adapters” on page 32.

Only two system ports are active in a multiple drawer configuration and are the ports logically connected to the active service processor.

2.6.11 Ethernet ports

The two built-in Ethernet ports provide 10/100/1000 Mbps connectivity over CAT-5 cable for up to 100 meters. Table 2-7 on page 33 lists the attributes of the LEDs that are visible on the side of the jack.

Table 2-7 Ethernet LED descriptions

LED	Light	Description
Link	Off Green	No link; could indicate a bad cable, not selected, or configuration error Connection established
Activity	On Off	Data activity Idle

2.7 Internal storage

Two Ultra320 SCSI controllers under EADS-X chips that are integrated into the system planar are used to drive the internal disk drives. The six internal drives plug into the disk drive backplane, which has two separate SCSI buses and controllers with three disk drives per bus. Each of these controllers can be dynamically assigned to partitions if required.

The internal disk drive bays can be used in two different modes, depending on whether the SCSI RAID Enablement Card (FC 5728) is installed. (See 2.6.4, “Integrated RAID options” on page 28.)

The p5-570 supports a split 6-pack disk drive backplane, which is designed for hot-pluggable disk drives. The disk drive backplane docks directly to the system planar. The virtual SCSI Enclosure Services (VSES) hot plug control functions are provided by the Ultra320 SCSI controllers.

2.7.1 Internal hot swappable SCSI disks

The p5-570 can have up to six hot-swappable disk drives plugged in the two logical 3-pack disk drive backplanes. The hot-swap process is controlled by the virtual SCSI Enclosure Services (VSES), which is located in the logical 3-pack disk drive backplane. (AIX assigns the name vses0 to the first 3-pack, and vses1 to the second, if present.) The two logical 3-pack disk drive backplanes can accommodate the devices listed in Table 2-8.

Table 2-8 Hot-swappable disk options

Feature code	Description
3277	36.4 GB 15,000 RPM Ultra320 SCSI hot-swappable disk drive
3274	73.4 GB 10,000 RPM Ultra320 SCSI hot-swappable disk drive
3278	73.4 GB 15,000 RPM Ultra320 SCSI hot-swappable disk drive
3275	146.8 GB 10,000 RPM Ultra320 SCSI hot-swappable disk drive
3279	146.8 GB 150,000 RPM Ultra320 SCSI hot-swappable disk drive
3278	300 GB 10,000 RPM Ultra320 SCSI hot-swappable disk drive

At the time of writing, if a new order is placed with more than one disk, the system configuration that is shipped from manufacturing may balance the total number of SCSI disks between the two logical 3-pack SCSI backplanes. In this case, this is for manufacturing test purposes and not because of any limitation. Having the disks balanced between the two 3-pack disk drive backplanes enables the manufacturing process to systematically test the SCSI paths and the devices related to them.

Prior to the hot-swap of a disk drive in the hot-swappable-capable bay, all necessary operating system actions must be undertaken to ensure that the disk is capable of being deconfigured. After the disk drive has been deconfigured, the SCSI enclosure device will power-off the slot, enabling safe removal of the disk. You should ensure that the appropriate planning has been given to any operating-system-related disk layout, such as the AIX Logical Volume Manager, when using disk hot-swap capabilities. For more information, see *Problem Solving and Troubleshooting in AIX 5L*, SG24-5496.

Note: We recommend that you follow this procedure, after the disk has been deconfigured, when removing a hot-swappable disk drive:

1. Release the tray handle on the disk assembly.
2. Pull out the disk assembly a little bit from the original position.
3. Wait up to 20 seconds until the internal disk stops spinning.

Now you can safely remove the disk from the DASD backplane.

After the SCSI disk hot-swap procedure, you can expect to find `SCSI_ERR10` logged in the AIX error log, with the second word of the sense data equal to `0017`. It is generated from a SCSI bus reset that is issued by the VSES to reset all processes when a drive is inserted, and it is not an issue.

Hot-swappable disks and Linux

Hot-swappable disk drives on IBM System p5 systems are supported with SUSE Linux Enterprise Server 9 for POWER, or later, and Red Hat Enterprise Linux AS for POWER Version 3, or later.

2.7.2 Internal media devices

Inside each CEC drawer in the p5-570 there is an optional media backplane with two media bays. These devices are treated as a group with respect to logical partitioning. The two internal IDE media bays in separate CEC drawers can be allocated or assigned to a different partition together as a pair. The media backplane inside each CEC drawer cannot be split between two logical partitions.

2.8 External I/O subsystems

This section describes the external I/O subsystems, which include the 7311 I/O drawers and the external storage solutions that p5-570 supports.

2.8.1 I/O drawers

As described in Chapter 1, “General description” on page 1, the p5-570 system has six internal PCI-X slots. If more PCI-X slots are needed to dedicate more adapters to a partition or to increase the bandwidth of network adapters, up to 20 7311 model D10, 7311 model D11, and 7311 model D20 I/O drawers can be added to the p5-570 system.

The p5-570 building block system configures a default RIO-2 bus to connect the internal PCI-X slots through the PCI-X to PCI-X bridges, and supports up to four external I/O drawers. To support more I/O drawers in one p5-570 building block, the RIO-2 expansion card (FC 1800) is needed. The RIO-2 expansion card supports up to four additional I/O drawers. If the combined system is made of more than one p5-570 building block, the optional RIO-2

expansion card may be not required if I/O drawers can be shared between the default RIO-2 loop of the p5-570 building blocks.

2.8.2 7311 Model D11 I/O drawers

The 7311-D11 provides six additional PCI-X slots supporting an enhanced blind-swap mechanism. The 7311-D11 has six slots that are PCI-X capable. Drawers must have a RIO-2 adapter to connect to the server.

Each primary PCI-X bus is connected to a PCI-X-to-PCI-X bridge, which provides three slots with Extended Error Handling (EEH) for error recovering. In the 7311 Model D11 I/O drawer, slots 1 to 6 are PCI-X slots that operate at 133 MHz and 3.3 V signaling. Figure 2-9 shows a conceptual diagram of the 7311 Model D11 I/O drawer.

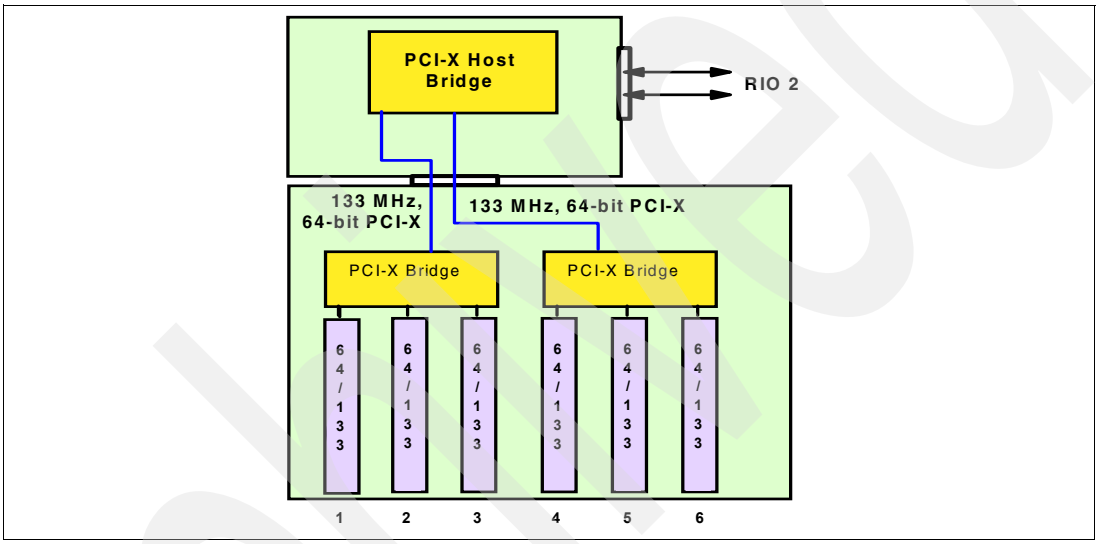


Figure 2-9 Conceptual diagram of the 7311-D11 I/O drawer

7311 Model D11 features

This I/O drawer model provides the following features:

- Six hot-plug 64-bit, 133 MHz, 3.3 V PCI-X slots, full length, enhanced blind-swap cassette
- Default redundant hot-plug power and cooling
- Two default remote (RIO-2) ports and two SPCN ports

2.8.3 7311 Model D20 I/O drawer

The 7311 Model D20 I/O drawer must have the RIO-2 loop adapter (FC 6417) to be connected to the p5-570 system. The PCI-X host bridge inside the I/O drawer provides two primary 64-bit PCI-X buses running at 133 MHz. Therefore, a maximum bandwidth of 1 GBps is provided by each of the buses. To avoid overloading an I/O drawer, follow the recommendation in the IBM @server Hardware Information Center at:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/

Figure 2-10 shows a conceptual diagram of the 7311 Model D20 I/O drawer subsystem.

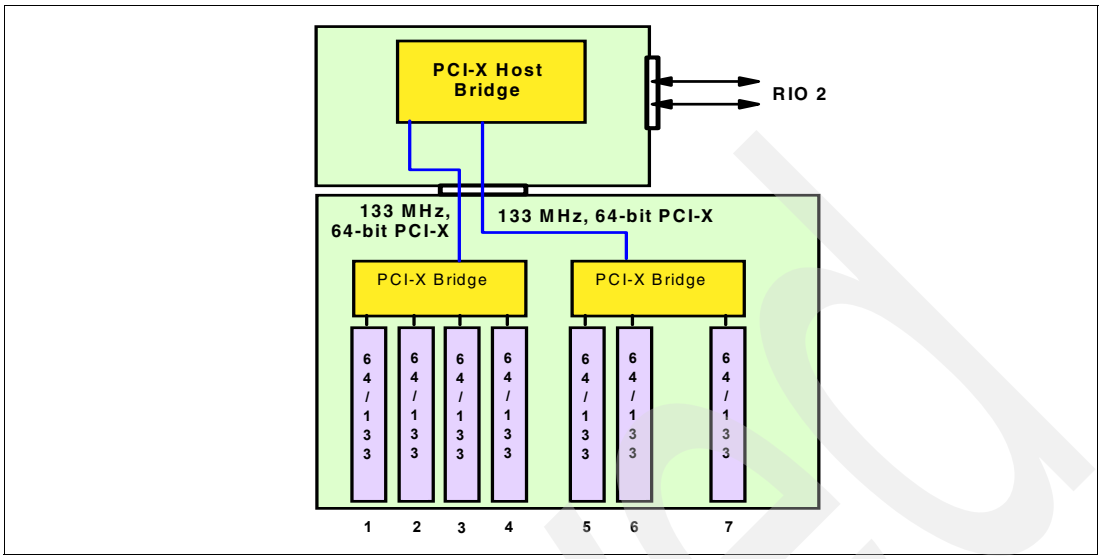


Figure 2-10 Conceptual diagram of the 7311-D20 I/O drawer

7311 Model D20 internal SCSI cabling

A 7311 Model D20 supports hot-swappable disks using two 6-pack disk bays for a total of 12 disks. Additionally, the SCSI cables (FC 4257) are used to connect a SCSI adapter (any of various features) in slot 7 to each of the 6-packs, or two SCSI adapters, one in slot 4 and one in slot 7. (See Figure 2-11.)

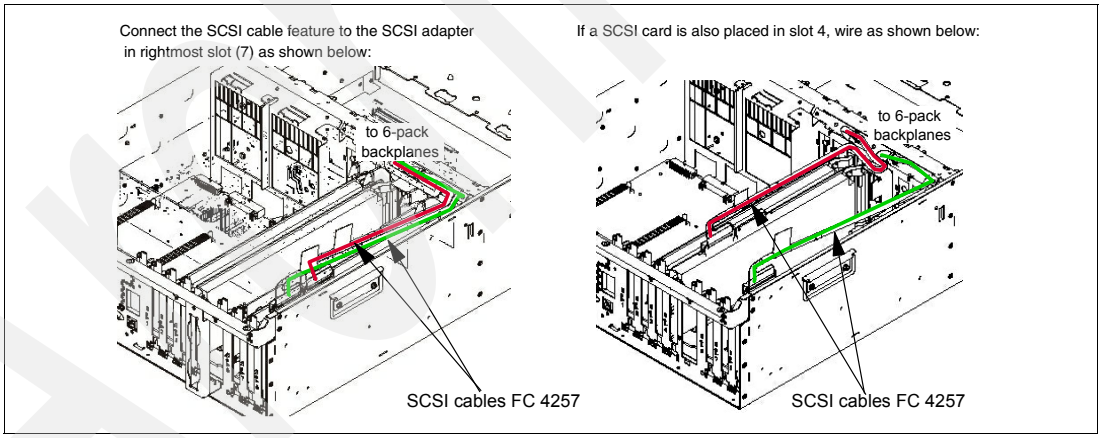


Figure 2-11 7311 Model D20 internal SCSI cabling

Note: Any 6-packs and the related SCSI adapter can be assigned to a partition. If one SCSI adapter is connected to both 6-packs, then both 6-packs can be assigned only to the same partition.

2.8.4 7311 I/O drawer and RIO-2 cabling

As described in 2.8, “External I/O subsystems” on page 34, we can connect up to four I/O drawers in the same loop, and up to 20 I/O drawers to the p5-570 system.

Each RIO-2 port can operate at 1 GHz in bidirectional mode and is capable of passing data in each direction on each cycle of the port. Therefore, the maximum data rate is 4 GBps per I/O drawer in double barrel mode (using two ports).

There is one default primary RIO-2 loop in any p5-570 building block. This feature provides two Remote I/O ports for attaching up to four 7311 Model D11 or 7311 Model D20 I/O drawers to the system in a single loop. Different I/O drawer models can be used in the same loop, but the combination of I/O drawers must be a total of four per single loop. The optional RIO-2 expansion card may be used to increase the number of I/O drawers that can be connected to one p5-570 building block, and the same rules of the default RIO-2 loop must be considered. The method that is used to connect the drawers to the RIO-2 loop is important for performance.

Figure 2-12 shows how you could connect four I/O drawers to one p5-570 building block. This is a logical view; actual cables should be wired according to the installation instructions.

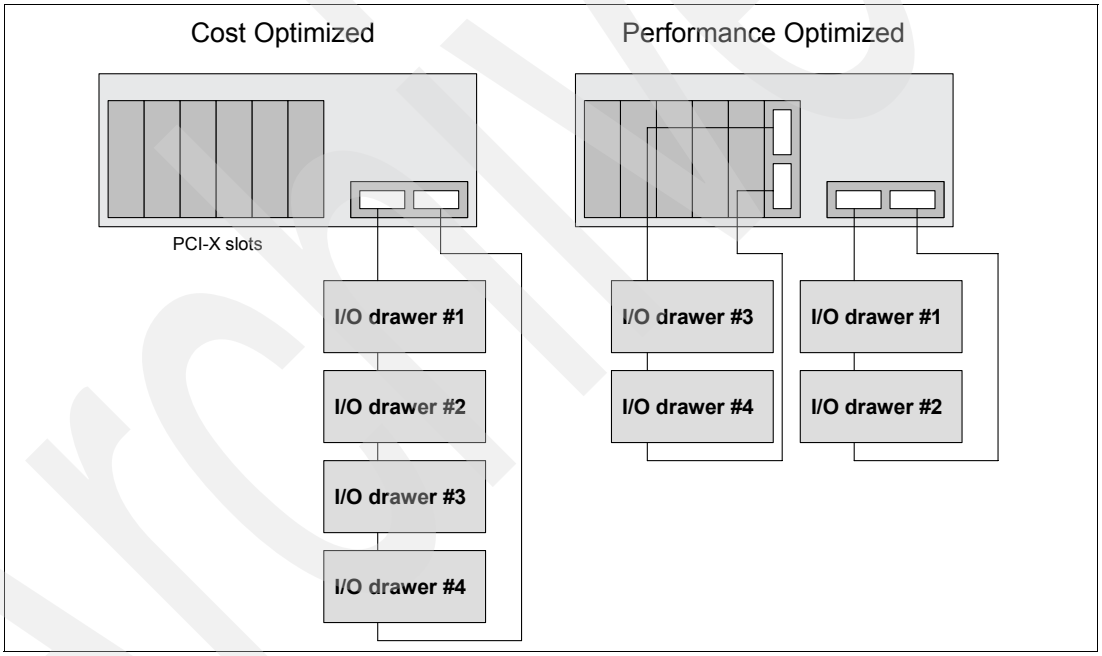


Figure 2-12 RIO-2 cabling examples

Note: If you have 20 I/O drawers, although there are no restrictions on their placement, this can affect performance.

RIO-2 cables are available in different lengths to satisfy different connection requirements:

- Remote I/O cable, 1.2 m (FC 3146, for between D11 drawers only)
- Remote I/O cable, 1.75 m (FC 3156)
- Remote I/O cable, 2.5 m (FC 3168)
- Remote I/O cable, 3.5 m (FC 3147)
- Remote I/O cable, 10 m (FC 3148)

2.8.5 7311 I/O drawer and SPCN cabling

SPCN[3] is used to control and monitor the status of power and cooling within the I/O drawer. SPCN is a loop: Cabling starts from SPCN port 0 on the p5-570 to SPCN port 0 on the first I/O drawer. The loop is closed, connecting the SPCN port 1 of the I/O drawer back to port 1 of the p5-570 system. If you have more than one I/O drawer, you continue the loop, connecting the next drawer (or drawers) with the same rule.

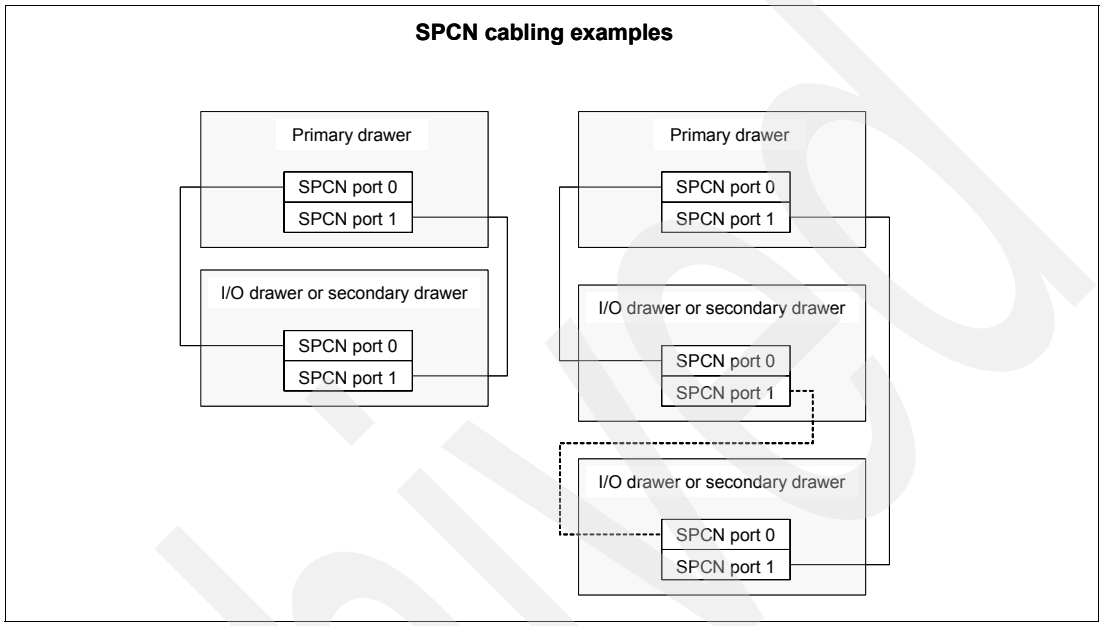


Figure 2-13 SPCN cabling examples

There are different SPCN cables to satisfy different length requirements:

- SPCN cable drawer-to-drawer, 2 m (FC 6001)
- SPCN cable drawer-to-drawer, 3 m (FC 6006)
- SPCN cable rack-to-rack, 6 m (FC 6008)
- SPCN cable rack-to-rack, 15 m (FC 6007)
- SPCN cable rack-to-rack, 30 m (FC 6029)

2.9 External disk subsystems

The p5-570 has internal hot-swappable drives. When the AIX 5L operating system is installed in a IBM System p5 server, the internal disks are usually used for the AIX 5L rootvg volume group and paging space. Specific client requirements can be satisfied with the several external disk possibilities that the p5-570 supports.

2.9.1 IBM TotalStorage EXP24 Expandable Storage

The IBM TotalStorage® EXP24 Expandable Storage disk enclosure, Model D24 or T24, can be purchased together with the p5-570 and will provide low-cost Ultra320 (LVD) SCSI disk storage. This disk storage enclosure device provides more than 7 TB of disk storage in a 4U rack-mount (Model D24) or compact deskside (Model T24) unit. Whether high availability

[3] System Power Control Network

storage solutions or simply high capacity storage for a single server installation, the unit provides a cost-effective solution. It provides 24 hot-swappable disk bays, 12 accessible from the front and 12 from the rear. Disk options that can be accommodated in any of the four six-packs disk drive enclosure are 73.4 GB, 146.8 GB, or 300 GB 10 K rpm or 36.4 GB, 73.4 GB, or 146.8 GB 15 K rpm drives. Each of the four six-packs disk drive enclosure might be attached independently to an Ultra320 SCSI or Ultra320 SCSI RAID adapter. For high available configurations, a dual bus repeater card (FC 5742) allows each six-pack to be attached to two SCSI adapters, installed in one or multiple servers or logical partitions. Optionally, the two front or two rear six-packs might be connected together to form a single Ultra320 SCSI bus of 12 drives.

2.9.2 IBM System Storage N3000 and N5000

The IBM System Storage N3000 and N5000 line of iSCSI enabled storage offerings provide a flexible way to implement a Storage Area Network over an Ethernet network. Flexible-Fibre Channel and SATA disk drive capabilities allow for deployment in multiple solution environments, including data compliant retention, nearline storage, disk-to-disk backup scenarios, and high-performance mission-critical I/O intensive operations. See the following link for more information:

http://www.ibm.com/servers/storage/nas

2.9.3 IBM TotalStorage Storage DS4000 Series

The IBM System Storage DS4000™ line of Fibre Channel enabled Storage offerings provides a wide range of storage solutions for your Storage Area Network. The IBM TotalStorage DS4000 Storage server family consists of the following models: DS4100, DS4300, DS4500, and DS4800. The Model DS4100 Express Model is the smallest model and scales up to 44.8 TB; the Model DS4800 is the largest and scales up to 89.6 TB of disk storage at the time of this writing. Model DS4300 provides up to 16 bootable partitions, or 64 bootable partitions if the turbo option is selected, that are attached with the Gigabit Fibre Channel Adapter (FC 1977). Model DS4500 provides up to 64 bootable partitions. Model DS4800 provides 4 GB switched interfaces. In most cases, both the IBM TotalStorage DS4000 family and the IBM System p5 servers are connected to a storage area network (SAN). If only space for the rootvg is needed, the Model DS4100 is a good solution.

For support of additional features and for further information about the IBM TotalStorage DS4000 Storage Server family, refer to the following Web site:

http://www.ibm.com/servers/storage/disk/ds4000/index.html

2.9.4 IBM TotalStorage Enterprise Storage Server

The IBM TotalStorage Enterprise Storage Server® (ESS) Models DS6000™ and DS8000™ are the high-end premier storage solution for use in storage area networks and use POWER technology-based design to provide fast and efficient serving of data. The IBM TotalStorage DS6000 provides enterprise class capabilities in a space-efficient modular package. It scales to 67.2 TB of physical storage capacity by adding storage expansion enclosures. The Model DS8000 series is the flagship of the IBM TotalStorage DS family. The DS8000 scales to 192 TB. However, the system architecture is designed to scale to over one petabyte. The Model DS6000 and DS8000 systems can also be used to provide disk space for booting LPARs or partitions using Micro-Partitioning technology. ESS and the IBM System p5 servers are usually connected together to a storage area network.

For further information about ESS, refer to the following Web site:

http://www.ibm.com/servers/storage/disk/enterprise/ds_family.html

2.10 Logical partitioning

Dynamic logical partitions (LPARs) and virtualization increase utilization of system resources and add a new level of configuration possibilities. This section provides details and configuration specifications about this topic. The virtualization discussion includes virtualization enabling technologies that are standard on the system, such as the POWER Hypervisor, and optional ones, such as the Advanced POWER Virtualization feature.

2.10.1 Dynamic logical partitioning

Logical partitioning (LPAR) was introduced with the POWER4 processor-based product line and the AIX 5L Version 5.1 operating system. This technology offered the capability to divide a pSeries system into separate logical systems, allowing each LPAR to run an operating environment on dedicated attached devices, such as processors, memory, and I/O components.

Later, dynamic LPAR increased the flexibility, allowing selected system resources, such as processors, memory, and I/O components, to be added and deleted from dedicated partitions while they are executing. AIX 5L Version 5.2, with all the necessary enhancements to enable dynamic LPAR, was introduced in 2002. The ability to reconfigure dynamic LPARs encourages system administrators to dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

Operating system support for dynamic LPAR

Table 2-9 lists AIX 5L and Linux support for dynamic LPAR capabilities.

Table 2-9 Operating system supported function

Function	AIX 5L Version 5.2	AIX 5L Version 5.3	Linux SLES 9	Linux RHEL AS 3	Linux RHEL AS 4
Dynamic LPAR capabilities (add, remove, and move operations)					
Processor	Y	Y	Y	N	Y
Memory	Y	Y	N	N	N
I/O slot	Y	Y	Y	N	Y

2.11 Virtualization

With the introduction of the POWER5 processor, partitioning technology moved from a dedicated resource allocation model to a virtualized shared resource model. This section briefly discusses the key components of virtualization on System p5 and @server p5 servers.

For more information about virtualization, see the following Web site:

http://www.ibm.com/servers/eserver/about/virtualization/systems/pseries.html

You can also consult the following IBM Redbooks:

http://www.redbooks.ibm.com/abstracts/sg247940.html?Open
http://www.redbooks.ibm.com/abstracts/sg245768.html?Open

2.11.1 POWER Hypervisor

Combined with features designed into the POWER5 and POWER5+ processors, the POWER Hypervisor delivers functions that enable other system technologies, including Micro-Partitioning technology, virtualized processors, IEEE VLAN, compatible virtual switch, virtual SCSI adapters, and virtual consoles. The POWER Hypervisor is a basic component of system firmware that is always active, regardless of the system configuration.

The POWER Hypervisor provides the following functions:

- Provides an abstraction between the physical hardware resources and the logical partitions that uses them.
- Enforces partition integrity by providing a security layer between logical partitions.
- Controls the dispatch of virtual processors to physical processors (see 2.12.2, “Logical, virtual, and physical processor mapping” on page 44).
- Saves and restores all processor state information during a logical processor context switch.
- Controls hardware I/O interrupt management facilities for logical partitions.
- Provides virtual LAN channels between physical partitions that help to reduce the need for physical Ethernet adapters for inter-partition communication.

The POWER Hypervisor is always active when the server is running partitioned or not and also when not connected to the HMC. It requires memory to support the logical partitions on the server. The amount of memory required by the POWER Hypervisor firmware varies according to several factors. Factors influencing the POWER Hypervisor memory requirements include the following:

- Number of logical partitions
- Partition environments of the logical partitions
- Number of physical and virtual I/O devices used by the logical partitions
- Maximum memory values given to the logical partitions

Note: Use the System Planning Tool to estimate the memory requirements of the POWER Hypervisor.

In AIX 5L V5.3, the `lparstat` command using the -h and -H flags displays the POWER Hypervisor statistical data. Using the -h flag adds summary POWER Hypervisor statistics to the default `lparstat` output.

The minimum amount of physical memory for each partition is 128 MB, but in most cases, the actual requirements and recommendations are between 256 MB and 512 MB for AIX 5L, Red Hat, and Novell SUSE Linux. Physical memory is assigned to partitions in increments of Logical Memory Block (LMB). For POWER5+ processor-based systems, LMB might be adjusted from 16 MB to 256 MB.

The POWER Hypervisor provides the following types of virtual I/O adapters:

- Virtual SCSI
- Virtual Ethernet
- Virtual (TTY) console

Virtual SCSI

The POWER Hypervisor provides a virtual SCSI mechanism for virtualization of storage devices (a special logical partition to install the Virtual I/O Server is required to use this feature, as described in 2.12.3, “Virtual I/O Server” on page 46). The storage virtualization is accomplished using two, paired, adapters: a virtual SCSI server adapter and a virtual SCSI client adapter. Only the Virtual I/O Server partition can define virtual SCSI server adapters, other partitions are *client* partitions. The Virtual I/O Server is available with the optional Advanced POWER Virtualization feature (FC 7942).

Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions on the same server to use a fast and secure communication without any need for physical interconnection. The virtual Ethernet allows a transmission speed in the range of 1 to 3 GBps. depending on the MTU[4] size and CPU entitlement. Virtual Ethernet requires system with either AIX 5L Version 5.3 or appropriate level of Linux supporting virtual Ethernet devices (see chapter 2.14, “Operating system support” on page 54). The virtual Ethernet is part of the base system configuration.

Virtual Ethernet has the following major features:

- The virtual Ethernet adapters can be used for both IPv4 and IPv6 communication and can transmit packets with a size up to 65408 bytes. Therefore, the maximum MTU for the corresponding interface can be up to 65394 (65390 if VLAN tagging is used).
- The POWER Hypervisor presents itself to partitions as a virtual 802.1Q compliant switch. The maximum number of VLANs is 4096. Virtual Ethernet adapters can be configured as either untagged or tagged (following the IEEE 802.1Q VLAN standard).
- A partition supports 256 virtual Ethernet adapters. Besides a default port VLAN ID, the number of additional VLAN ID values that can be assigned per Virtual Ethernet adapter is 20, which implies that each Virtual Ethernet adapter can be used to access 21 virtual networks.
- Each partition operating system detects the virtual local area network (VLAN) switch as an Ethernet adapter without the physical link properties and asynchronous data transmit operations.

Any virtual Ethernet can also have connection outside of the server if a layer-2 bridging to a physical Ethernet adapter is set in one Virtual I/O server partition (see 2.12.3, “Virtual I/O Server” on page 46 for more details about shared Ethernet).

Note: Virtual Ethernet is based on the IEEE 802.1Q VLAN standard. No physical I/O adapter is required when creating a VLAN connection between partitions, and no access to an outside network is required.

Virtual (TTY) console

Each partition needs to have access to a system console. Tasks such as operating system installation, network setup, and some problem analysis activities require a dedicated system console. The POWER Hypervisor provides the virtual console using a virtual TTY or serial adapter and a set of Hypervisor calls to operate on them. Virtual TTY does not require the purchase of any additional features or software such as the Advanced POWER Virtualization feature.

[4] Maximum transmission unit

Depending on the system configuration, the operating system console can be provided by the Hardware Management Console virtual TTY, IVM virtual TTY, or from a terminal emulator that is connected to a system port.

2.12 Advanced POWER Virtualization feature

The Advanced POWER Virtualization feature (FC 7942) is an optional, additional cost feature. This feature enables the implementation of more fine-grained virtual partitions on IBM System p5 servers.

The Advanced POWER Virtualization feature includes:

- Firmware enablement for Micro-Partitioning technology.

  Support for up to 10 partitions per processor using 1/100 of the processor granularity. The minimum CPU requirement per partition is 1/10. All processors are enabled for micro-partitions (number of processors on system equals the number of Advanced POWER Virtualization features ordered).
- An Installation image for the Virtual I/O Server software that is shipped as a system image on a DVD. Client partitions can be either AIX 5L V5.3 or Linux. It supports:
  - Ethernet adapter sharing (Ethernet bridge from virtual Ethernet to external network)
  - Virtual SCSI Server
  - Partition management using Integrated Virtualization Manager (Virtual I/O Server Version 1.2 or later only)
- Partition Load Manager (AIX 5L Version 5.3 only)
  - Automated CPU and memory reconfiguration
  - Real-time partition configuration and load statistics
  - Graphical user interface

For more details about Advanced POWER Virtualization and virtualization in general, see:

http://www.ibm.com/servers/eserver/pseries/ondemand/ve/resources.html

2.12.1 Micro-Partitioning technology

The concept of Micro-Partitioning technology allows you to allocate fractions of processors to the partition. The Micro-Partitioning technology is only available with POWER5 and POWER5+ processor-based systems. From an operating system perspective, a virtual processor cannot be distinguished from a physical processor, unless the operating system has been enhanced to be made aware of the difference. Physical processors are abstracted into virtual processors that are available to partitions. See 2.12.2, “Logical, virtual, and physical processor mapping” on page 44 for more details.

When defining a shared partition, several options have to be defined:

- The minimum, desired, and maximum processing units. Processing units are defined as processing power, or fraction of time, the partition is dispatched on physical processors.
- The processing sharing mode, either capped or uncapped.
- The weight (preference) in the case of uncapped partition.
- The minimum, desired, and maximum number of virtual processors.

The POWER Hypervisor calculates partition's processing *entitlement* based on minimum, desired, and maximum values, sharing mode and also based on other active partitions' requirements. The actual entitlement is never smaller than the minimum value but can exceed the maximum value in the case of an uncapped partition.

A partition can be defined with a processor capacity as small as 0.10 processing units. This represents one-tenth of a physical processor. Each physical processor can be shared by up to 10 shared processor partitions and the partition's entitlement can be incremented fractionally by as little as one-hundredth of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC or Integrated Virtualization Management (included with Virtual I/O Server software version 1.2 or later). There is only one pool of shared processors at the time of the writing of this Redpaper and all shared partitions are dispatched by the Hypervisor within this pool. Dedicated partitions and micro-partitions can coexist on the same POWER5+ processor-based server as long as enough processors are available.

The p5-570 supports up to a 16-core configuration, therefore up to sixteen dedicated partitions, or up to 160 micro-partitions, can be created. It is important to point out that the maximums stated are supported by the hardware, but the practical limits depend from the application workload demands.

2.12.2 Logical, virtual, and physical processor mapping

The meaning of the term *physical processor* in this section is a *processor core*. For example, in a 2-core server with a Dual-Core Module (DCM), there are two physical processors.

In dedicated mode, physical processors are assigned as a whole to partitions. The simultaneous multithreading feature in the POWER5+ processor core allows the core to execute instructions from two independent software threads simultaneously. To support this feature, the concept of *logical processors* was introduced. The operating system (AIX 5L or Linux) sees one physical processor as two logical processors if the simultaneous multithreading feature is on. It can be turned off while the operating system is executing (for AIX 5L, use the **`smtctl`** command). If simultaneous multithreading is off, then each physical processor is presented as one logical processor and thus only one thread is executed on the physical processor at the time.

In a micro-partitioned environment with shared mode partitions, an additional concept of *virtual processors* was introduced. Shared partitions can define any number of virtual processors (maximum number is 10 times the number of processing units assigned to the partition). From the POWER Hypervisor point of view, the virtual processors represent dispatching objects (for example, the POWER Hypervisor dispatches virtual processors to physical processors according to partition's processing units entitlement). At the end of the POWER Hypervisor's dispatch cycle (10 ms), all partitions should receive total CPU time equal to their processing units entitlement. Virtual processors are either running (dispatched) on a physical processor or standby (waiting). An operating system is able to dispatch its software threads to these virtual processors and is completely screened from the actual number of physical processors. The logical processors are defined on top of virtual processors in the same way as though they are physical processors. So, even with a virtual processor, the concept of logical processor exists and the number of logical processor depends whether the simultaneous multithreading is turned on or off.

Some additional information related to the virtual processors is as follows:

- There is one-to-one mapping of running virtual processors to physical processors at any given time. The number of virtual processors that can be active at any given time cannot exceed the total number of physical processors in shared processor pool.
- A virtual processor can be either running (dispatched) on a physical processor or standby waiting for a physical processor to became available.
- Virtual processors do not introduce any additional abstraction level; they really are only a dispatch entity. When running on a physical processor, virtual processors run at the same speed as the physical processor.
- Each partition's profile defines CPU entitlement that determines how much processing power any given partition should receive. The total sum of CPU entitlement of all partitions cannot exceed the number of available physical processors in a shared processor pool.
- A partition has the same amount of processing power regardless of the number of virtual processors that it defines.
- A partition can use more processing power, regardless of its entitlement, if it is defined as an *uncapped* partition in the partition profile. If there is spare processing power available in shared processor pool or other partitions are not using their entitlement, an uncapped partition can use additional processing units if its entitlement is not enough to satisfy its application processing demand in the given processing entitlement.
- When the partition in uncapped, the number of defined virtual processors determines the limitation of the maximum processing power it can receive. For example, if the number of virtual processors is two, then the maximum usable processor units is two.
- You are allowed to define more virtual processors than physical processors. In that case, a virtual processor will be waiting for dispatch more often and some performance impact caused by redispatching virtual processors on physical processors should be considered. It is also true that some applications can benefit from using more virtual processors than the physical processors.
- The number of virtual processors can be changed dynamically through a dynamic LPAR operation.

Virtual processor recommendations

For each partition, you can define a number of virtual processors set to the maximum processing power the partition could ever request. If there are, for example, four physical processors installed in the system, and one production partition and three test partitions, then:

- Define the production LPAR with four virtual processors so that it can receive the full processing power of all four physical processors during the time the other partitions are idle.
- If you know that the test system will never consume more than one processor computing unit, then they should be defined with one virtual processor. Some test systems might require additional virtual processors, such as four, in order to use idle processing power left over by a production system during off-business hours.

Figure 2-14 shows logical, virtual, and physical processor mapping, and an example of how the virtual processor and logical processor can be dispatched to the physical processor.

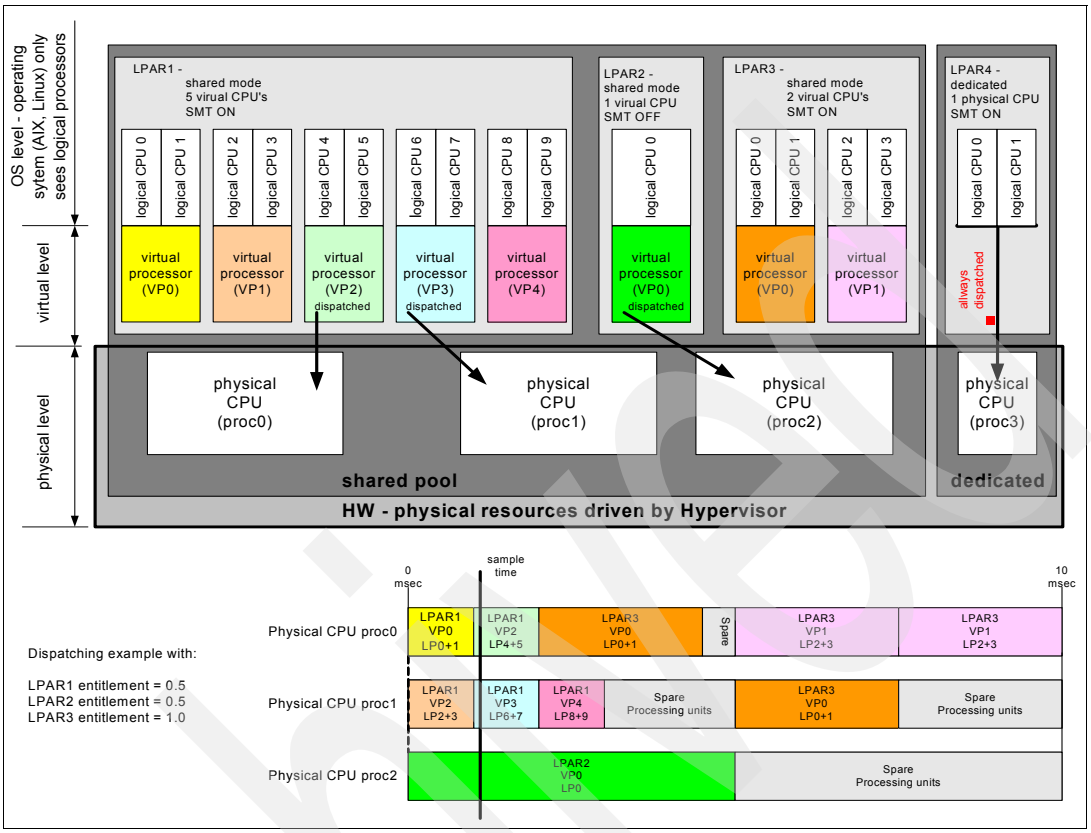


Figure 2-14 Logical, virtual, and physical processor mapping

In Figure 2-14, a system with four physical processors and four partitions is presented; one partition (LPAR4) is in dedicated mode and three partitions (LPAR1, LPAR2, and LPAR3) are running in shared mode. Dedicated mode LPAR4 is using one physical processor and thus three processors are available for the shared processor pool. LPAR1 defines five virtual processors and the simultaneous multithreading feature is on (and thus sees 10 logical processors), LPAR2 defines one virtual processor and simultaneous multithreading is off (one logical processor). LPAR3 defines two virtual processors and simultaneous multithreading is on. Currently (sample time), virtual processors 2 and 3 of LPAR1 and virtual processor 0 of LPAR2 are dispatched on physical processors in the shared pool. Other virtual processors are idle and waiting for dispatch by the Hypervisor. When more virtual processors are defined within a partition, any virtual processor shares equal parts of the partition processing entitlement.

2.12.3 Virtual I/O Server

The Virtual I/O Server is a special purpose partition that provides virtual I/O resources to other partitions. The Virtual I/O Server owns the physical resources (SCSI, Fibre Channel, and network adapters, and optical devices) and allows client partitions to share access to them, thus minimizing the number of physical adapters in the system. The Virtual I/O Server eliminates the requirement that every partition own a dedicated network adapter, disk adapter, and disk drive.

Figure 2-15 shows an organization view of a micro-partitioned system, including the Virtual I/O Server. The figure also includes virtual SCSI and Ethernet connections and mixed operating system partitions.

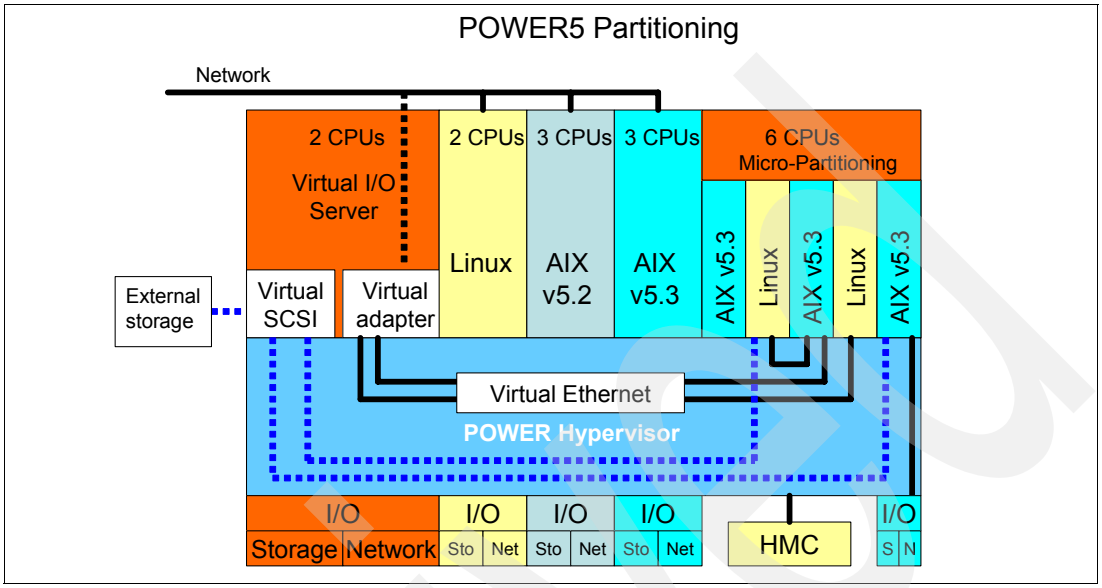


Figure 2-15 Micro-Partitioning technology and VIOS

Because the Virtual I/O server is an operating system-based appliance server, redundancy for physical devices attached to the Virtual I/O Server can be provided by using capabilities such as Multipath I/O and IEEE 802.3ad Link Aggregation.

Installation of the Virtual I/O Server partition is performed from a special system backup DVD that is provided to clients that order the Advanced POWER Virtualization feature. This dedicated software is only for the Virtual I/O Server (and IVM, in case it is used) and is only supported in special Virtual I/O Server partitions.

The Virtual I/O Server can be installed by:

- Media (assigning the DVD-ROM drive to the partition and booting from the media)
- The HMC (inserting the media in the DVD-ROM drive on the HMC and using the **`installios`** command)
- Using the Network Install Manager (NIM)

Note: To increase the performance of I/O-intensive applications, use dedicated physical adapters that use dedicated partitions.

We recommend that you install the Virtual I/O Server in a partition with dedicated resources or at least 0.5 processor entitlement to help ensure consistent performance.

The Virtual I/O Server supports RAID configurations and SAN attached devices (possibly with multipath driver). Logical volumes created on RAID or JBOD configurations are bootable, and the number of logical volumes is limited to the amount of storage available and architectural limits of the Logical Volume Manager.

Two major functions are provided with the Virtual I/O Server: a shared Ethernet adapter and Virtual SCSI.

Shared Ethernet adapter

A shared Ethernet adapter (SEA) is a Virtual I/O Server service that acts as a layer 2 network bridge between a physical Ethernet adapter or aggregation of physical adapters (EtherChannel) and one or more virtual Ethernet adapters defined by Hypervisor on the Virtual I/O Server. A SEA enables LPARs on the virtual Ethernet to share access to the physical Ethernet and communicate with stand-alone servers and LPARs on other systems. The shared Ethernet network provides this access by connecting the internal Hypervisor VLANs with the VLANs on the external switches. Because the shared Ethernet network processes packets at layer 2, the original MAC address and VLAN tags of the packet is visible to other systems on the physical network. IEEE 802.1 VLAN tagging is supported.

The virtual Ethernet adapters that are used to configure a shared Ethernet adapter are required to have the trunk setting enabled. The trunk setting causes these virtual Ethernet adapters to operate in a special mode so that they can deliver and accept external packets from the POWER5 internal switch to the external physical switches. The trunk setting should only be used for the virtual Ethernet adapters that are part of a shared Ethernet setup in the Virtual I/O Server.

A single SEA setup can have up to 16 Virtual Ethernet trunk adapters and each virtual Ethernet trunk adapter can support up to 20 VLAN networks. Therefore, it is possible for a single physical Ethernet to be shared between 320 internal VLAN networks. The number of shared Ethernet adapters that can be set up in a Virtual I/O server partition is limited only by the resource availability, as there are no configuration limits.

For a more detailed discussion about virtual networking, see:

`http://www.ibm.com/servers/aix/whitepapers/aix_vn.pdf`

Virtual SCSI

Access to real storage devices is implemented through the virtual SCSI services, a part of the Virtual I/O Server partition. This is accomplished using a pair of virtual adapters: a virtual SCSI server adapter and a virtual SCSI client adapter. The virtual SCSI server and client adapters are configured using an HMC or through Integrated Virtualization Manager on smaller systems. The virtual SCSI server (target) adapter is responsible for executing any SCSI commands it receives. It is owned by the Virtual I/O Server partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN attached devices and LUNs that are assigned to the client partition.

Physical disks owned by the Virtual I/O Server partition can either be exported and assigned to a client partition as a whole device, or can be configured into a volume group and partitioned into several logical volumes. These logical volumes can then be assigned to individual partitions. From a client partition point of view, these two options are equivalent.

The Virtual I/O Server provides mapping between *backing devices* (physical devices or logical volumes assigned to client partitions in VIOS nomenclature) and client partitions by a command-line interface. The appropriate command is the **`mkvdev`** command. For syntax and semantics, see the Virtual I/O server documentation.

All current storage device types, such as SAN, SCSI, and RAID, are supported; SSA and iSCSI are not supported at the time of writing.

For more information about the specific storage devices supported, see:

`http://techsupport.services.ibm.com/server/vios/home.html`

Note: Mirrored Logical Volumes (LVs) on Virtual I/O Server level are not recommended as backing devices. If mirroring is required, two independent devices (possibly from two separate VIO servers) should be assigned to the client partition, and the client partition should define a mirror on top of them.

Virtual I/O Server Version 1.3

Virtual I/O Server Version 1.3 brings a host of new enhancements, including improved monitoring, such as additional **`topas`** and **`viostat`** performance metrics, and the bundling of the Performance ToolKit (PTX®) agent. Virtual SCSI and Virtual Ethernet performance increases, and command-line enhancements and enablement of additional storage solutions are also included.

Virtual I/O Server V1.3 introduced several enhancements for Virtual SCSI and shared Fibre Channel adapters support:

- Independent Software Vendor / Independent Hardware Vendor Virtual I/O enablement
- iSCSI TOE adapter
- iSCSI direct attached n3700 storage subsystem
- HP storage
- Virtual SCSI functional enhancements
  - Support for SCSI Reserve/Release for limited configurations
  - Changeable queue depth
  - Updating virtual device capacity non disruptively so that the virtual disk can "grow" without requiring a reconfiguration
  - Configurable fast fail time (number of retries on failure)
  - Error log enhancements

Virtual I/O Server V1.3 also introduced several enhancements for Virtual Ethernet and shared Ethernet adapter support.

2.12.4 Partition Load Manager

Partition Load Manager (PLM) provides automated processor and memory distribution between a dynamic LPAR and a Micro-Partitioning technology-capable logical partition running AIX 5L. The PLM application is based on a client/server model to share system information, such as processor or memory events, across the concurrent present logical partitions.

The following events are registered on all managed partition nodes:

- Memory-pages-steal high thresholds and low thresholds
- Memory-usage high thresholds and low thresholds
- Processor-load-average high threshold and low threshold

Note: PLM is supported on AIX 5L Version 5.2 and AIX 5L Version 5.3; it is not supported on Linux.

2.12.5 Operating system support for advanced virtualization

Table 2-10 lists AIX 5L and Linux support for advanced virtualization.

Table 2-10 Operating system supported functions

Advanced virtualization feature	AIX 5L Version 5.2	AIX 5L Version 5.3	Linux SLES 9	Linux RHEL AS 3	Linux RHEL AS 4
Micro-partitions (1/10th of processor)	N	Y	Y	Y	Y
Virtual Storage	N	Y	Y	Y	Y
Virtual Ethernet	N	Y	Y	Y	Y
Partition Load Manager	Y	Y	N	N	N

2.13 Hardware Management Console

The HMC is a dedicated workstation that provides a graphical user interface for configuring, operating, and performing basic system tasks for the System p5 servers functioning in either non-partitioned, LPAR, or clustered environments. In addition, the Hardware Management Console is used to configure and manage partitions. One HMC is capable of controlling multiple POWER5 and POWER5+ processor-based systems.

At the time of writing, one HMC supports up to 48 POWER5 and POWER5+ processor-based systems and up to 254 LPARs using the HMC machine code Version 5.1. For updates of the machine code and HMC functions and hardware prerequisites, refer to the following Web site:

http://techsupport.services.ibm.com/server/hmc

POWER5 and POWER5+ processor-based system HMCs require Ethernet connectivity between HMC and server's service processor; moreover, if dynamic LPAR operations are required, all AIX 5L and Linux partitions must be enabled to communicate over network to HMC. Ensure that sufficient Ethernet adapters are available to enable public and private networks, if you need both:

- The HMC 7310 Model C04 is a desktop model with only one integrated 10/100/1000 Mbps Ethernet port, but two additional PCI slots.
- The HMC 7310 Model C05 is a deskside model with only one integrated 10/100/1000 Mbps Ethernet port, but two additional PCI slots.
- The 7310 Model CR3 is a 1U, 19-inch rack-mountable drawer that has two native 10/100/1000 Mbps Ethernet ports and two additional PCI slots.

For any partition in a server, it is possible to use the shared Ethernet adapter in Virtual I/O Server for a unique connection from the HMC to the partitions. Therefore, client partitions do not require their own physical adapter in order to be able to communicate to the HMC.

It is a good practice to connect the HMC to the first HMC port on the system, labeled as HMC Port 1, although other network configurations are possible. A second HMC can be attached to HMC Port 2 of the server for redundancy (or vice versa). Figure 2-16 shows a simple network configuration to enable the connection from HMC to server, and to enable dynamic LPAR operations. For more details about the HMC and the possible network connections, refer to:

http://www.redbooks.ibm.com/abstracts/redp3999.html

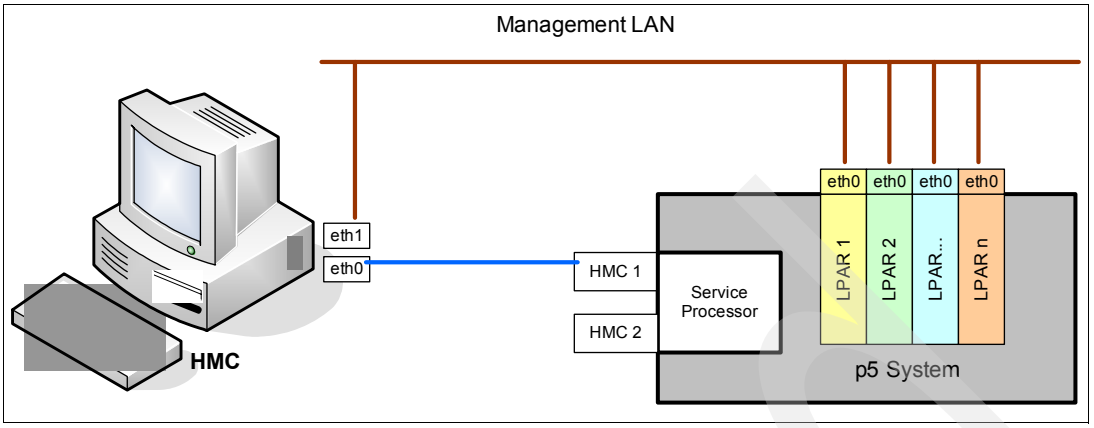


Figure 2-16 HMC to service processor and LPARs network connection

The default mechanism for allocation of the IP addresses for the service processor HMC ports is dynamic. The HMC can be configured as a DHCP server, providing the IP address at the time the managed server is powered on. If the service processor of the managed server does not receive a DHCP reply before timeout, predefined IP addresses will be set up on both ports. Static IP address allocation is also an option. You can configure the IP address of the service processor ports with a static IP address by using the Advanced System Management Interface (ASMI) menus. See 2.15.4, “Service processor” on page 62 for predefined IP addresses and additional information.

Note: If you need to access ASMI (for example, to set up an IP address of a new POWER5+ processor-based server when the HMC is not available or not providing DHCP services), you can connect any client to one of the service processor HMC ports with any kind of Ethernet cable, and use a Web browser to access the predefined IP address, such as the following example:

`https://192.168.2.147`

Functions performed by the HMC include:

- Creating and maintaining a multiple partition environment
- Displaying a virtual operating system session terminal for each partition
- Displaying a virtual operator panel of contents for each partition
- Detecting, reporting, and storing changes in hardware conditions
- Powering managed systems on and off
- Acting as a service focal point

The HMC provides both a graphical and command-line interface for all management tasks. Remote connection to the HMC using Web-based System Manager or SSH are possible. For accessing the graphical interface, you can use the Web-based System Manager Remote Client running on the AIX 5L, Linux, or Windows® operating system. The Web-based System Manager client installation image can be downloaded from the HMC itself from the following URL:

`http://<hmc_address_or_name>/remote_client.html`

Both un-encrypted and encrypted Web-based System Manager connections are supported. The command line interface is also available by using the SSH secure shell connection to the

HMC. It can be used by an external management system or a partition to perform HMC operations remotely.

2.13.1 High availability using the HMC

The HMC is an important hardware component. HACMP Version 5.3 High Availability cluster software can be used to automatically activate resources (where available), thus becoming an integral part of the cluster. For some environments, We recommend working with redundant HMCs. POWER5 and POWER5+ processor-based systems have two service processor interfaces (HMC port 1 and HMC port 2) available for connection to the HMC. We recommend using both of them for redundant network configuration. Depending on your environment, you have multiple options to configure the network. Figure 2-17 on page 52 shows one possible highly available configuration. Note the planning requirements for additional network hardware, such as switches or hubs.

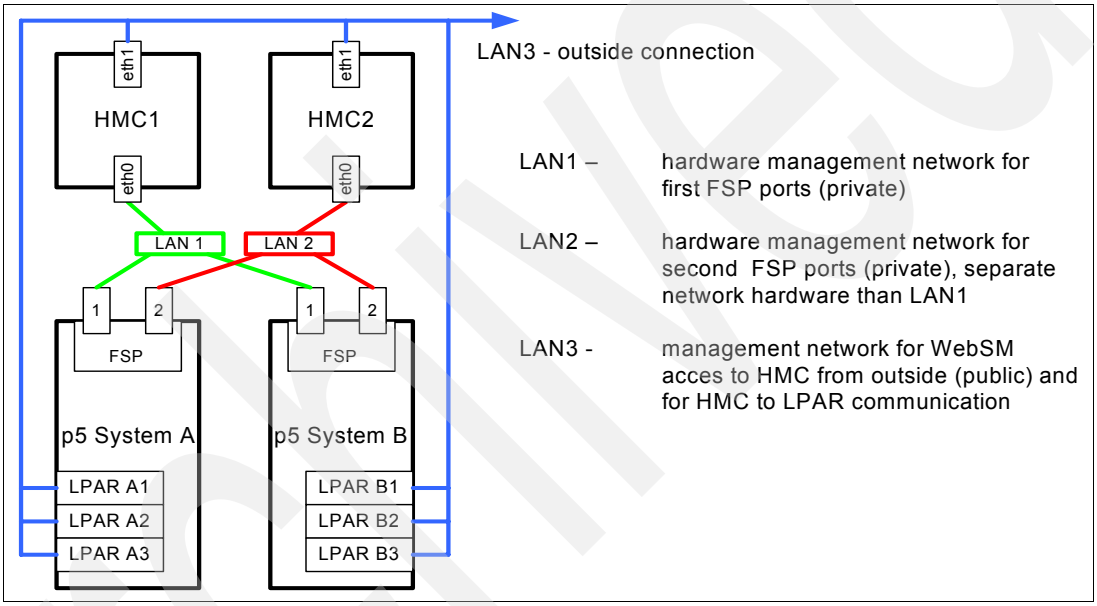


Figure 2-17 Highly available HMC and network architecture

Note that only hardware management network (LAN1 and LAN2) is highly available in Figure 2-17 in order to keep things simple. But the management network (LAN3) can also be made highly available by using a similar concept and adding more Ethernet adapters to LPARs and HMCs.

2.13.2 IBM System Planning Tool

The IBM System Planning Tool (SPT) is the next generation of the IBM LPAR Validation Tool (LVT). It contains all of the functions from the LVT and is integrated with the IBM Systems Workload Estimator (WLE). System plans generated by the SPT can be deployed on the system by the Hardware Management Console (HMC). The SPT is available to assist the user in system planning, design, validation, and to provide a system validation report that reflects the user's system requirements while not exceeding system recommendations. The SPT is a PC based browser application designed to be run in a stand-alone environment.

The IBM System Planning Tool can be downloaded at no additional charge from:

http://www.ibm.com/servers/eserver/support/tools/systemplanningtool/

The System Planning Tool (SPT) helps you design a system to fit your needs. You can use the SPT to design a logically partitioned system or you can use the SPT to design an unpartitioned system. You can create an entirely new system configuration, or you can create a system configuration based upon any of the following:

- Performance data from an existing system that the new system is to replace
- Performance estimates that anticipates future workloads that you must support
- Sample systems that you can customize to fit your needs

Integration between the SPT and both the Workload Estimator (WLE) and IBM Performance Management (PM) allows you to create a system that is based upon performance and capacity data from an existing system or that is based on new workloads that you specify.

You can use the SPT before you order a system to determine what you must order to support your workload. You can also use the SPT to determine how you can partition a system that you already have.

Important: We recommend using the IBM System Planning Tool to estimate Hypervisor requirements and determine the memory resources that are required for all partitioned and non-partitioned servers.

Figure 2-18 shows the estimated Hypervisor memory requirements based on sample partition requirements.

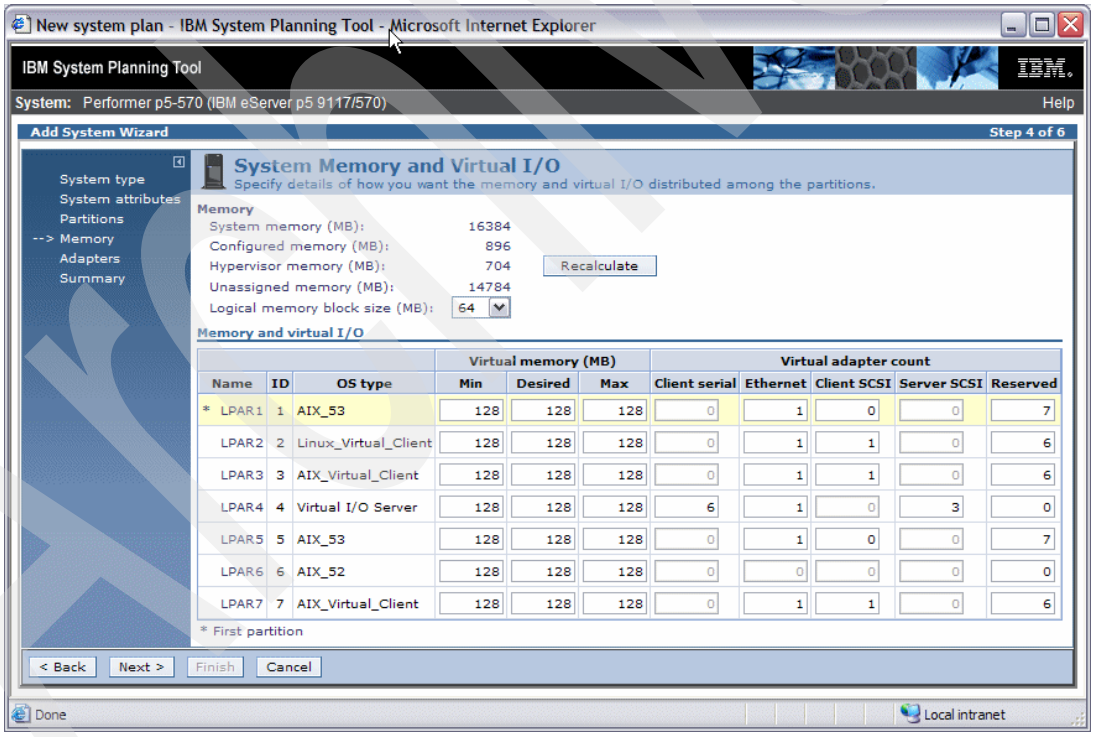


Figure 2-18 IBM System Planning Tool window showing Hypervisor requirements

2.14 Operating system support

The p5-570 is capable of running the AIX 5L and Linux operating systems. The AIX 5L operating system has been specifically developed and enhanced to exploit and support the extensive RAS features on IBM System p systems.

2.14.1 AIX 5L

If installing AIX 5L on the p5-570, the following minimum requirements must be met:

- AIX 5L for POWER V5.2 with 5200-08 Technology Level (APAR IY77270), CD# LCD4-1133-08 or later (DVD LCD4-7549-01 media is also available.)
- AIX 5L for POWER V5.3 with 5300-04 Technology Level (APAR IY77273), CD#LCD4-7463-05 or later (DVD LCD4-7544-01 media is also available.)

Note: The Advanced POWER Virtualization feature (FC 7942) is not supported on AIX 5L V5.2; it requires AIX 5L V5.3.

IBM releases maintenance packages for the AIX 5L operating system periodically. These packages are available on CD-ROM, or you can download them from:

http://www.ibm.com/servers/eserver/support/pseries/aixfixes.html

The Web page provides information about how to obtain the CD-ROM.

You can also get individual operating system fixes and information about obtaining AIX 5L service at this site. In AIX 5L V5.3, the **suma** command is also available, which helps the administrator to automate the task of checking and downloading operating system downloads. For more information about the **suma** command functionality, refer to:

http://techsupport.services.ibm.com/server/suma/home.html

On 18 November 2005, Electronic Software Delivery (ESD) for AIX 5L V5.2 and V5.3 for POWER5 systems was made available. This delivery method is a way for clients to receive software and associated publications online, instead of waiting for a physical shipment to arrive. Clients requesting ESD should order FC 3450.

ESD has the following requirements:

- POWER5 system
- Internet connectivity from a POWER5 system or PC, an a reasonable connection speed for downloading large products, such as AIX 5L
- Registration on the ESD Web site

For additional information, contact your IBM sales representative.

Software support for new features in the POWER5+ processor

For a complete list of new features introduced in the POWER5+ processor, see 2.1, “The POWER5+ processor” on page 20. Support for two new virtual memory page sizes was introduced: 64 KB and 16 GB, as well as support for the 1 TB segment size. While 16 GB pages are intended to only be used in very high performance environments, 64 KB pages are general purpose. AIX 5L Version 5.3 with the 5300-04 Technology Level 64-bit kernel is required for 64 KB and 16 GB page size support.

As with all previous versions of AIX 5L, 4 KB is the default page size. A process will continue to use 4 KB pages unless a user specifically requests another page size be used. AIX 5L has

rich support of 64 KB pages. They are easy to use, and it is expected that many applications will see performance benefits when using 64 KB pages rather than 4 KB pages. No system configuration changes are necessary to enable a system to use 64 KB pages, they are fully pageable, and the size of the pool of 64 KB page frames on a system is dynamic and fully managed by AIX 5L.

The main benefit of a larger page size is improved performance for applications that allocate and repeatedly access large amounts of memory. The performance improvement from larger page sizes is due to the overhead of translating a page address as it is used in an application, to a page address that is understood by the computer's memory subsystem. To improve performance, the information needed to translate a given page is usually cached in the processor. In POWER5+, this cache takes the form of a translation lookaside buffer (TLB). Since there are a limited number of TLB entries, using a large page size increases the amount of address space that can be accessed without incurring translation delays. Also, the size of TLB in POWER5+ has been doubled compared to POWER5.

Huge pages (16 GB) are intended to be used only in very high performance environments, and AIX 5L will not automatically configure a system to use these page sizes. A system administrator must configure AIX 5L to use these page sizes and specify their number via HMC before partition start.

A user can specify page sizes to use for three regions process's address space with an environment variable or with settings in an application's XCOFF binary using the **ldedit** or **ld** commands. These three regions are: data, stack and program text. An application programmer can also select the page size to use for System V shared memory via a new SHM_PAGESIZE command to the shmctl() system call.

Here is an example of using system variables to start a program with 64 KB page size support:

```
LDR_CNTRL=DATAPSIZE=64K@TEXTPSIZE=64K@STACKPSIZE=64K <program>
```

System commands (**ps**, **vmstat**, **svmon**, and **pagesize**) have been enhanced to report various page size usage.

2.14.2 Linux

For the p5-570, Linux distributions are available through Novel SUSE and Red Hat at the time this Redpaper was written. The p5-570 requires the following version of Linux distributions:

- SUSE Linux Enterprise Server 9 for IBM POWER Service Pack 3 or later
- Red Hat Enterprise Linux AS 4 for IBM POWER Service Update 2 or later

Note: Not all p5-570 features available on AIX 5L are available on Linux.

For information about the features and external devices supported by Linux, refer to:

http://www.ibm.com/servers/eserver/pseries/linux/

For information about SUSE Linux Enterprise Server 9, refer to:

http://www.novell.com/products/linuxenterpriseserver/

For information about Red Hat Enterprise Linux AS, refer to:

http://www.redhat.com/software/rhel/details/

Many of the features described in this Redpaper are operating system dependant and might not be available on Linux. For more information, see:

http://www.ibm.com/servers/eserver/linux/power/whitepapers/linux_overview.html

Note: IBM only supports the Linux systems of clients with a SupportLine contract covering Linux. Otherwise, contact the Linux distributor for support.

Specially-priced Linux subscriptions

Linux subscriptions are now available when ordered through IBM and combined with an IBM System p5 Express Edition. Clients can purchase a one-year specially priced subscription or a greater discount for a three-year subscription.

These new Linux options, available on System p5 Express servers, bring improved pricing and price performance to our clients interested in Linux as their primary operating system. Clients interested in AIX 5L can also obtain an Express Edition that fits their needs.

Clients are still encouraged to purchase support for their Linux subscription either though IBM Global Services or through the distributor to receive updates and technical assistance as needed. Support is not included in the price of the subscription.

The new lower-priced Linux subscriptions, when combined with the lower package prices of the System p5 Express Edition, make these products an exceptional value for our smaller to mid-market clients, as well as larger enterprises.

Refer to the following Web site for Red Hat information:

http://www.redhat.com/software/

For additional information about Linux on POWER, visit:

http://www.ibm.com/servers/eserver/linux/power/

2.14.3 i5/OS

The i5/OS® operating system is supported in an IBM System p5 570 with a single POWER5+ 2.2 GHz processor dedicated to a logical partition.

If installing the i5/OS in this partition, i5/OS V5R3 or i5/OS V5R4 is required.

2.15 Service information

The p5-570 is not a client setup server (CSU). Therefore, the IBM service representative completes the system installation.

2.15.1 Touch point colors

Blue (IBM blue) or terra-cotta (orange) on a component indicates a touch point (for electronic parts) where you can grip the hardware to remove it from or to install it into the system, to open or to close a latch, and so on. IBM defines the touch point colors as follows:

Blue This requires a shutdown of the system before the task can be performed, for example, installing additional processors contained in the second processor book.

Terra-cotta	The system can remain powered on while this task is being performed. Keep in mind that some tasks might require that other steps have to be performed first. One example is deconfiguring a physical volume in the operating system before removing a disk from a 4-pack disk enclosure of the server.
Blue and terra-cotta	Terra-cotta takes precedence over this color combination, and the rules for a terra-cotta-only touch point apply.

Important: It is important to adhere to the touch point colors on the system. Not doing so can compromise your safety and damage the system.

2.15.2 Operator control panel

The service processor provides an interface to the control panel that is used to display server status and diagnostic information. See Figure 2-19 on page 57 for operator control panel physical details and buttons.

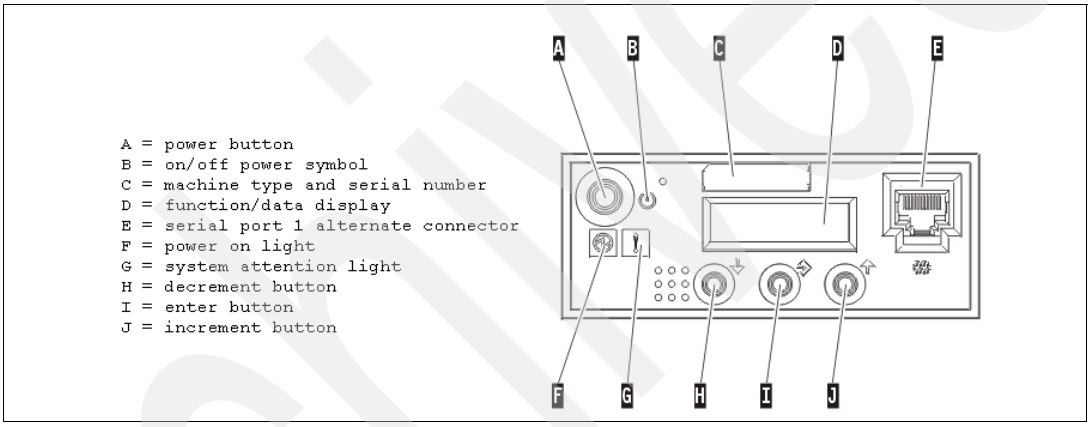


Figure 2-19 Operator control panel physical details and button

Note: For servers managed by the HMC, use it to perform control panel functions.

Primary control panel functions

The primary control panel functions are defined as functions 01 to 20, including options to view and manipulate IPL modes, server operating modes, IPL speed, and IPL type.

The following list describes the primary functions:

- Function 01: Display the selected IPL type, system operating mode, and IPL speed
- Function 02: Select the IPL type, IPL speed override, and system operating mode
- Function 03: Start IPL
- Function 04: Lamp Test
- Function 05: Reserved
- Function 06: Reserved
- Function 07: SPCN functions
- Function 08: Fast Power Off
- Functions 09 to 10: Reserved
- Functions 11 to 19: System Reference Code
- Function 20: System type, model, feature code, and IPL type

All the functions mentioned are accessible using the Advanced System Management Interface (ASMI), HMC, or the control panel.

Extended control panel functions

The extended control panel functions consist of two major groups:

- Functions 21 through 49, which are available when you select Manual mode from Function 02.
- Support service representative Functions 50 through 99, which are available when you select Manual mode from Function 02, then select and enter the client service switch 1 (Function 25), followed by service switch 2 (Function 26).

Function 30 – CEC SP IP address and location

Function 30 is one of the Extended control panel functions and is only available when Manual mode is selected. This function can be used to display the central electronic complex (CEC) Service Processor IP address and location segment. The Table 2-11 shows an example of how to use the Function 03.

Table 2-11 CEC SP IP address and location

Information on operator panel	Action or description
3 0	Use the increment or decrement buttons to scroll to Function 30.
3 0 * *	Press Enter to enter sub-function mode.
3 0 0 0	Use the increment or decrement buttons to select an IP address: 0 0 = Service Processor ETH0 or HMC1 port 0 1 = Service Processor ETH1 or HMC2 port
S P A : E T H 0 : _ _ _ T 5 1 9 2 . 1 6 8 . 2 . 1 4 7	Press Enter to display the selected IP address.
3 0 * *	Use the increment or decrement buttons to select sub-function exit.
3 0	Press Enter to exit sub-function mode.

2.15.3 System firmware

Server firmware is the part of the Licensed Internal Code that enables hardware, such as the service processor. Depending on your service environment, you can download, install, and manage your server firmware fixes using different interfaces and methods, including the HMC, or by using functions specific to your operating system. See 3.2.4, “IBM System p5 firmware maintenance” on page 77 for a detailed description of System p5 firmware.

Note: Normally, installing the server firmware fixes through the operating system is a nonconcurrent process.

Temporary and permanent firmware sides

The service processor maintains two copies of the server firmware:

- One copy is considered the permanent or backup copy and is stored on the permanent side, sometimes referred to as the *p* side.
- The other copy is considered the installed or temporary copy and is stored on the temporary side, sometimes referred to as the *t* side. We recommend that you start and run the server from the temporary side.

The copy actually booted from is called the activated level, sometimes referred to as *b*.

Note: The default value, from which the system boots, is temporary.

The following examples are the output of the **lsmcode** command for AIX 5L and Linux, showing the firmware levels as they are displayed in the outputs.

- AIX 5L:

```
The current permanent system firmware image is SF220_005.
The current temporary system firmware image is SF220_006.
The system is currently booted from the temporary image.
```

- Linux:

```
system:SF220_006 (t) SF220_005 (p) SF220_006 (b)
```

When you install a server firmware fix, it is installed on the temporary side.

Note: The following points are of special interest:

- The server firmware fix is installed on the temporary side only after the existing contents of the temporary side are permanently installed on the permanent side (the service processor performs this process automatically when you install a server firmware fix).
- If you want to preserve the contents of the permanent side, you need to remove the current level of firmware (copy the contents of the permanent side to the temporary side) before you install the fix.
- However, if you get your fixes using the Advanced features on the HMC interface and you indicate that you do not want the service processor to automatically accept the firmware level, the contents of the temporary side are not automatically installed on the permanent side. In this situation, you do not need to remove the current level of firmware to preserve the contents of the permanent side before you install the fix.

You might want to use the new level of firmware for a period of time to verify that it works correctly. When you are sure that the new level of firmware works correctly, you can permanently install the server firmware fix. When you permanently install a server firmware fix, you copy the temporary firmware level from the temporary side to the permanent side.

Conversely, if you decide that you do not want to keep the new level of server firmware, you can remove the current level of firmware. When you remove the current level of firmware, you copy the firmware level that is currently installed on the permanent side from the permanent side to the temporary side.

System firmware download site

For the system firmware download site for the p5-570, go to:

http://techsupport.services.ibm.com/server/mdownload

Receive server firmware fixes using an HMC

If you use an HMC to manage your server and you periodically configure several partitions on the server, you need to download and install fixes for your server and power subsystem firmware.

How you get the fix depends on whether the HMC or server is connected to the Internet:

- The HMC or server is connected to the Internet.

  There are several repository locations from which you can download the fixes using the HMC. For example, you can download the fixes from your service provider's Web site or support system, from optical media that you order from your service provider, or from an FTP server on which you previously placed the fixes.

- Neither the HMC nor your server is connected to the Internet (server firmware only).

  You need to download your new server firmware level to a CD-ROM media or FTP server.

For both of these options, you can use the interface on the HMC to install the firmware fix (from one of the repository locations or from the optical media). The Change Internal Code wizard on the HMC provides a step-by-step process for you to perform the procedure to install the fix. Perform these steps:

1. Ensure that you have a connection to the service provider (if you have an Internet connection from the HMC or server).
2. Determine the available levels of server and power subsystem firmware.
3. Create the optical media (if you do not have an Internet connection from the HMC or server).
4. Use the Change Internal Code wizard to update your server and power subsystem firmware.
5. Verify that the fix installed successfully.

For a detailed description of each task, select **Customer service, support, and troubleshooting** → **Fixes and upgrades** → **Getting fixes and upgrades** from the IBM Systems Hardware Information Center Web site at:

`http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp?lang=en`

Receive server firmware fixes without an HMC

Periodically, you need to install fixes for your server firmware. If you do not use an HMC to manage your server, you must get your fixes through your operating system. In this situation, you can get server firmware fixes through the operating system regardless of whether your operating system is AIX 5L or Linux.

To do this, complete the following tasks:

1. Determine the existing level of server firmware using the **`lsmcode`** command.
2. Determine the available levels of server firmware.
3. Get the server firmware.
4. Install the server firmware fix to the temporary side.
5. Verify that the server firmware fix installed successfully.
6. Install the server firmware fix permanently (optional).

Note: To view existing levels of server firmware using the **lsmcode** command, you need to have the following service tools installed on your server:

- AIX 5L

  You must have AIX 5L diagnostics installed on your server to perform this task. AIX 5L diagnostics are installed when you install AIX 5L on your server. However, it is possible to deselect the diagnostics. Therefore, you need to ensure that the online AIX 5L diagnostics are installed before proceeding with this task.

- Linux
  - Platform Enablement Library: librtas-*nnnnn*.rpm
  - Service Aids: ppc64-utils-*nnnnn*.rpm
  - Hardware Inventory: lsvpd-*nnnnn*.rpm

  Where *nnnnn* represents a specific version of the RPM file.

  If you do not have the service tools on your server, you can download them at the following Web site:

  http://techsupport.services.ibm.com/server/lopdiags

See 3.2.4, “IBM System p5 firmware maintenance” on page 77 for additional information.

2.15.4 Service processor

The service processor is an embedded controller running the service processor internal operating system. The service processor operating system contains specific programs and device drivers for the service processor hardware. The host interface is a 32-bit PCI-X interface connected to the Enhanced I/O Controller.

Service processor is used to monitor and manage the system hardware resources and devices. The service processor offers the following connections:

Two Ethernet 10/100 Mbps ports

- Both Ethernet ports are only visible to the service processor and can be used to attach the p5-570 to a HMC or to access the Advanced System Management Interface (ASMI) options from a client Web browser, using the HTTP server integrated into the service processor internal operating system.
- Both Ethernet ports have a default IP address:
  - Service processor Eth0 or HMC1 port is configured as 192.168.2.147.
  - Service processor Eth1 or HMC2 port is configured as 192.168.3.147.

2.15.5 Redundant service processor

Redundant service processors are available on the p5-570, and provide the continuation of services processor functions in the event one service processor becomes unreachable. This function requires the system to be shut down, if the failure is determined to be the SP itself, to service the hardware to restore redundancy. The failover itself is nondisruptive.

The redundant FSP feature is comprised of two FC 7997. The following minimum requirements apply:

- The SP level must be 03N6355 or higher. (FRU’s 80P6027, 80P5319, and 80P5560 are *not* supported.)

- The firmware and hardware levels must be at least:
  - 01SF240_ 201 _201 with FC 8338 or FC 7782 for POWER5+ processor-based systems.
- The HMC must be at the following levels: Version 5, Release 1 with PTF MH000607.
- There are two building block (CEC) enclosures.

Note: The firmware and HMC levels listed are strictly minimums. Installation of the latest available Service Packs for your code stream is strongly recommended, as they contain fixes and enhancements that will provide more robust performance and improve reliability.

Both service processors are capable of providing all the functions, with one being the primary (master) and the other the secondary (backup).

The configuration for this requires unique connections (interconnecting and network switches) to provide communication between SPs and also connections to each of the HMCs in a redundant HMC environment. See the online product documentation for additional information, starting with the following resource:

http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp?topic=/iphae/planredundantfsp.htm

2.15.6 Hardware management user interfaces

This section provides a brief overview of the different p5-570 hardware management user interfaces available.

Advanced System Management Interface

The Advanced System Management Interface (ASMI) is the interface to the service processor that enables you to set flags that affect the operation of the server, such as auto power restart, and to view information about the server, such as the error log and vital product data.

This interface is accessible using a Web browser on a client system that is connected directly to the service processor (in this case, a standard Ethernet cable or a crossed cable can be both used) or through an Ethernet network. Using the *network configuration menu*, the ASMI enables the ability to change the service processor IP addresses or to apply some security policies and avoid the access from undesired IP addresses or range. The ASMI can also be accessed using a terminal attached to the system service processor ports on the server, if the server is not HMC managed. The service processor and the ASMI are standard on all IBM System p servers.

You might be able to use the service processor's default settings. In that case, accessing the ASMI is not necessary.

Accessing the ASMI using a Web browser

The Web interface to the Advanced System Management Interface is accessible through, at the time of writing, Microsoft® Internet Explorer® 6.0, Netscape 7.1, Mozilla Firefox, or Opera 7.23 running on a PC or mobile computer connected to the service processor. The Web interface is available during all phases of system operation, including the initial program load and runtime. However, some of the menu options in the Web interface are unavailable during IPL or runtime to prevent usage or ownership conflicts if the system resources are in use during that phase.

Accessing the ASMI using an ASCII console

The Advanced System Management Interface on an ASCII console supports a subset of the functions provided by the Web interface and is available only when the system is in the platform standby state. The ASMI on an ASCII console is not available during some phases of system operation, such as the initial program load and runtime.

Accessing the ASMI using an HMC

To access the Advanced System Management Interface using the Hardware Management Console, complete the following steps:

1. Ensure that the HMC is set up and configured.
2. In the navigation area, expand the managed system with which you want to work.
3. Expand **Service Applications** and click **Service Focal Point**.
4. In the content area, click **Service Utilities**.
5. From the Service Utilities window, select the managed system with which you want to work with.
6. From the Selected menu on the Service Utilities window, select **Launch ASM menu**.

System Management Services

Use the System Management Services (SMS) menus to view information about your system or partition and to perform tasks, such as changing the boot list or setting the network parameters.

To start System Management Services, perform the following steps:

1. For a server that is connected to an HMC, use the HMC to restart the server or partition.

   If the server is not connected to an HMC, stop the system, and then restart the server by pressing the power button on the control panel.
2. For a partitioned server, watch the virtual terminal window on the HMC.

   For a full server partition, watch the firmware console.
3. Look for the POST[5] indicators (memory, keyboard, network, SCSI, and speaker) that appear across the bottom of the screen. Press the numeric 1 key after the word `keyboard` appears and before the word `speaker` appears.

The SMS menus is useful to defining the operating system installation method, choosing the installation boot device, or setting the boot device priority list for a full managed server or a logical partition. In the case of a network boot, SMS menus are provided to set up the network parameters and network adapter IP address.

HMC

The Hardware Management Console is a system that controls managed systems, including IBM System p5 hardware, and logical partitions. To provide flexibility and availability, there are different ways to implement HMCs.

Web-based System Manager Remote Client

The Web-based System Manager Remote Client is an application that is usually installed on a PC and can be downloaded directly from an installed HMC. When an HMC is installed and HMC Ethernet IP addresses have been assigned, it is possible to download the Web-based System Manager Remote Client from a web browser, using the following URL:

```
http://HMC_IP_address/remote_client.html
```

[5] POST stands for power-on-self-test.

You can then use the PC to access other HMCs remotely. Web-based System Manager Remote Clients can be present in private and open networks. You can perform most management tasks using the Web-based System Manager Remote Client. The remote HMC and the Web-based System Manager Remote Client allow you the flexibility to access your managed systems (including HMCs) from multiple locations using multiple HMCs.

For more detailed information about the use of the HMC, refer to the IBM Systems Hardware Information Center.

Open Firmware

A System p5 server has one instance of Open Firmware both when in the partitioned environment and when running as a full system partition. Open Firmware has access to all devices and data in the server. Open Firmware is started when the server goes through a power-on reset. Open Firmware, which runs in addition to the POWER Hypervisor in a partitioned environment, runs in two modes: global and partition. Each mode of Open Firmware shares the same firmware binary that is stored in the flash memory.

In a partitioned environment, Open Firmware runs on top of the global Open Firmware instance. The partition Open Firmware is started when a partition is activated. Each partition has its own instance of Open Firmware and has access to all the devices assigned to that partition. However, each instance of Open Firmware has no access to devices outside of the partition in which it runs. Partition firmware resides within the partition memory and is replaced when AIX 5L or Linux takes control. Partition firmware is needed only for the time that is necessary to load AIX 5L or Linux into the partition server memory.

The global Open Firmware environment includes the partition manager component. That component is an application in the global Open Firmware that establishes partitions and their corresponding resources (such as CPU, memory, and I/O slots), which are defined in partition profiles. The partition manager manages the operational partitioning transactions. It responds to commands from the service processor external command interface that originates in the application running on the HMC. The ASMI can be accessed during boot time or using the ASMI and selecting the boot to Open Firmware prompt.

For more information about Open Firmware, refer to *Partitioning Implementations for IBM* eServer*Partitioning Implementations for IBM @server p5 Servers*, SG24-7039, which is available at:

http://www.redbooks.ibm.com/abstracts/sg247039.html

3

RAS and manageability

This chapter provides information about IBM System p5 design features that help lower the total cost of ownership (TCO). The state of art IBM RAS (Reliability, Availability, and Service ability) technology allows the possibility to improve your TCO architecture by reducing unplanned down time. This chapter includes several features that are based on the benefits available when using AIX 5L. Support of these features when using Linux can vary.

3.1 Reliability, availability, and serviceability

Excellent quality and reliability are inherent in all aspects of the IBM System p5 design and manufacturing. The fundamental objective of the design approach is to minimize outages. The RAS features help to ensure that the system operates when required, performs reliably, and efficiently handles any failures that might occur. This is achieved using capabilities provided by both the hardware and the operating system AIX 5L.

The p5-570 as a POWER5+ processor-based server enhances the RAS capabilities implemented in POWER4 processor-based servers. RAS enhancements available are:

- Most firmware updates allow the system to remain operational.
- The ECC has been extended to inter-chip connections for the fabric and processor bus.
- Partial L2 cache deallocation is possible.
- The number of L3 cache line deletes improved from two to ten for better self-healing capability.

The following sections describe the concepts that form the basis of leadership RAS features of IBM System p5 product line in more details.

3.1.1 Fault avoidance

System p5 servers are built on a quality-based design intended to keep errors from happening. This design includes the following features:

- Reduced power consumption, cooler operating temperatures for increased reliability, enabled by the use of copper chip circuitry, silicon-on-insulator, and dynamic clock gating
- Mainframe-inspired components and technologies

3.1.2 First Failure Data Capture

If a problem should occur, the ability to correctly diagnose it is a fundamental requirement upon which improved availability is based. The p5-570 incorporate advanced capability in start-up diagnostics and in run-time First Failure Data Capture (FDDC) that is based on strategic error checkers built into the chips.

Any errors that are detected by the pervasive error checkers are captured into Fault Isolation Registers (FIRs), which can be interrogated by the service processor. The service processor has the capability to access system components using special purpose ports or by access to the error registers. Figure 3-1 on page 69 shows a schematic of a Fault Register Implementation.

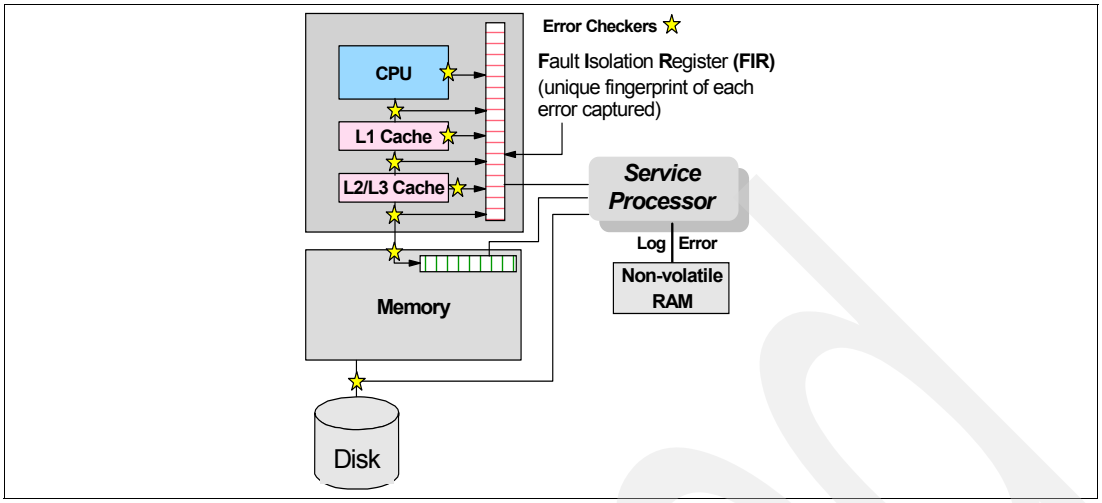


Figure 3-1 Schematic of Fault Isolation Register implementation

The FIRs are important because they enable an error to be uniquely identified, thus enabling the appropriate action to be taken. Appropriate actions might include such things as a bus retry, ECC correction, or system firmware recovery routines. Recovery routines can include dynamic deallocation of potentially failing components.

Errors are logged into the system non-volatile random access memory (NVRAM) and the service processor event history log, along with a notification of the event to AIX 5L for capture in the operating system error log. Diagnostic Error Log Analysis (*diagela*) routines analyze the error log entries and invoke a suitable action, such as issuing a warning message. If the error can be recovered, or after suitable maintenance, the service processor resets the FIRs so that they can accurately record any future errors.

The ability to correctly diagnose any pending or firm errors is a key requirement before any dynamic or persistent component deallocation or any other reconfiguration can take place.

For further details, see 3.1.7, “Resource deallocation” on page 71.

3.1.3 Permanent monitoring

The service processor provides a way to monitor the system even when the main processor is inoperable.

Mutual surveillance

The service processor can monitor the operation of the firmware during the boot process, and it can monitor the operating system for loss of control. This allows the service processor to take appropriate action, including calling for service, when it detects that the firmware or the operating system has lost control. Mutual surveillance also allows the operating system to monitor for service processor activity and can request a service processor repair action if necessary.

Environmental monitoring

Environmental monitoring related to power, fans, and temperature is done by the System Power Control Network (SPCN). Environmental critical and non-critical conditions generate Early Power-Off Warning (EPOW) events. Critical events (for example, loss of primary power) trigger appropriate signals from hardware to impacted components so as to prevent any data

loss without the operating system or firmware involvement. Non-critical environmental events are logged and reported using Event Scan.

The operating system cannot program or access the temperature threshold using the service processor.

EPOW events can, for example, trigger the following actions.

- Temperature monitoring, which increases the fans speed rotation when ambient temperature is above a preset operating range.
- Temperature monitoring warns the system administrator of potential environmental-related problems. It also performs an orderly system shutdown when the operating temperature exceeds a critical level.
- Voltage monitoring provides warning and an orderly system shutdown when the voltage is out of the operational specification.

3.1.4 Self-healing

For a system to be self-healing, it must be able to recover from a failing component by first detecting and isolating the failed component, taking it offline, fixing or isolating it, and reintroducing the fixed or replacement component into service without any application disruption. Examples include:

- *Bit steering* to redundant memory in the event of a failed memory module to keep the server operational
- *Bit-scattering*, thus allowing for error correction and continued operation in the presence of a complete chip failure (Chipkill™ recovery)
- ECC on the data received on the cache chip from the processor, which protects the interface for data from the processor to the cache
- ECC on the data read out of the eDRAM, which flags an array error
- ECC on the processor receive interface, which protects the interface for data from the cache to the processor
- L3 cache line deletes extended from 2 to 10 for additional self-healing
- ECC extended to inter-chip connections on fabric and processor bus
- *Memory scrubbing* to help prevent soft-error memory faults

Memory reliability, fault tolerance, and integrity

The p5-570 uses Error Checking and Correcting (ECC) circuitry for system memory to correct single-bit and to detect double-bit memory failures. Detection of double-bit memory failures helps maintain data integrity. Furthermore, the memory chips are organized such that the failure of any specific memory module only affects a single bit within a four-bit ECC word (*bit-scattering*), thus allowing for error correction and continued operation in the presence of a complete chip failure (*Chipkill recovery*). The memory DIMMs also use *memory scrubbing* and thresholding to determine when spare memory modules within each bank of memory should be used to replace ones that have exceeded their threshold of error count (*dynamic bit-steering*). Memory scrubbing is the process of reading the contents of the memory during idle time and checking and correcting any single-bit errors that have accumulated by passing the data through the ECC logic. This function is a hardware function on the memory controller chip and does not influence normal system memory performance.

3.1.5 N+1 redundancy

The use of redundant parts allows the p5-570 to remain operational with full resources:

- Redundant spare memory bits in L1, L2, L3, and main memory
- Redundant fans
- Redundant service processors (optional)
- Redundant power supplies

Note: With this standard feature, every p5-570 building block requires two power cords, which are not included in the base order. For maximum availability, we highly recommend connecting power cords from the same p5-570 building block to two separate PDUs in the rack; these PDUs are connected to two independent client power sources.

3.1.6 Fault masking

If corrections and retries succeed and do not exceed the threshold limits, the system remains operational with full resources, and no intervention is required. The following items are examples of fault masking:

- CEC bus retry and recovery
- PCI-X bus recovery
- ECC Chipkill soft error

3.1.7 Resource deallocation

If recoverable errors exceed threshold limits, resources can be deallocated with the system remaining operational, allowing deferred maintenance at a convenient time.

Dynamic or persistent deallocation

Dynamic deallocation of potentially failing components is nondisruptive, allowing the system to continue to run. Persistent deallocation occurs when a failed component is detected, which is then deactivated at a subsequent reboot.

Dynamic deallocation functions include:

- Processor
- L3 cache line delete
- Partial L2 cache deallocation
- PCI-X bus and slots

For dynamic processor deallocation, the service processor performs a predictive failure analysis based on any recoverable processor errors that have been recorded. If these transient errors exceed a defined threshold, the event is logged and the processor is deallocated from the system while the operating system continues to run. This feature (named *CPU Guard*) enables maintenance to be deferred until a suitable time. Processor deallocation can only occur if there are sufficient functional processors (at least two).

To verify whether CPU Guard has been enabled, run the following command:

```
lsattr -El sys0 | grep cpuguard
```

If enabled, the output will be similar to the following:

```
cpuguard    enable      CPU Guard     True
```

If the output shows CPU Guard as disabled, enter the following command to enable it:

```
chdev -l sys0 -a cpuguard='enable'
```

Cache or cache-line deallocation is aimed at performing dynamic reconfiguration to bypass potentially failing components. This capability is provided for both L2 and L3 caches. Dynamic runtime deconfiguration is provided if a threshold of L1 or L2 recovered errors is exceeded.

In the case of an L3 cache runtime array single-bit solid error, the spare chip resources are used to perform a line delete on the failing line.

PCI-X hot-plug slot fault tracking helps prevent slot errors from causing a system machine check interrupt and subsequent reboot. This provides superior fault isolation, and the error affects only the single adapter. Runtime errors on the PCI bus caused by failing adapters will result in recovery action. If this is unsuccessful, the PCI device will be gracefully shut down. Parity errors on the PCI bus itself will result in bus retry, and if uncorrected, the bus and any I/O adapters or devices on that bus will be deconfigured.

The p5-570 supports PCI Extended Error Handling (EEH) if it is supported by the PCI-X adapter. In the past, PCI bus parity errors caused a global machine check interrupt, which eventually required a system reboot in order to continue. In the p5-570 system, hardware, system firmware, and AIX 5L interaction have been designed to allow transparent recovery of intermittent PCI bus parity errors and graceful transition to the I/O device available state in the case of a permanent parity error in the PCI bus.

EEH-enabled adapters respond to a special data packet generated from the affected PCI-X slot hardware by calling system firmware, which will examine the affected bus, allow the device driver to reset it, and continue without a system reboot.

Persistent deallocation functions include:

- Processor
- Memory
- Deconfigure or bypass failing I/O adapters
- L3 cache

Following a hardware error that has been flagged by the service processor, the subsequent reboot of the system will invoke extended diagnostics. If a processor or L3 cache has been marked for deconfiguration by persistent processor deallocation, the boot process will attempt to proceed to completion with the faulty device automatically deconfigured. Failing I/O adapters will be deconfigured or bypassed during the boot process.

Note: The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable software error, software hang, hardware failure, or environmentally induced failure (such as loss of power supply).

3.1.8 Serviceability

Increasing service productivity means the system is up and running for a longer time. p5-570 improve service productivity by providing the functions described in the following sections.

Error indication and LED indicators

The p5-570 is not designed for client setup and most of the hardware features can be handled by the IBM service representative. You might like to contact the IBM service representative to know which features can be a Client Replaceable Unit (CRU). The p5-570 provides internal and external LED diagnostics that will identify parts that require service. Attenuation of the

error is provided through a series of light attention signals, starting on the exterior of the system (System Attention LED) located on the front of the system, and ending with an LED near the failing unit.

For more information about Client Replaceable Units, including videos, see:

http://publib.boulder.ibm.com/eserver

System Attention LED

The attention indicator is represented externally by an amber LED on the operator panel and the back of the system unit. It is used to indicate that the system is in one of the following states:

- Normal state, LED is off.
- Fault state, LED is on solid.
- Identify state, LED is blinking.

Additional LEDs on I/O components such as PCI-X slots and disk drives, provide status information, such as power, hot-swap, and need for service.

Concurrent maintenance

Concurrent maintenance provides replacement of the following parts while the system remains running:

- Disk drives
- Cooling fans
- Power subsystems
- PCI-X adapter cards
- Operator Panel (requires HMC guided support)

Note that the optional second service processor requires the system to be powered down in order for service.

Remember to take in consideration the touch point colors, as described in 2.15.1, “Touch point colors” on page 56.

3.2 Manageability

The functions and tools provided for IBM System p5 servers to ease management are described in the next sections.

3.2.1 Service processor

The Service processor (SP) is always working regardless of main p5 Central Electronic Complex (CEC) state. CEC can be in the following states:

- Power standby mode (power off)
- Operating, ready to start partitions
- Operating with some partitions running and an AIX 5L or Linux system in control of the machine

The SP is still working and checking the system for errors, ensuring the connection to the HMC (if present) for manageability purposes and accepting Advanced System Management Interface (ASMI) SSL network connections. The SP provides the ability to view and manage the machine-wide settings using the ASMI and allows complete system and partition

management from HMC. Also, the surveillance function of the SP is monitoring the operating system to check that it is still running and has not stalled.

Note: The IBM System p5 service processor enables the analysis of a system that will not boot. It can be performed either from the ASMI, HMC, or ASCI console (depending on presence of HMC). ASMI is provided in any case.

See Figure 3-2 for an example of the ASMI accessed from a Web browser.

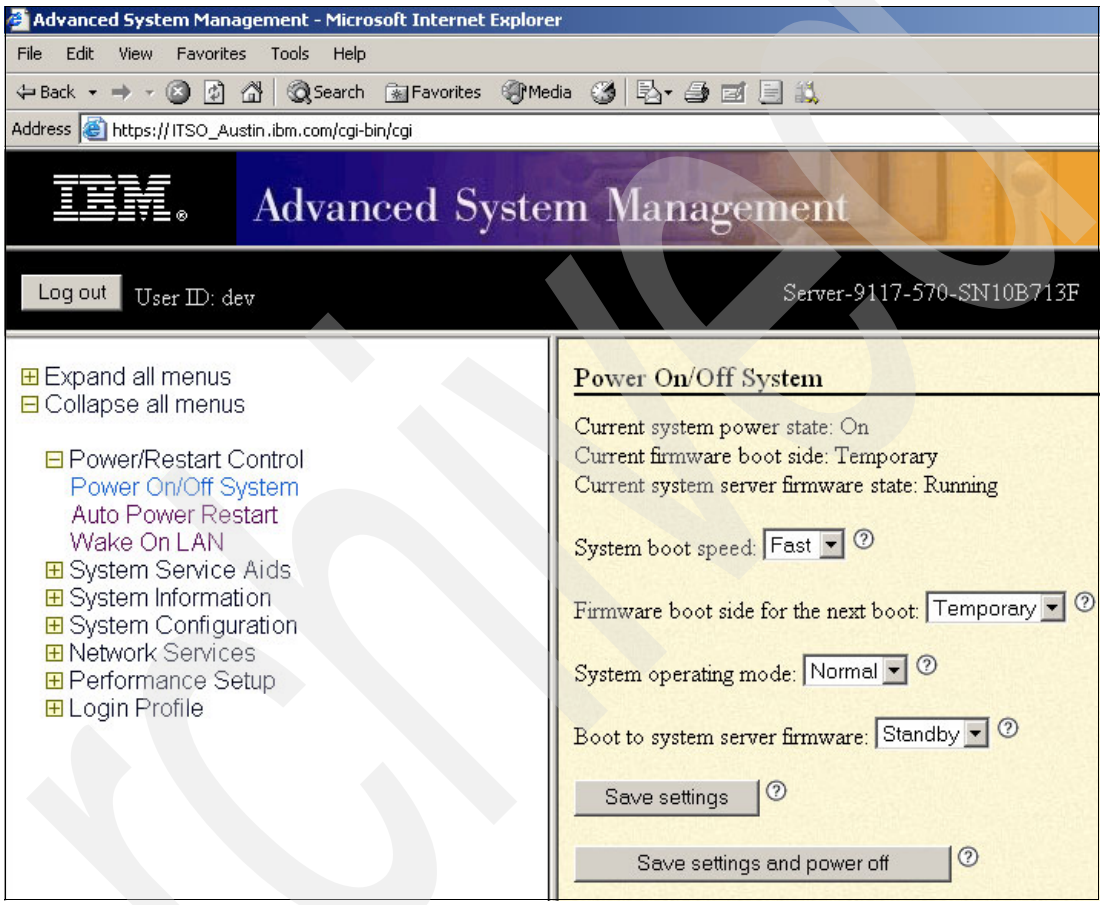


Figure 3-2 Advanced System Management main menu

3.2.2 Partition diagnostics

The diagnostics consist of stand-alone diagnostics, which are loaded from the DVD-ROM drive, and online diagnostics (available in AIX 5L).

- Online diagnostics, when installed, are resident with AIX 5L on the disk or server. They can be booted in single-user mode (service mode), run in maintenance mode, or run concurrently (concurrent mode) with other applications. They have access to the AIX 5L error log and the AIX 5L configuration data.
  - Service mode (requires service mode boot) enables the checking of system devices and features. Service mode provides the most complete checkout of the system resources. All system resources, except the SCSI adapter and the disk drives used for paging, can be tested.

- Concurrent mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, some devices might require additional actions by the user or diagnostic application before testing can be done.
- Maintenance mode enables the checking of most system resources. Maintenance mode provides the exact same test coverage as Service Mode. The difference between the two modes is the way they are invoked. Maintenance mode requires that all activity on the operating system be stopped. The **shutdown -m** command is used to stop all activity on the operating system and put the operating system into maintenance mode.

- The System Management Services (SMS) error log is accessible from the SMS menu for tests performed through SMS programs. For results of service processor tests, access the error log from the service processor menu.

Note: Because the p5-570 system has an optional DVD-ROM (FC 1994) and DVD-RAM (FC 1993), alternate methods for maintaining and servicing the system need to be available if the DVD-ROM or DVD-RAM is not ordered. It is possible to use Network Install Manager (NIM) server for this purpose.

3.2.3 Service Agent

Service Agent is an application program that operates on an IBM System p computer and monitors them for hardware errors. It reports detected errors, assuming they meet certain criteria for severity, to IBM for service with no intervention. It is an enhanced version of Service Director™ with a graphical user interface.

The key things you can accomplish using Service Agent for System p5, pSeries, and RS/6000 include:

- Automatic VPD collection
- Automatic problem analysis
- Problem-definable threshold levels for error reporting
- Automatic problem reporting; service calls placed to IBM without intervention
- Automatic client notification

In addition:

- Commonly viewed hardware errors. You can view hardware event logs for any monitored machine in the network from any Service Agent host user interface.
- High-availability cluster multiprocessing (HACMP) support for full fallback. Includes high-availability cluster workstation (HACWS) for 9076.
- Network environment support with minimum telephone lines for modems.
- Provides a communication base for the performance data collection and reporting tool Performance Management (PM/AIX). For more information about PM/AIX, see:

  `http://www.ibm.com/servers/aix/pmaix.html`

Machines are defined by using the Service Agent user interface. After the machines are defined, they are registered with the IBM Service Agent Server (SAS). During the registration process, an electronic key is created that becomes part of your resident Service Agent program. This key is used each time the Service Agent places a call for service. The IBM Service Agent Server checks the current client service status from the IBM entitlement database; if this reveals that you are not on Warranty or MA, the service call is refused and posted back using an e-mail notification.

Service Agent can be configured to connect to IBM either using a modem or network connection. In any case, the communication is encrypted and strong authentication is used. Service Agent sends outbound transmissions only and does not allow any inbound connection attempts. Only hardware machine configuration, machine status, or error information is transmitted. Service Agent does not access or transmit any other data on the monitored systems.

Thee principal ways of communication are possible:

- Dial-up using an attached modem device (uses the AT&T Global Network dialer for modem access; does not accept incoming calls to modem)
- VPN (IPsec is used in this case)
- HTTPS (can be configured to work with firewalls and authenticating proxies)

Figure 3-3 shows the possible communication paths that an IBM System p5 system be configured to use to utilize all features of Service Agent. The communication to IBM support can be either modem or network. If an HMC is present, Service Agent is an integral part of it and if activated will collect hardware related information and error messages about the whole system and partitions. If software level information (like performance data, for example) is also required, Service Agent can also be installed on any of the partitions and can be configured to act as either a gateway and connection manager or a client. The gateway and connection manager gathers data from clients and communicates to IBM on behalf of them.

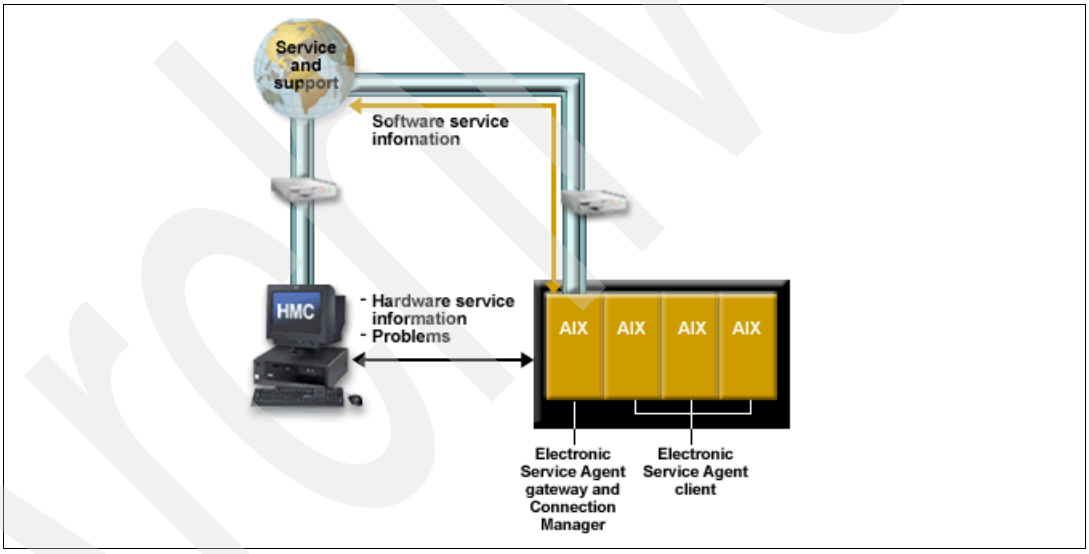


Figure 3-3 Service agent and possible connections to IBM

Additional services provided by Service Agent:

- My Systems: The client and IBM persons authorized by the client can view hardware information and error messages gathered by Service Agent on the Electronic Services WWW pages (http://www.ibm.com/support/electronic)
- Premium Search: A search service using information gathered by Service Agents (a paid service that requires a special contract).
- Performance Management: Service Agent provides the means for collecting long term performance data. The data is collected in reports accessed by the client on the WWW pages of Electronic Services (a paid service that requires a special contract).

You can download the latest version of Service Agent at:

ftp://ftp.software.ibm.com/aix/service_agent_code

Service Focal Point

Traditional service strategies become more complicated in a partitioned environment. Each logical partition reports errors it detects, without determining whether other logical partitions also detect and report the errors. For example, if one logical partition reports an error for a shared resource, such as a managed system power supply, other active logical partitions might report the same error. The Service Focal Point application helps you to avoid long lists of repetitive call-home information by recognizing that these are repeated errors and correlating them into one error.

Service Focal Point is an application on the HMC that enables you to diagnose and repair problems on the system. In addition, you can use Service Focal Point to initiate service functions on systems and logical partitions that are not associated with a particular problem. You can configure the HMC to use the Service Agent call-home feature to send IBM event information. Service Focal Point is available also in Integrated Virtualization Manager. It allows you to manage serviceable events, create serviceable events, manage dumps, and collect vital product data (VPD) but no reporting via Service Agent is possible.

3.2.4 IBM System p5 firmware maintenance

IBM System p5, pSeries, and RS/6000 Customer-Managed Microcode is a methodology that enables you to manage and install microcode updates on System p5, pSeries, and RS/6000 systems and associated I/O adapters. The IBM System p5 Microcode can be installed either from HMC or from a running partition; for update details, see 2.15.3, “System firmware” on page 59.

If you use an HMC to manage your server, you can use the HMC interface to view the levels of server firmware and power subsystem firmware that are installed on your server, and are available to download and install.

Each System p5 server has the following levels of server firmware and power subsystem firmware:

- Installed level: This is the level of server firmware or power subsystem firmware that has been installed and will be installed into memory after the managed system is powered off and powered on. It is installed on the *t* side of system firmware (for an additional discussion about firmware sides, see 2.15.3, “System firmware” on page 59).
- Activated level: This is the level of server firmware or power subsystem firmware that is active and running in memory.
- Accepted level: This is the backup level of server or power subsystem firmware. You can return to this level of server or power subsystem firmware if you decide to remove the installed level. It is installed on the *p* side of system firmware (for an additional discussion about firmware sides, see 3.2.1, “Service processor” on page 73).

IBM introduced the Concurrent Firmware Maintenance (CFM) function on System p5 servers in system firmware level 01SF230_126_120, which was released on 16 June 2005. This function supports nondisruptive system firmware service packs to be applied to the system concurrently (without requiring a reboot to activate changes). For systems that are not managed by an HMC, the installation of system firmware is always disruptive.

The concurrent levels of system firmware might, on occasion, contain fixes that are known as deferred. These deferred fixes can be installed concurrently, but will not be activated until the next IPL. Deferred fixes, if any, will be identified in the Firmware Update Descriptions table of

the firmware release. For deferred fixes within a service pack, only the fixes in the service pack that cannot be concurrently activated are deferred.

Use the following information as a reference to determine whether your installation will be concurrent or disruptive.

Figure 3-4 shows the system firmware file naming convention.

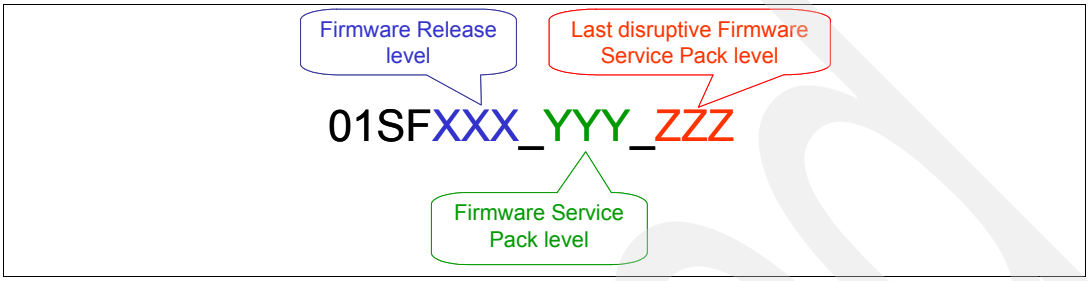


Figure 3-4 System firmware file naming convention

An installation is disruptive if:

- The release levels (XXX) of currently installed and new firmware are different.
- The service pack level (YYY) and the last disruptive service pack level (ZZZ) are equal in new firmware.

Otherwise, an installation is concurrent if:

- If the service pack level (YYY) of new firmware is higher than the service pack level currently installed on the system and the above conditions for disruptive installation are not met.

3.3 Cluster solution

Today's IT infrastructure requires that servers meet increasing demands, while offering the flexibility and manageability to rapidly develop and deploy new services. IBM clustering hardware and software provide the building blocks, with availability, scalability, security, and single-point-of-management control, to satisfy these needs. The advantages of clusters are:

- Large-capacity data and transaction volumes, including support of mixed workloads
- Scale-up (add processors) or scale-out (add servers) without downtime
- Single point-of-control for distributed and clustered server management
- Simplified use of IT resources
- Designed for 24 x 7 access to data applications
- Business continuity in the event of disaster

The POWER5+ processor-based AIX 5L and Linux cluster targets scientific and technical computing, large-scale databases, and workload consolidation. IBM Cluster Systems Management software (CSM) is designed to provide a robust, powerful, and centralized way to manage a large number of POWER5 processor-based servers, all from one single point-of-control. Cluster Systems Management can help lower the overall cost of IT ownership by helping to simplify the tasks of installing, operating, and maintaining clusters of servers. Cluster Systems Management can provide one consistent interface for managing both AIX 5L and Linux nodes (physical systems or logical partitions), with capabilities for remote parallel network install, remote hardware control, and distributed command execution.

Cluster Systems Management for AIX 5L and Linux on POWER processor-based servers is supported on the p5-570. For hardware control, an HMC is required. One HMC can also control several servers that are part of the cluster. If a p5-570 that is configured in partition mode (with physical or virtual resources) is part of the cluster, all partitions must be part of the cluster.

Monitoring is much easier to use, and the system administrator can monitor all of the network interfaces, not just the switch and administrative interfaces. The management server pushes information out to the nodes, which releases the management server from having to trust the node. In addition, the nodes do not have to be network-connected to each other. This means that giving root access on one node does not mean giving root access on all nodes. The base security setup is all done automatically at install time.

For information regarding the IBM Cluster Systems Management for AIX 5L, HMC control, cluster building block servers, and cluster software available, visit the following links:

- IBM System Cluster 1600

  http://www.ibm.com/servers/eserver/clusters/hardware/1600.html

- IBM System Cluster 1350™

  http://www.ibm.com/servers/eserver/clusters/hardware/1350.html

The CSM ships with AIX 5L itself (a 60-day Try and Buy license is shipped with AIX). The CSM client side is automatically installed and ready when you install AIX, so each system or logical partition is cluster-ready.

The CSM V1.4 on AIX 5L and Linux introduces an optional IBM CSM High Availability Management Server (HA MS) feature, which is designed to allow automated failover of the CSM management server to a backup management server. In addition, sample scripts for setting up NTP[1] and network tuning (AIX 5L only) configurations, and the capability to copy files across nodes or node groups in the cluster, can improve cluster ease of use and site customization.

[1] Network Time Protocol

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this Redpaper.

IBM Redbooks

For information about ordering these publications, see “How to get IBM Redbooks” on page 83. Note that some of the documents referenced here may be available in softcopy only.

- *Advanced POWER Virtualization on IBM @server p5 Servers: Architecture and Performance Considerations*, SG24-5768
- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *LPAR Simplification Tools Handbook*, SG24-7231
- *Partitioning Implementations for IBM @server p5 Servers*, SG24-7039
- *Problem Solving and Troubleshooting in AIX 5L*, SG24-5496
- *Virtual I/O Server Integrated Virtualization Manager*, SG24-4061
- *IBM @server p5 510 Technical Overview and Introduction*, REDP-4001
- *IBM @server p5 520 Technical Overview and Introduction*, REDP-9111
- *IBM @server p5 550 Technical Overview and Introduction*, REDP-9113
- *IBM @server p5 570 Technical Overview and Introduction*, REDP-9117
- *IBM @server p5 590 and 595 System Handbook*, SG24-9119
- *IBM @server p5 590 and 595 Technical Overview and Introduction*, REDP-4024
- *IBM @server pSeries Sizing and Capacity Planning: A Practical Guide*, SG24-7071
- *IBM System p5 505 and 505Q Technical Overview and Introduction*, REDP-4079
- *IBM System p5 510 and 510Q Technical Overview and Introduction*, REDP-4136
- *IBM System p5 520 and 520Q Technical Overview and Introduction*, REDP-4137
- *IBM System p5 550 and 550Q Technical Overview and Introduction*, REDP-4138
- *IBM System p5 560Q Technical Overview and Introduction*, REDP-4139

Other publications

These publications are also relevant as further information sources:

- *7014 Series Model T00 and T42 Rack Installation and Service Guide*, SA38-0577, contains information regarding the 7014 Model T00 and T42 Rack, in which this server can be installed.
- *Planning for Partitioned-System Operations*, SA38-0626, provides information to planners, system administrators, and operators about how to plan for installing and using a partitioned server. It also discusses some issues associated with the planning and implementing of partitioning.

- *RS/6000 and eServer pSeries Diagnostics Information for Multiple Bus Systems*, SA38-0509, contains diagnostic information, service request numbers (SRNs), and failing function codes (FFCs).
- *System p5, eServer p5 Customer Service Support and Troubleshooting*, SA38-0538, contains information regarding slot restrictions for adapters that can be used in this system.
- *System Unit Safety Information*, SA23-2652, contains translations of safety information used throughout the system documentation.

Online resources

These Web sites and URLs are also relevant as further information sources:

- AIX 5L operating system maintenance packages downloads

  http://www.ibm.com/servers/eserver/support/unixservers/aixfixes.html
- News on new computer technologies

  http://www.ibm.com/chips/micronews
- Copper circuitry

  http://domino.research.ibm.com/comm/pr.nsf/pages/rsc.copper.html
- IBM Systems Hardware Information Center documentation

  http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp
- IBM Systems Information Centers

  http://publib.boulder.ibm.com/eserver/
- IBM microcode downloads

  http://www14.software.ibm.com/webapp/set2/firmware/gjsn
- Support for IBM System p servers

  http://www.ibm.com/servers/eserver/support/unixservers/index.html
- Technical help database for AIX 5L

  http://www14.software.ibm.com/webapp/set2/srchBroker/views/srchBroker.jsp?rs=111
- IBMlink

  http://www.ibmlink.ibm.com
- Linux for IBM System p5

  http://www.ibm.com/systems/p/linux/
- Microcode Discovery Service

  http://www14.software.ibm.com/webapp/set2/mds/fetch?page=mds.html

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM System p5 570 Technical Overview and Introduction

Redpaper

Finer system granulation using Micro-Partitioning technology to help lower TCO

Modular midrange solution for managing On Demand Business

New POWER5+ processor options using DDR2 memory technology

This IBM Redpaper is a comprehensive guide covering the IBM System p5 570 UNIX server. It introduces major hardware offerings and discusses their prominent functions.

Professionals wishing to acquire a better understanding of IBM System p5 products should read this Redpaper. The intended audience includes:

- Customers
- Sales and marketing professionals
- Technical support professionals
- IBM Business Partners
- Independent software vendors

This Redpaper expands the current set of IBM System p documentation by providing a desktop reference that offers a detailed technical description of the p5-570 system.

This Redpaper does not replace the latest marketing materials and tools. It is intended as an additional source of information that, together with existing sources, may be used to enhance your knowledge of IBM server solutions.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks