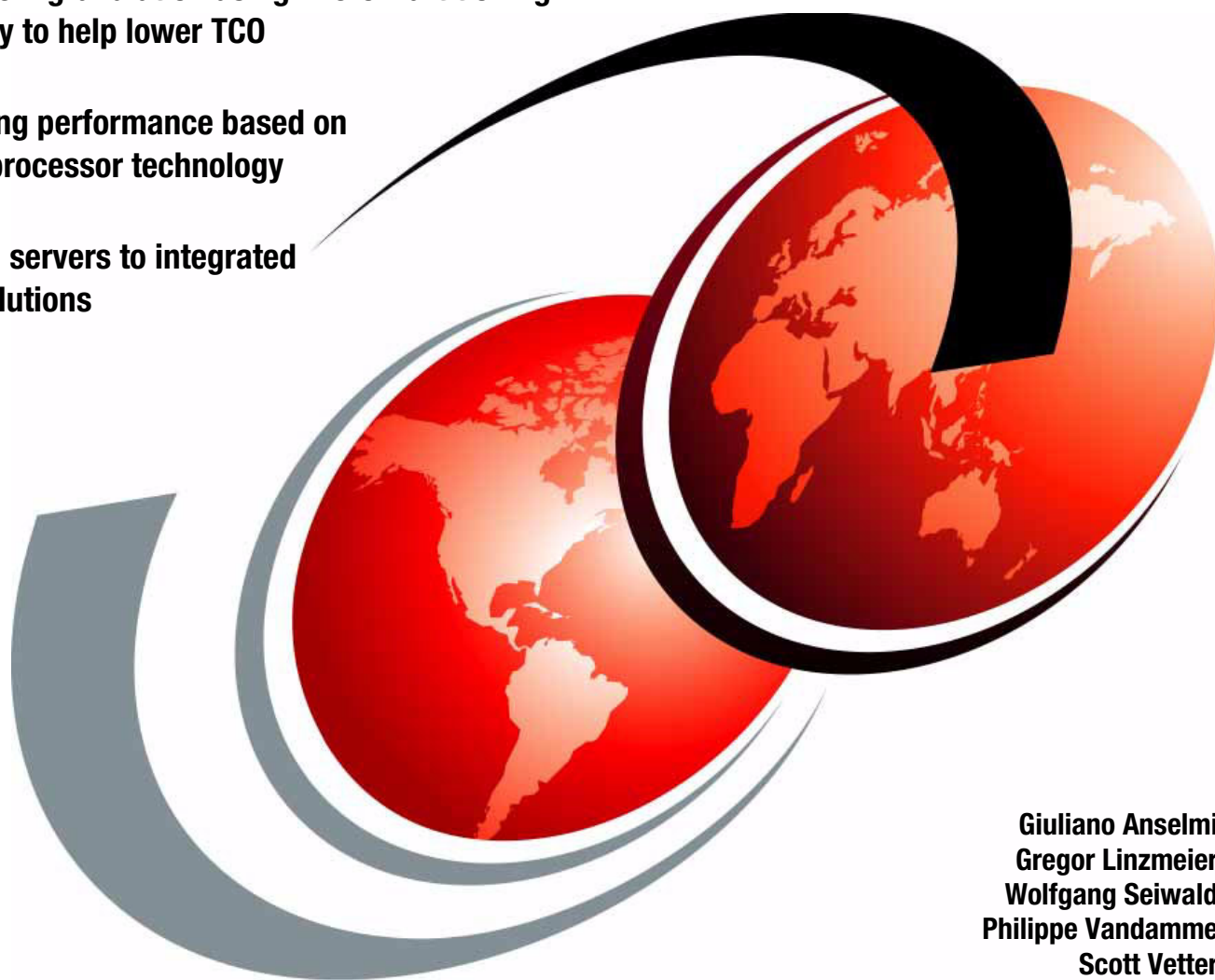


IBM *e*server p5 520 Technical Overview and Introduction

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

Outstanding performance based on POWER5 processor technology

From Web servers to integrated cluster solutions



Giuliano Anselmi
Gregor Linzmeier
Wolfgang Seiwald
Philippe Vandamme
Scott Vetter



International Technical Support Organization

**IBM @server p5 520 Technical Overview and
Introduction**

October 2004

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

Second Edition (October 2004)

This edition applies to the IBM @server p5 520 and AIX 5L Version 5.3, product number 5765-G03.

© Copyright International Business Machines Corporation 2004. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	vii
Trademarks	viii
Preface	ix
The team that wrote this Redpaper	ix
Become a published author	x
Comments welcome	x
Chapter 1. General description	1
1.1 System specifications	3
1.2 Physical package	3
1.3 p5-520 desktside and rack model	3
1.3.1 IBM eServer p5 520 desktside	4
1.3.2 IBM eServer p5 520 rack-mounted	4
1.4 Minimum and optional features	5
1.4.1 Processor features	6
1.4.2 Memory features	7
1.4.3 Disk and media features	7
1.4.4 USB diskette drive	8
1.4.5 I/O drawers	8
1.4.6 Hardware Management Console models	9
1.5 Value Paks	10
1.6 System racks	10
1.6.1 IBM RS/6000 7014 Model T00 Enterprise Rack	11
1.6.2 IBM RS/6000 7014 Model T42 Enterprise Rack	11
1.6.3 AC Power Distribution Unit and rack content	12
1.6.4 Rack-mounting rules for p5-520	12
1.6.5 Additional options for rack	13
1.6.6 OEM rack	14
Chapter 2. Architecture and technical overview	17
2.1 The POWER5 chip	18
2.1.1 Simultaneous multi-threading	19
2.1.2 Dynamic power management	20
2.1.3 The POWER chip evolution	20
2.1.4 CMOS, copper, and SOI technology	21
2.2 Processor and cache	21
2.2.1 Available processor speeds	21
2.3 Memory subsystem	22
2.3.1 Memory placement rules	22
2.3.2 Memory restriction	23
2.3.3 Memory throughput	24
2.4 System buses	24
2.4.1 RIO buses and GX card	24
2.5 Internal I/O subsystem	24
2.5.1 PCI-X slots and adapters	24
2.5.2 LAN adapters	25
2.5.3 Graphic accelerators	25
2.5.4 Audio adapter	25

2.5.5 SCSI adapters	26
2.6 Internal serial ports	26
2.7 Internal storage	26
2.7.1 Internal media devices	26
2.7.2 Internal hot swappable SCSI disks	27
2.7.3 RAID options	28
2.8 External I/O subsystem	28
2.8.1 I/O drawers	28
2.8.2 7311 I/O drawer RIO-2 cabling	30
2.8.3 7311 Model D20 I/O drawer SPCN cabling	30
2.8.4 External disk subsystem	31
2.9 Dynamic logical partitioning	32
2.10 Virtualization	33
2.10.1 Virtual Ethernet	33
2.10.2 Advanced POWER Virtualization feature	33
2.11 Service processor	36
2.11.1 Service processor base	36
2.11.2 Service processor extender	36
2.12 Boot process	37
2.12.1 IPL flow without an HMC attached to the system	37
2.12.2 Hardware Management Console	38
2.12.3 IPL flow with an HMC attached to the system	38
2.12.4 Definitions of partitions	39
2.12.5 Hardware requirements for partitioning	40
2.12.6 Specific partition definitions used for Micro-Partitioning	40
2.12.7 System Management Services	41
2.12.8 Boot options	42
2.12.9 Additional boot options	43
2.12.10 Security	43
2.13 Operating system requirements	43
2.13.1 AIX 5L	43
2.13.2 Linux	44
Chapter 3. RAS and manageability	45
3.1 Reliability, availability, and serviceability	46
3.1.1 Fault avoidance	46
3.1.2 First Failure Data Capture	46
3.1.3 Permanent monitoring	47
3.1.4 Self-healing	48
3.1.5 N+1 redundancy	48
3.1.6 Fault masking	49
3.1.7 Resource deallocation	49
3.1.8 Serviceability	50
3.2 Manageability	51
3.2.1 Service processor	51
3.2.2 Service Agent	52
3.2.3 IBM eServer p5 Customer-Managed Microcode	53
3.2.4 Service Update Management Assistant	53
3.3 IBM eServer Cluster 1600	54
Related publications	57
IBM Redbooks	57
Other publications	57

Online resources	58
How to get IBM Redbooks	59
Help from IBM	59

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 illustrates 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. 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 IBM's application programming interfaces.

Trademarks

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

Advanced Micro-Partitioning™
AIX®
AIX 5L™
Chipkill™
Electronic Service Agent™
Enterprise Storage Server®
@server®
@server®

HACMP™
i5/OS™
IBM®
Micro-Partitioning™
POWER™
POWER4™
POWER4+™
POWER5™

PowerPC®
pSeries®
Redbooks™
Redbooks (logo) ™
RS/6000®
Service Director™
TotalStorage®

The following terms are trademarks of other companies:

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

Preface

This document is a comprehensive guide covering the IBM @server® p5 520 UNIX® servers. We introduce major hardware offerings and discuss their prominent functions.

Professionals wishing to acquire a better understanding of IBM @server p5 products should consider reading this document. The intended audience includes:

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

This document expands the current set of IBM @server documentation by providing a desktop reference that offers a detailed technical description of the p5-520 system.

This publication does not replace the latest IBM @server pSeries® marketing materials and tools. It is intended as an additional source of information that, together with existing sources, can 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 @server 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 12 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 RS/6000 and pSeries workstation and entry servers as part of the Systems and Technology Group in Mainz, Germany supporting IBM sales, Business Partners, and customers with pre-sales consultation and implementation of client/server environments. He has worked for more than 13 years as an infrastructure specialist for RT, RS/6000, 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 five years has been in the areas of the IBM @server pSeries systems and the IBM AIX operating system.

Philippe Vandamme is an IT Specialist working in pSeries Field Technical Support in Paris, France, EMEA West region. With 15 years of experience in semi-conductor fabrication and manufacturing and associated technologies, he is now in charge of pSeries 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 document was managed by:

Scott Vetter
IBM U.S.

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

Ron Arroyo, Steve Pittman, Barb Hewitt, Thoi Nguyen, Jan Palmer, Charlie Reeves, Craig Shempert, Scott Smylie, Joel Tandler, Ed Toutant, Jane Arbeitman, Tenley Jackson, Andy McLaughlin, Janine Tally.
IBM U.S.

Derrick Daines, Dave Williams
IBM U.K.

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 and/or customers.

Your efforts will help increase product acceptance and customer 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 Internet note to:

redbook@us.ibm.com

- ▶ Mail your comments to:

IBM® Corporation, International Technical Support Organization
Dept. JN9B Building 905 Internal Zip 9053D004
11501 Burnet Road
Austin, Texas 78758-3493



General description

The IBM *e*server p5 520 desktide and rack-mount server are designed for greater application flexibility, with innovative technology, to capitalize on the e-business revolution at the entry-level for server environments. Introduced with the POWER4™ and POWER4+™ technology in 2001, available from the 1-way entry-level through the 32-way high-end pSeries systems, the IBM POWER™ architecture achieved a new stage of capability characteristics by introducing features such as logical partitioning. With POWER5™ microprocessor technology, the p5-520 is the first cost-effective, high-performance, entry UNIX server that includes the next development of the IBM partitioning concept, Micro-Partitioning™ technology.

Dynamic logical partitioning (LPAR) on a 2-way p5-520 allows up to two dedicated partitions. In addition, the optional Advanced POWER Virtualization feature is designed to support up to 20 partitions on a 2-way system. 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. In addition to the Micro-Partitioning technology, the Advanced POWER Virtualization feature allows sharing of physical network adapters and enables the virtualization of SCSI storage.

In combination with the extraordinary POWER5 processor, Micro-Partitioning technology increases the system management efficiency and lowers the operating expenses by the multiple use of single physical resources installed in the p5-520 system. Simultaneous multi-threading, a standard feature of POWER5 technology, allows two threads to be executed at the same time on a single processor. Simultaneous multi-threading is selectable with dedicated processors or processors using Micro-Partitioning technology.

The symmetric multiprocessor (SMP) p5-520 system features 1-way (1.5 GHz) or 2-way (1.5 GHz or 1.65 GHz), state-of-the-art, 64-bit, copper and silicon on insulator (SOI)-based POWER5 microprocessors with 36 MB off-chip Level 3 cache soldered directly to the system planar on 2-way configurations. Main memory, starting at 512 MB for 1.5 GHz models and 1 GB for 1.65 GHz models, can be expanded up to 32 GB, 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.

Included in the p5-520 are six hot-plug PCI-X slots with Enhanced Error Handling (EEH), one dual-channel Ultra320 SCSI controller, a dual-port 10/100/1000 Mbps integrated Ethernet controller, two serial ports, two USB 2.0 capable ports, two HMC ports, two RIO-2 ports, and two System Power Control Network (SPCN) ports.

The p5-520 includes four front-accessible, hot-swap-capable disk bays in a minimum configuration with an additional four hot-swap-capable disk bays as an optional feature. The eight disk bays can accommodate up to 1.17 TB of disk storage using the 146.8 GB Ultra320 SCSI disk drives. Three non-hot-swappable media bays are used to accommodate additional devices. Two media bays only accept slim line media devices, such as DVD-ROM or DVD-RAM drives, and one half-height bay is used for a tape drive. The p5-520 also has I/O extension capability using the RIO-2 bus that allows attachment of the 7311 Model D20 I/O drawers.

Additional reliability and availability features include redundant hot-plug cooling fans and redundant power supply. Along with these hot-plug components, the p5-520 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.

1.1 System specifications

Table 1-1 lists the general system specifications of the p5-520 system.

Table 1-1 IBM @server p5-520 specifications

Description	Range
Operating temperature	5 to 35 degrees Celsius (41 to 95 F)
Relative humidity	8% to 80%
Operating voltage	100 to 127 or 200 to 240 V AC (auto-ranging)
Operating frequency	47/63 Hz
Maximum power consumption	600 watts maximum
Maximum thermal output	2047 Btu ^a /hour (maximum)

a. British Thermal Unit

1.2 Physical package

The following sections discuss the major physical attributes found on the p5-520 system in rack-mounted and deskside versions, as shown in Table 1-2. The p5-520 is a 4U¹, 19-inch rack-mounted system or deskside system depending on the feature code.

Table 1-2 Physical packaging of the p5-520

Dimension	Rack (FC 7918)	Deskside (FC 7919)
Height	178 mm (7.0 inches)	533 mm (21.0 inches)
Width	437 mm (17.2 inches)	201 mm (7.9 inches)
Depth	508 mm (20.0 inches)	584 mm (23.0 inches)
Weight		
Minimum configuration	35.5 kg (78 pounds)	
Maximum configuration	43.0 kg (95 pounds)	

1.3 p5-520 deskside and rack model

Figure 1-1 shows a detailed view of the p5 520 deskside and rack-mounted versions.

¹ One Electronic Industries Association Unit (1U) is 44.45 mm (1.75 inches).



Figure 1-1 The p5-520 rack-mount and deskside versions

1.3.1 IBM eServer p5 520 deskside

The p5-520, when configured as a deskside server, is ideal for environments requiring the user to have local access to the machine. A typical example of this would be applications requiring a native graphics display.

To order a p5-520 system as a deskside version, FC 7919 is required. The system is designed to be set up by the customer and, in most cases, will not require the use of any tools. Full set-up instructions are included with the system.

The GXT135P 2D graphics accelerator with analog and digital interfaces (FC 2849) is available and is supported for SMS, firmware menus, and other low-level functions, as well as when AIX or Linux® starts the X11-based graphical user interface. You can use graphical AIX system tools for configuration management if the adapter is connected to the primary console, such as the IBM L200p Flat Panel Monitor (FC 3636) or the IBM T541H 15-inch TFT Color Monitor (FC 3637).

1.3.2 IBM eServer p5 520 rack-mounted

The p5-520, when configured as a 4U rack-mounted server, is intended to be installed in a 19-inch rack, thereby enabling efficient use of computer room floor space. If the IBM 7014 T42 rack is used to mount the p5 520, it is possible to place up to 10 systems in an area of 644 mm (25.5 inches) x 1147 mm (45.2 inches).

To order a p5-520 system as a rack-mounted version, FC 7918 must be selected. In addition to the rack-mounted version, the p5-520 can be installed in either IBM or OEM racks. Therefore, you are required to select one of the following features:

- ▶ IBM Rack-mount Drawer Rail Kit (FC 7160)
- ▶ OEM Rack-mount Drawer Rail Kit (FC 7161)

Included with the p5 520 rack-mounted server packaging are all of the components and instructions necessary to enable installation in a 19-inch rack using suitable tools.

The GXT135P 2D graphics accelerator with analog and digital interfaces (FC 2849) is available and is supported for SMS, firmware menus, and other low-level functions, as well as when AIX or Linux starts the X11-based graphical user interface. You can use graphical AIX system tools for configuration management if the adapter is connected to a common maintenance console, such as the 7316-TF3 rack-mounted flat-panel display. Figure 1-2 shows detailed views of the p5-520 rack-mount system.

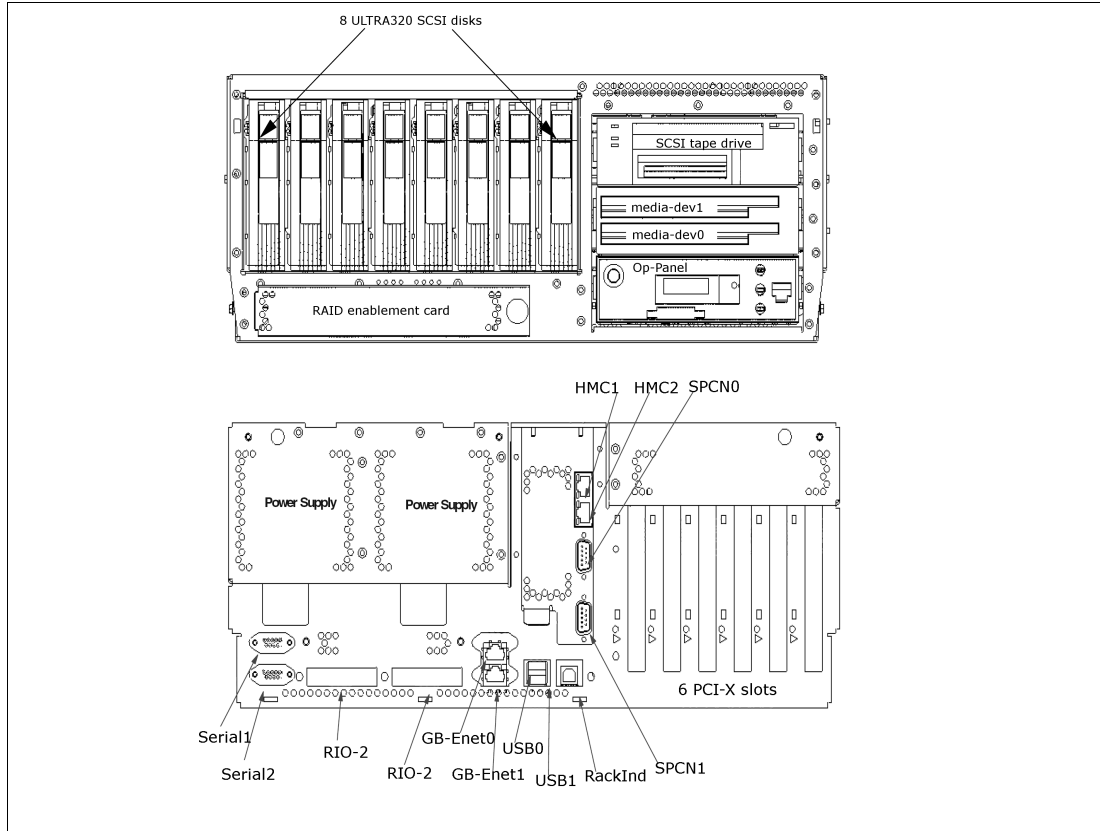


Figure 1-2 Detailed views of the p5-520 rack-mount system

1.4 Minimum and optional features

The p5-520 system is based on a flexible, modular design of one POWER5 chip packaged in a Dual Chip Module (DCM) and an integrated L3 cache, soldered directly to the system planar. The p5-520 is available only in a 2-way configuration, and it features:

- ▶ From 512 MB (1.5 GHz models) or 1 GB (1.65 GHz models) to 32 GB of total system memory capacity using DDR1 DIMM technology
- ▶ Four SCSI disk drives in a minimum configuration, eight SCSI disk drives with an optional second 4-pack enclosure for a total internal storage capacity of 1.17 TB using 146.8 GB disk drives
- ▶ Six PCI-X slots (three long and three short slots)
- ▶ Two slim-line media bays for optional storage devices
- ▶ One half-high bay for an optional tape device

The p5-520, including the service processor described in 2.11, “Service processor” on page 36, supports the following native ports:

- ▶ Two 10/100/1000 Ethernet ports
- ▶ Two serial ports
- ▶ Two USB 2.0 ports
 - Optionally, an external USB diskette drive 1.44 (FC 2591) is available.
- ▶ Two HMC ports
- ▶ Two remote I/O (RIO-2) ports and two SPCN ports

In addition, the p5-520 features one internal Ultra320 SCSI dual channel controller, redundant hot-swap power supply (optional), and cooling fans.

The system supports 32-bit and 64-bit applications and requires a specific level of operating system. See 2.13, “Operating system requirements” on page 43.

1.4.1 Processor features

The p5-520 features one POWER5 chip with two processor cores, running at 1.5 GHz on 1-way, 1.5 GHz 2-way, and 1.65 GHz 2-way features that share 1.9 MB of L2 on chip cache, 36 MB of L3 cache (on 2-way models), and eight slots for memory DIMMs using DDR1 technology. For a list of available features; see Table 1-3.

Table 1-3 Processor feature codes

Processor FC	Description
5231	1-way 1.5 GHz, no L3 cache, eight DDR1 DIMM sockets
5226	2-way 1.5 GHz, 36 MB L3 cache, eight DDR1 DIMM sockets
5229	2-way 1.65 GHz, 36 MB L3 cache, eight DDR1 DIMM sockets

The p5-520 POWER5 chip is mounted to the system planar and directly interfaced to the memory buffer SMI chips through an elastic interface, as shown in Figure 1-3.

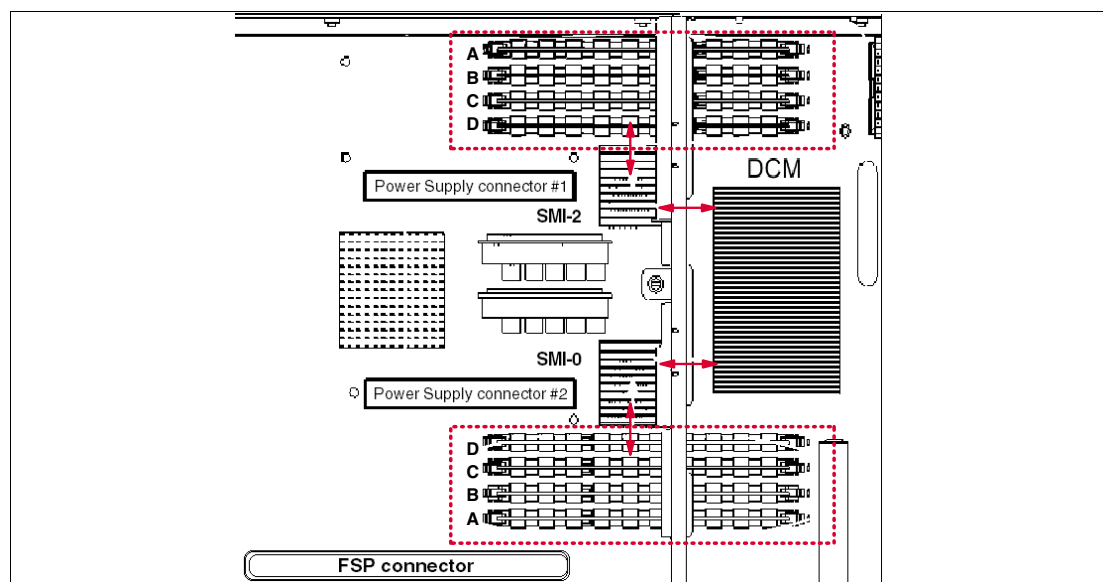


Figure 1-3 Enlarged view of the processor subsystem on the system planar

1.4.2 Memory features

The system planar of the p5-520 system has eight sockets for memory DIMMs. The minimum memory requirement is 512 GB, and the maximum capacity is 36 GB. Table 1-4 lists the available memory features. FC 4443 is only available on 1.5 GHz processors.

Table 1-4 Memory feature codes

Feature code	Description
4443	512 MB (2x 256 MB) DIMMS, 266 MHz DDR1 SDRAM
4444	1 GB (4x 256 MB) DIMMS, 266 MHz DDR1 SDRAM
4447	2 GB (4x 512 MB), DIMMS, 266 MHz DDR1 SDRAM
4445	4 GB (4x 1024 MB), DIMMS, 266 MHz DDR1 SDRAM
4449	8 GB (4x 2048 MB), DIMMS, 266 MHz DDR1 SDRAM
4450	16 GB (4x 4096 MB) DIMMS, 266 MHz DDR1 SDRAM

1.4.3 Disk and media features

The minimum p5-520 configuration includes a 4-pack disk drive enclosure. A second 4-pack disk drive enclosure can be installed by ordering FC 6574 or FC 6594. The p5-520 features up to eight disk drive bays, two slim-line media device bays, and one half-height media bay. The minimum configuration requires at least one disk drive. Table 1-5 shows the disk drive feature codes that each bay can contain.

Table 1-5 Disk drive feature code description

Feature code	Description
3273	36.4 GB 10 K RPM Ultra3 SCSI disk drive assembly
3277	36.4 GB 15 K RPM Ultra3 SCSI disk drive assembly
3274	73.4 GB 10 K RPM Ultra3 SCSI disk drive assembly
3278	73.4 GB 15 K RPM Ultra3 SCSI disk drive assembly
3275	146.8 GB 10 K RPM Ultra3 SCSI disk drive assembly

Any combination of DVD-ROM and DVD-RAM drives of the following devices can be installed in the two slim-line bays:

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

A logical partition running a supported release of LINUX requires a DVD-ROM drive or DVD-RAM drive to provide a way to run a **diag** of the CD for hardware diagnostics from the CD. Concurrent diagnostics, as provided by AIX, is not available on Linux at the time of writing.

Supplementary devices can be installed in the half-height media bay, such as:

- ▶ IBM 80/160 GB Internal Tape Drive with VXA Technology, FC 6120
- ▶ 60/150 GB 16-bit 8 mm Internal Tape Drive, FC 6134
- ▶ 36/72 GB 4 mm Internal Tape Drive, FC 6258

Devices installed in the media bays must be assigned as a group to a single LPAR on a partitioned system.

1.4.4 USB diskette drive

For today's administration tasks, an internal diskette drive is not state-of-the-art. In some situations, the external USB 1.44 MB diskette drive for p5-520 systems (FC 2591) is helpful. This super-slim-line and lightweight USB V2 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.

1.4.5 I/O drawers

The p5-520 has six internal PCI-X slots, where three of them are long slots and three are short slots. If more PCI-X slots are needed, especially well-suited to extend the number of LPARs and partitions, up to four 7311 Model D20 drawers can be connected to the two RIO-2 ports on the rear of the system that are provided in a minimum configuration.

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 p5-520 system with increasing I/O requirements. When a p5-520 is fully configured with four attached 7311 Model D20 drawers, the combined system supports up to 34 PCI-X adapters (in a maximum configuration (Remote I/O expansion cards are required) and 56 hot-swappable SCSI disks, for a total internal capacity of 8.2 TB using 146.8 GB disks.

PCI-X and PCI cards are inserted from the top of the I/O drawer down into the slot from the drawers front service position. The installed adapters are protected by plastic separators, 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 volt, keyed, 133 MHz hot-pluggable
- ▶ Two 6-pack hot-swappable SCSI bays (optional)
- ▶ Optional redundant hot-plug power
- ▶ Two RIO-2 ports and two SPCN ports

Note: A 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 inches)
- ▶ Depth: 610 mm (24.0 inches)
- ▶ Height: 178 mm (7.0 inches)

- ▶ Weight: 45.9 kg (101 pounds)

Figure 1-4 on page 9 shows the different views of the 7311-D20 I/O drawer.

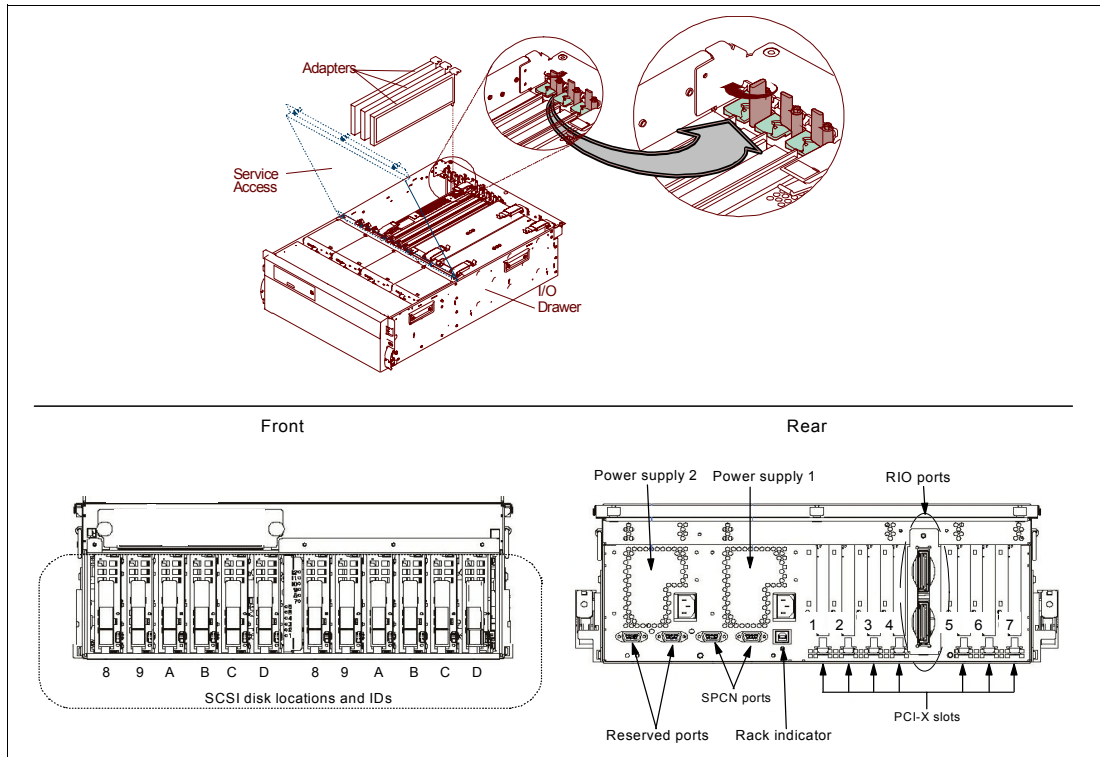


Figure 1-4 7311-D20 I/O drawer views

Note: The 7311 Model D20 I/O drawer is designed to be installed by an IBM service representative.

I/O drawers and usable PCI slots

Only the 7311 Model D20 I/O drawer is supported on a p5-520 system. Table 1-6 summarizes the maximum number of I/O drawers supported and the total number of PCI-X slots available.

Table 1-6 Maximum number of I/O drawers supported and total number of PCI slots

Maximum number of I/O drawers	Total number of PCI-X slots
4	34

1.4.6 Hardware Management Console models

The Hardware Management Console (HMC) provides a set of functions that is necessary to manage the p5-520 system when LPAR, IBM @server 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 allocated to an individual partition.

The 7310 Model CR2 or the 7310 Model C03 HMCs are specifically for POWER5 processor-based systems. However, an existing 7315 Model CR2 and the 7315 Model C03 (POWER4 processor-based systems HMC) can be converted for POWER5 processor-based

system use when it is loaded with the HMC software required for POWER5 processor-based systems (FC 0961).

POWER5 processor-based system HMCs require Ethernet connectivity. Ensure that sufficient Ethernet adapters are available to enable public and private networks, if you need both. The 7310 Model C03 is a desktop model with only one native 10/100/1000 Ethernet port, but three additional PCI slots. The 7310 Model CR2 is a 1U, 19-inch rack-mountable drawer that has two native Ethernet ports and two additional PCI slots.

When an HMC is connected to the p5-520, the p5-520 integrated serial ports are disabled. If you need serial connections, for example, non-Ethernet HACMP™ heartbeat, you need to provide an async adapter to provide additional serial connections.

Note: It is not possible to connect POWER4 and POWER5 processor-based systems simultaneously to the same HMC, however one HMC can manage several systems.

1.5 Value Paks

Value Paks are a new offering available on an initial order only. They provide a predefined configuration, designed to meet typical customer requirements. Special reduced pricing is available when a system order satisfies specific configuration requirements for memory, disk drives, and processors. When a Value Pak is ordered, it is still possible to select additional features.

If a p5-520 is ordered in a minimum configuration, you will be entitled to a discounted AIX operating system license or you can choose to purchase the system with no operating system. The minimum features, as shown in Table 1-7, can be upgraded as needed and still receive the discounted AIX operating system. A DVD-RAM or DVD-ROM feature is required.

Note: The Value Pak is available in an initial order only. Value Pak configurations are subject to change as new features are introduced.

Table 1-7 Value Pak configuration

Value Paks	Processors	Memory (MB)	Disk	DVD
1.5 GHz	1-way, FC 5231 x 1	1024, FC 4444 x 1	2 x 36.4 GB (FC 3273)	ROM or RAM
1.5 GHz	2-way, FC 5226 x 1	2048, FC 4447 x 1	2 x 36.4 GB (FC 3273)	ROM or RAM
1.65 GHz	2-way, FC 5229 x 1	2048, FC 4447 x 1	2 x 36.4 GB (FC 3273)	ROM or RAM

1.6 System racks

The Enterprise Rack Models T00 and T42 are 19-inch wide racks for general use with IBM @server p5, pSeries, and RS/6000 rack-based or rack drawer-based systems. The racks provide increased capacity, greater flexibility, and improved floor space utilization.

The p5-520, when featured, uses a 4U rack-mounted server drawer.

If a p5 system is to be installed in a non-IBM rack or cabinet, you should ensure that the rack conforms to the EIA² standard EIA-310-D (see 1.6.6, “OEM rack” on page 14).

² 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.

Note: It is the customer'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.

1.6.1 IBM RS/6000 7014 Model T00 Enterprise Rack

The 1.8-meter (71-inch) Model T00 is compatible with past and present p5, pSeries, and RS/6000 19-inch racks and is designed for use in all situations that have previously used the older rack models R00 and S00. 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 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.
- ▶ DC rack height is increased to 1926 mm (75.8 inches) 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 proper bays, but others can fit inside the rack. See 1.6.3, "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 p5-520, and other components, such as a 7311 I/O drawer, inside the rack. Servers can be programmed to illuminate the beacon in response to a detected problem or changes in system status.
- ▶ A rack status beacon junction box (FC 4693) should be used to connect multiple servers and I/O drawers 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).
- ▶ Weight:
 - T00 base empty rack: 244 kg (535 pounds)
 - T00 full rack: 816 kg (1795 pounds)

1.6.2 IBM RS/6000 7014 Model T42 Enterprise Rack

The 2.0-meter (79.3-inch) Model T42 is the rack that will address the special requirements of customers who want 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 the following:

- ▶ 42 EIA units (42U) of usable space
- ▶ AC power support only

- ▶ Weight:
 - T42 base empty rack: 261 kg (575 pounds)
 - T42 full rack: 930 kg (2045 pounds)

1.6.3 AC Power Distribution Unit and rack content

For rack models T00 and T42 9-outlet PDUs are available.

PDUs with nine outlets (FC 9176, 9177, 9178, 7176, 7177, and 7178) are available. A T42 rack configured for the maximum number of power outlets would have six PDUs (two mounted horizontally requiring 2U of rack space), for a total of 54 power outlets.

The p5-520 can be connected to any PDU that is available for the 7014-T00 or 7014-T42 racks.

For detailed power cords requirements and power cord feature codes, see the publication *Site and Hardware Planning Information*, SA38-0508. An online copy can be found at:

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

The first four PDUs ordered for a rack will be mounted vertically in the sides of the rack, occupying all the four PDU bays available. Any additional PDU will be mounted horizontally in the rear of the rack and will occupy 1U of rack space.

Note: Each p5-520, or a system drawer to be mounted in the rack, requires two power cords (with redundant power feature) that are not included in the base system order.

Universal PDU (FC 7188) and the optional Universal PDU to be mounted horizontally in the rack (FC 9188) will be available on December 31, 2004, supporting a wide range of country requirements and electrical power specifications. Each Universal PDU provides 12 C13 power outlets for use within a 7014-T00 or 7014-T42 rack, compared to nine C13 power outlets provided by FC 7176 or FC 7177 PDUs. Nine different power cord features can be used to connect the PDU to a wall power outlet. Each power cord provides the unique design characteristics for the different 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.

1.6.4 Rack-mounting rules for p5-520

The primary rules that should be followed when mounting the p5-520 into a rack are:

- ▶ The p5-520 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.
- ▶ 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 p5-520 into the service position, it is essential that the rack manufacturer's safety instructions have been followed regarding rack stability.

Depending on current implementation and future enhancements of additional 7311 Model D20 drawers connected to the p5-520 or single installed p5-520 systems, Table 1-8 on page 13 shows examples of minimum and maximum configurations for different combinations of p5-520s and attached 7311 Model D20 I/O drawers.

Table 1-8 Minimum and maximum configurations for p5-520s and 7311-D20s

	Only p5-520s	One p5-520, one 7311-D20	One p5-520, four 7311-D20s	One p5-520, eight 7311-D20s
7014-T00 rack	9	4	1	1
7014-T42 rack	10	5	2	1

1.6.5 Additional options for rack

The intention of this section is to highlight some solutions available to provide a single point of management for environments composed of multiple p5-520 servers or other IBM @server p5, pSeries, and RS/6000 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 select IBM @server p5, pSeries, and RS/6000 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:

- ▶ Digital Data Storage (DDS) Gen 5 DAT72 Tape Drive provides physical storage capacity of 36 GB (72 GB with 2:1 compression) per data cartridge.
- ▶ 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.
- ▶ Digital Data Storage (DDS-4) tape drive with 20 GB native data capacity per tape cartridge and a native physical data transfer rate of up to 3 MB/sec that uses a 2:1 compression so that a single tape cartridge can store up to 40 GB of data.
- ▶ 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 IPL³ (IBM-distributed CD-ROM media only) and program distribution.
- ▶ DVD-RAM drive with up to 2.7 MB/sec 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 pels⁴ 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 Space Saver 2, 14.5-inch keyboard that mounts in the rack keyboard tray and is available as a feature in 16 language configurations (the track point mouse is integrated into the keyboard).

³ Initial Program Load

⁴ Picture elements

Note: 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 a POWER GXT135P graphics accelerator (FC 2848 or FC 2849) to be installed in the server, or other graphic accelerator, if supported.

Hardware Management Console 7310 Model CR2

The 7310 Model CR2 is a 1U, 19-inch rack-mountable drawer supported in the 7014 Model T00 and T42 racks. The 7310 Model CR2 provides one serial port, two integrated Ethernet ports, and two additional PCI slots. The HMC 7310 Model CR2 has USB ports to connect USB keyboard and mouse devices.

Note: The HMC serial port can be used for external modem attachment if the Service Agent call-home function is implemented, and the Ethernet ports are used to communicate to the service processor in p5-520 systems. An Ethernet cable (FC 7801 or 7802) is required to attach the HMC to the p5-520 system it controls.

1.6.6 OEM rack

The p5-520 can be installed in a suitable OEM rack, provided that the rack conforms to the EIA-310-D standard. This standard is published by the Electrical Industries Alliance, and a summary of this standard is available in the publication *Site and Hardware Planning Information*, SA38-0508.

The key points mentioned in this standard are as follows:

- ▶ Any rack used must be capable of supporting 15.9 kg (35 pounds) per EIA unit (44.5 mm [1.75 inches] of rack height).
- ▶ To ensure proper rail alignment, the rack must have mounting flanges that are at least 494 mm (19.45 inches) across the width of the rack and 719 mm (28.3 inches) between the front and rear rack flanges.
- ▶ It might be necessary to supply additional hardware, such as fasteners, for use in some manufacturer's racks.

Figure 1-5 shows the drawing specifications for OEM racks.

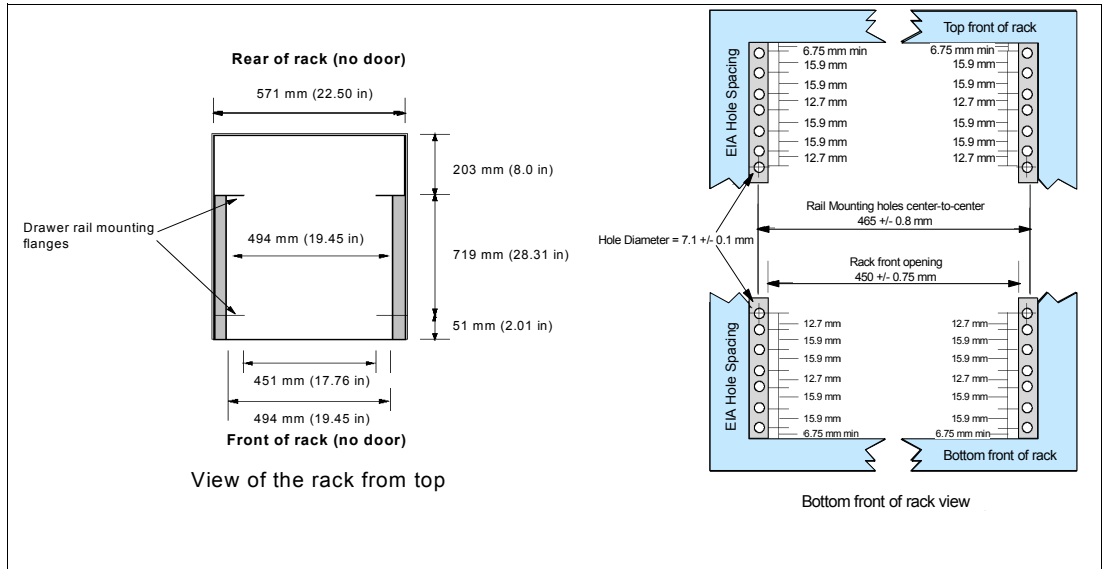


Figure 1-5 Reference drawing for OEM rack specifications

Architecture and technical overview

This chapter discusses the overall system architecture represented by Figure 2-1. The major components of this diagram are described in the following sections. The bandwidths provided throughout this section are theoretical maximums provided for reference. We always recommend that you obtain real-world performance measurements using production workloads.

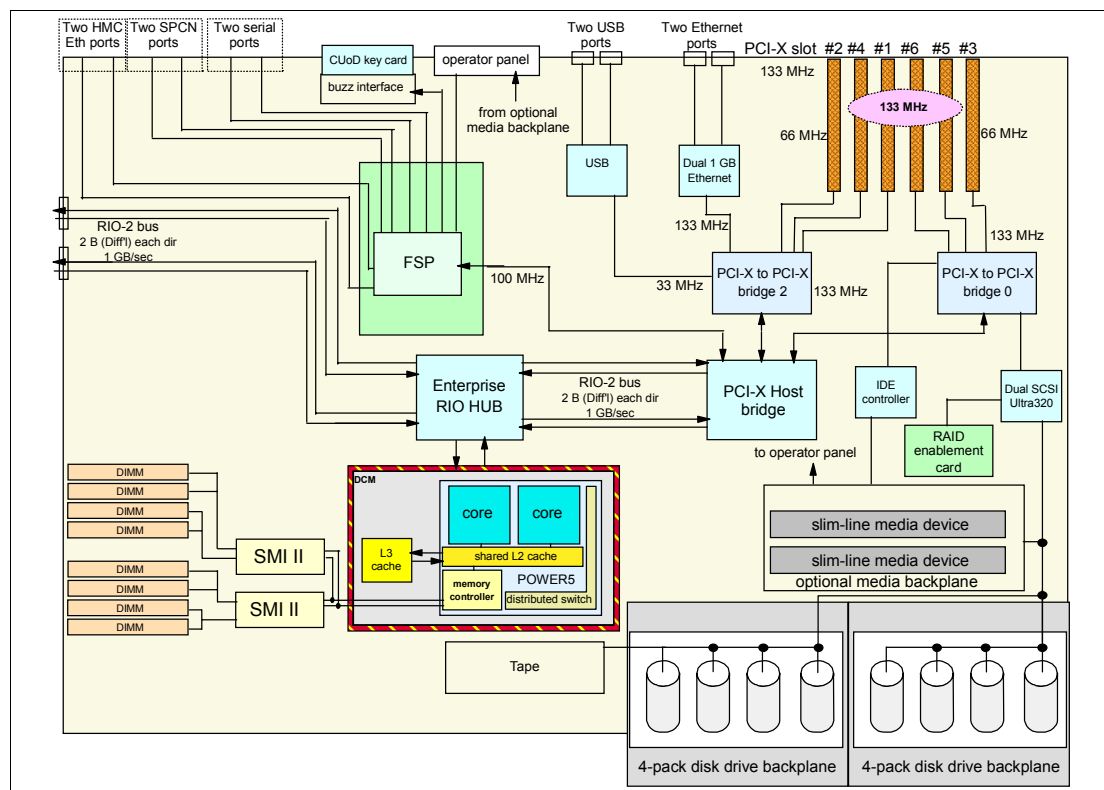


Figure 2-1 p5-520 logic data flow

2.1 The POWER5 chip

The POWER5 chip features single and multi-threaded execution, providing higher performance in the single-threaded mode than its POWER4 predecessor at equivalent frequencies provides. POWER5 maintains both binary and architectural compatibility with existing POWER4 systems to ensure that binaries continue executing properly and all application optimizations carry forward to newer systems. The POWER5 provides additional enhancements such as virtualization, reliability, availability, and serviceability at both chip and system levels.

Figure 2-2 shows the high-level structures of POWER4 and POWER5 processor-based systems. Though this discussion does not directly apply to a 1 or 2-way system with no chip-to-chip fabric, it does provide an overview of the POWER5 enhancements. The POWER4 scales up to a 32-way symmetric multiprocessor. Going beyond 32 processors increases interprocessor communication, resulting in higher utilization on the interconnection fabric bus. This greater contention negatively affects system scalability.

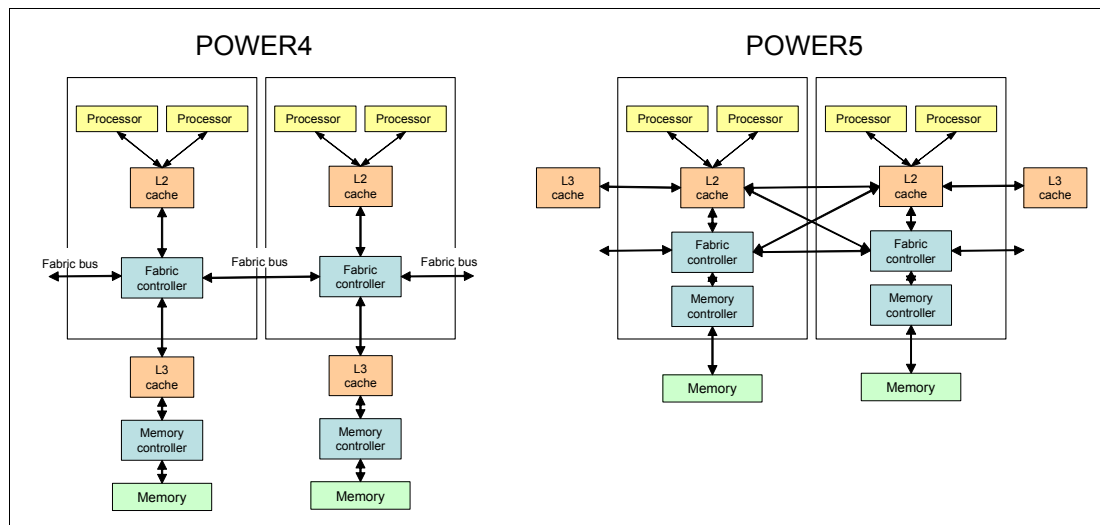


Figure 2-2 POWER4 and POWER5 system structures

Moving the L3 cache (2-way processors only) from inline with the memory to outboard provides significantly more cache on the processor side than previously available, thus reducing traffic on the fabric bus and allowing POWER5 processor-based systems to scale to higher levels of symmetric multiprocessing. The POWER5 supports a 1.9 MB on-chip L2 cache, implemented as three identical slices with separate controllers for each. Either processor core can independently access each L2 controller. The available L3 cache, with a capacity of 36 MB, operates as a backdoor with separate buses for reads and writes that operate at half processor speed.

Because of the higher transistor density of the POWER5 0.13- μm technology, it was possible to move the memory controller on chip and eliminate a chip previously needed for the memory controller function. These changes in the POWER5 also have the significant side benefits of reducing latency to the L3 cache and main memory, as well as reducing the number of chips necessary to build a system.

The POWER5 processor supports the 64-bit PowerPC® architecture. A single die contains two identical processor cores, each supporting two logical threads. This architecture makes the chip appear as a 4-way symmetric multiprocessor to the operating system. The POWER5

processor core has been designed to support both enhanced simultaneous multi-threading and single threaded (ST) operation modes.

2.1.1 Simultaneous multi-threading

As a requirement for performance improvements at the application level, simultaneous multi-threading functionality is embedded in the POWER5 chip technology. Applications developed to use process level parallelism (multi-tasking) and thread-level parallelism (multi-threads) can shorten their overall execution time. Simultaneous multi-threading is the next stage of processor saturation for throughput-oriented applications to introduce the method of instruction level parallelism to support multiple pipelines to the processor.

If simultaneous multi-threading is activated, on a 2-way POWER5 processor-based system, the operating system discovers the available processors as a 4-way system. To achieve a higher performance level, simultaneous multi-threading is also applicable in Micro-Partition, capped or uncapped, and dedicated in partition environments.

Simultaneous multi-threading is supported on POWER5 processor-based systems running AIX 5L Version 5.3 or Linux-based systems at a required 2.6 kernel. AIX provides the `smtctl` command that turns simultaneous multi-threading on and off without a subsequent reboot. For Linux, an additional boot option must be set to activate simultaneous multi-threading after a reboot.

The simultaneous multi-threading mode maximizes the usage of the execution units. In the POWER5 chip, more rename registers have been introduced (for floating-point operation, rename registers increased to 120), which are essential for out of order execution, and then vital for the simultaneous multi-threading.

Enhanced simultaneous multi-threading features

To improve simultaneous multi-threading performance for various workloads and provide robust quality of service, POWER5 provides two features:

- ▶ Dynamic resource balancing
 - The objective of dynamic resource balancing is to ensure that the two threads executing on the same processor flow smoothly through the system.
 - Depending on the situation, the POWER5 processor resource balancing logic has different thread throttling mechanisms (a thread reached threshold of L2 cache misses and will be throttled to allow other threads to pass the stalled thread).
- ▶ Adjustable thread priority
 - Adjustable thread priority lets software determine when one thread should have a greater (or lesser) share of execution resources.
 - The POWER5 supports eight software-controlled priority levels for each thread.

ST operation

Not all applications benefit from simultaneous multi-threading. Having threads executing on the same processor will not increase the performance of applications with execution unit limited performance or applications that consume all the chip's memory bandwidth. For this reason, the POWER5 supports the ST execution mode. In this mode, the POWER5 processor gives all the physical resources to the active thread, allowing it to achieve higher performance than a POWER4 based-system at equivalent frequencies. Highly optimized scientific codes are one example where ST operation is ideal.

2.1.2 Dynamic power management

In current Complementary Metal Oxide Semiconductor (CMOS) technologies, chip power is one of the most important design parameters. With the introduction of simultaneous multi-threading, more instructions execute per cycle per processor core, thus increasing the core's and the chip's total switching power. To reduce switching power, POWER5 chips use a fine-grained, dynamic clock gating mechanism extensively. This mechanism gates off clocks to a local clock buffer if dynamic power management logic knows the set of latches driven by the buffer will not be used in the next cycle. This allows substantial power saving with no performance impact. In every cycle, the dynamic power management logic determines whether a local clock buffer that drives a set of latches can be clock gated in the next cycle.

In addition to the switching power, leakage power has become a performance limiter. To reduce leakage power, the POWER5 chip uses transistors with low threshold voltage only in critical paths. The POWER5 chip also has a low-power mode, enabled when the system software instructs the hardware to execute both threads at the lowest available priority. In low power mode, instructions are dispatched once every 32 cycles at most, further reducing switching power. The POWER5 chip uses this mode only when there is no ready task to run on either thread.

2.1.3 The POWER chip evolution

The p5-520 system complies with the RS/6000 platform architecture, which is an evolution of the PowerPC Common Hardware Reference Platform (CHRP) specifications. Figure 2-3 on page 20 shows the POWER evolution of the IBM UNIX server.

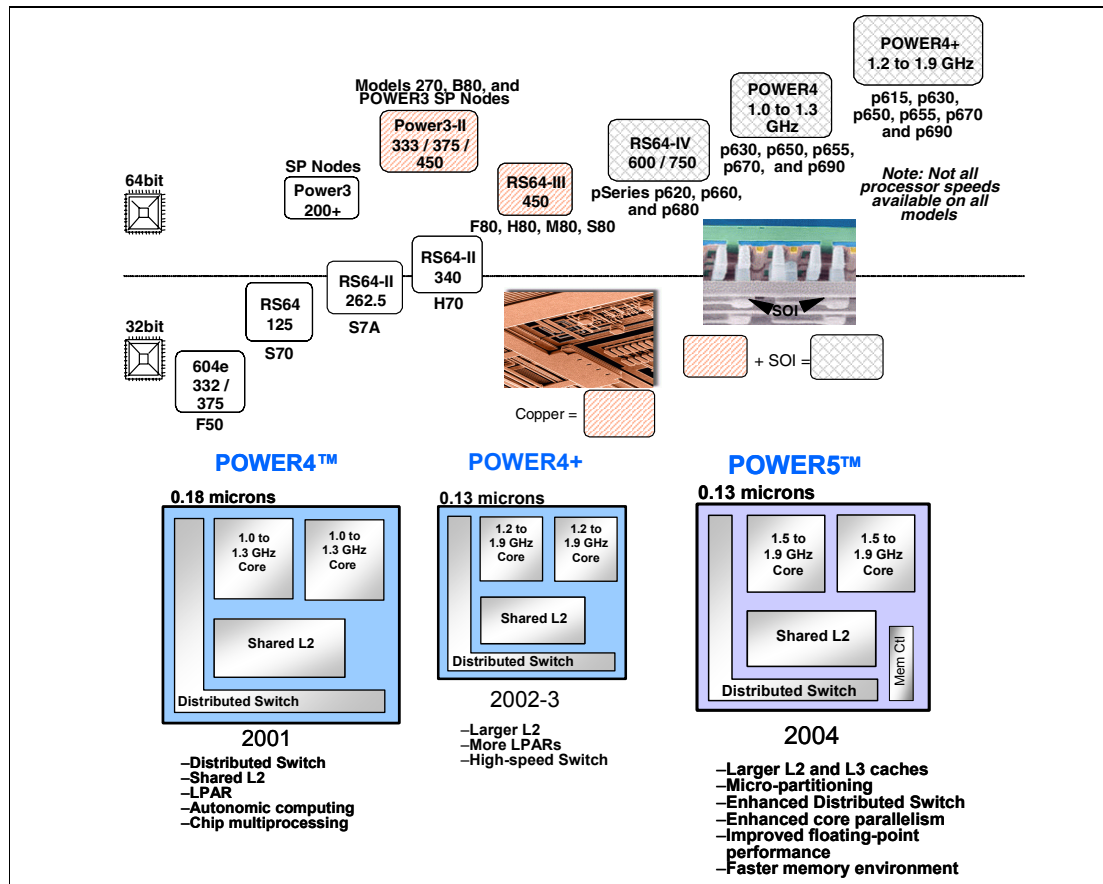


Figure 2-3 The POWER chip evolution

2.1.4 CMOS, copper, and SOI technology

The POWER5 processor design is a result of a close collaboration between IBM Systems Group and IBM Microelectronics technologies that enables IBM @server p5 systems to give customers improved performance, reduced power consumption, and decreased IT footprint size through logical partitioning. The POWER5 processor chip takes advantage of IBM leadership technology. It is made using IBM 0.13- μm -lithography Complementary Metal Oxide Semiconductor (CMOS) technology. The POWER5 processor also uses silicon-on-insulator (SOI) technology to allow a higher operating frequency for improved performance yet with reduced power consumption and improved reliability compared to processors not using this technology.

2.2 Processor and cache

In the p5-520 system, the POWER5 chip has been packaged with the L3 cache chip (on 2-way models) into a cost-effective Dual Chip Module (DCM) package. In p5-520 systems, the DCM is directly soldered to the system planar. The storage structure for the POWER5 chip is a distributed memory architecture which provides high memory bandwidth. Although each processor can address all memory and sees a single shared memory resource. The DCM and its optional L3 cache are directly soldered to the system planar. They are interfaced to eight memory slots, controlled by two SMI-2 controllers, which are located in close physical proximity to the DCM. The p5-520 supports one processor core (the core is either a 1-way or 2-way) and optional integrated 36 MB L3 cache module. I/O connects to the p5-520 I/O subsystem using the GX+ bus. The DCM provides a single GX+ bus. The GX+ bus provides an interface to a single device such as the RIO-2 buses.

The processor core contains a single DCM and the local memory storage subsystem for that DCM. See Figure 2-4 for a POWER5 processor core layout view.

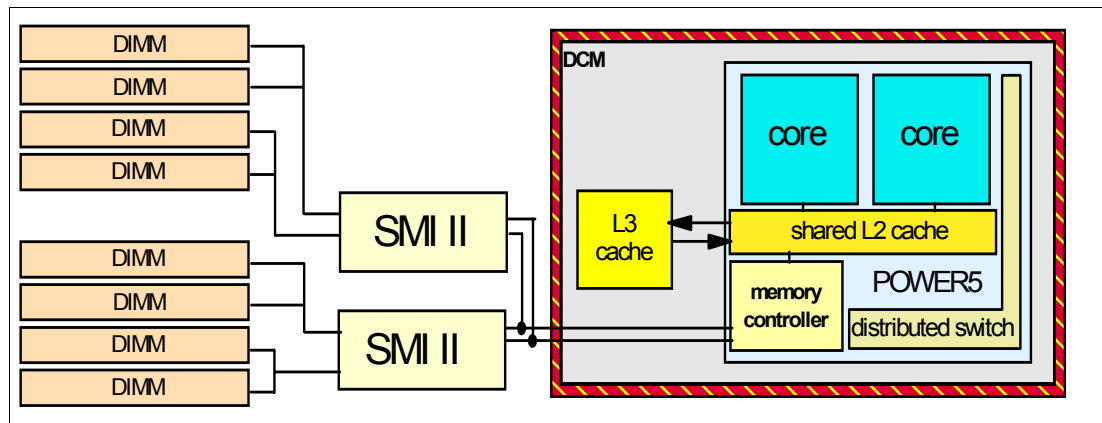


Figure 2-4 2-way processor core with DDR1 memory socket layout view

2.2.1 Available processor speeds

The p5-520 operates only at a processor clock rate of 1.5 GHz or 1.65 GHz for 2-way systems and 1.5 GHz for 1-way processor-based systems.

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

lsattr -El procX Where X is the number of the processor, for example, proc0 is the first processor in the system. The output from the command¹ would be

similar to the following (False, as used in this output, signifies that the value cannot be changed through an AIX command interface):

```
type powerPC_POWER5      Processor type      False
frequency 165600000      Processor Speed     False
smt_enabled true         Processor SMT enabled False
smt_threads 2            Processor SMT threads False
state enable             Processor state     False
```

pmcycles -m

This command (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 is the output of a 2-way p5-520 running at 1.65 GHz system and simultaneous multi-threading functionality enabled:

```
Cpu 0 runs at 1656 MHz
Cpu 1 runs at 1656 MHz
Cpu 2 runs at 1656 MHz
Cpu 3 runs at 1656 MHz
```

Note: The `pmcycles` command is part of the `bos.pmap` fileset. First check whether that component is installed using the `lspp -l bos.pmap` command.

2.3 Memory subsystem

The p5-520 system offers pluggable DIMMs for memory. The system planar provides eight slots for up to eight pluggable DIMMs. The minimum memory for a p5-520 1.5 GHz processor-based system is 512 GB and 32 GB as maximum installable memory option. Figure 2-5 shows the offerings and memory slot availability.

2.3.1 Memory placement rules

The memory features available at the time of writing for the p5-520 are listed in 1.4.2, “Memory features” on page 7.

Each memory feature consists of four DIMMs, or quad, and must be installed according to Figure 2-5. The first quad slots are J0A, J1A, J0C, and J1C, and for the second quad, the slots are J0B, J1B, J0D, and J1D. The 512 MB memory feature must be installed in J0A and J1A.

Note: A quad must consist of a single feature (that is made of identical DIMMs). Mixed DIMM capacities in a quad will result in reduced RAS.

¹ The output of the `lsattr` command has been expanded with AIX 5L™ to include the processor clock rate.

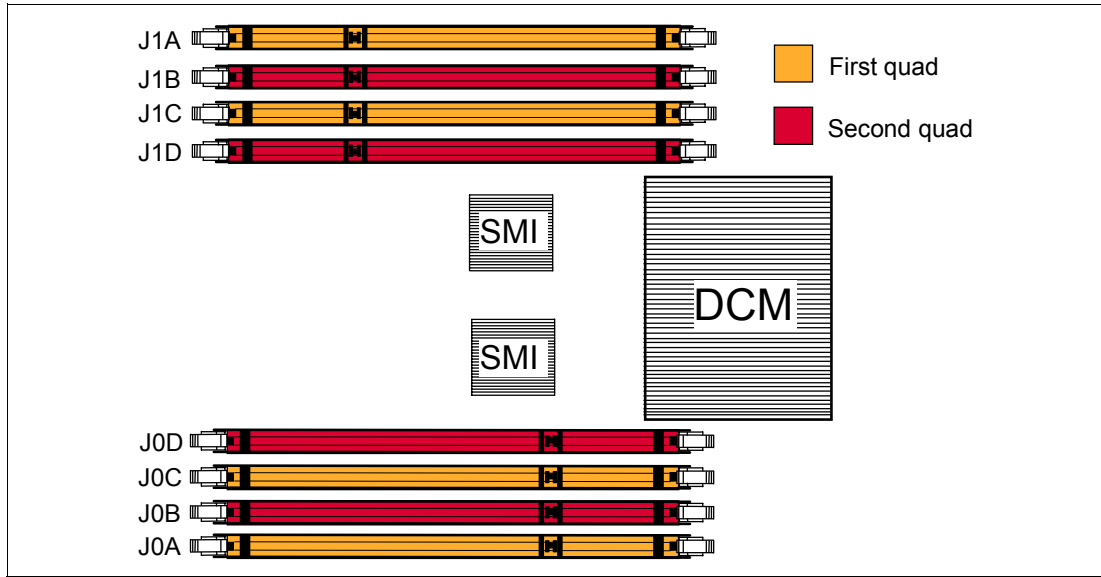


Figure 2-5 Memory placement for the p5-520

2.3.2 Memory restriction

The p5-520 does not support OEM memory, and there is no exception to this rule. OEM memory is never certified for the use in pSeries and the new p5 servers. If the p5-520 is populated with OEM memory, you could experience unexpected and unpredictable behavior, especially when the system is planned to use Micro-Partitioning technology.

All IBM memory is identified by an IBM logo and a white label printed with a barcode on top and an alphanumeric string on the bottom, created according to the rule reported in Figure 2-6.

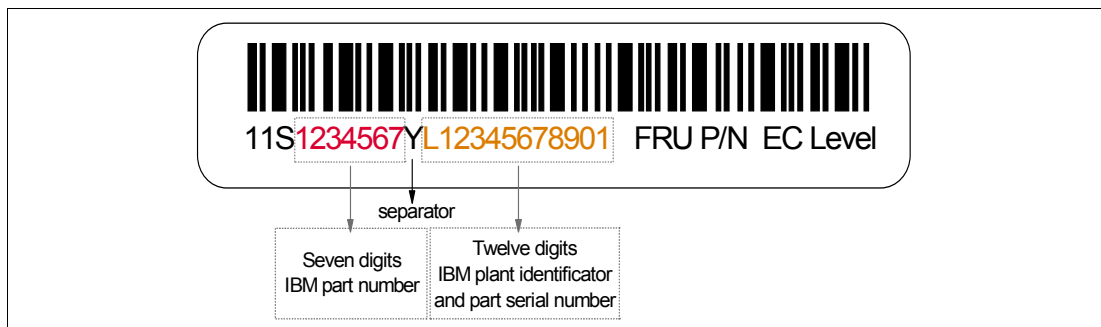


Figure 2-6 IBM memory certification label

Sometimes, OEM vendors put a label reporting the IBM memory part number but not the barcode or the alphanumeric string, or both, on their DIMMs.

In case of system failure caused by OEM memory installed in the system, the first thing to do is to replace the suspected memory with IBM memory and check whether the problem is corrected. Contact your IBM representative for further assistance if needed.

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 chip, buffers reads and writes to and from memory and the processor. There are two SMIs, each with a single 8 byte read and 2 byte write DDR bus to the processor. A DDR bus allows double reads or writes per clock cycle. If 266 MHz memory is installed (operating at 266.5 MHz), the throughput is $(16 \times 2 \times 266.5) + (4 \times 2 \times 266.5)$ or 10660 MB/second or 10.41 GB/second between the processor and memory controller. These values are maximum theoretical throughputs for comparison purposes only.

There are four 8 Byte paths to the memory DIMMs from the SMIs, therefore the throughput is 8.32 GB/s.

The POWER5 processor's integrated memory controller further reduces latency over the previous outboard controller on POWER4 systems to the SMI chips by requiring fewer cycles in order to set up memory addressing in the hardware.

2.4 System buses

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

2.4.1 RIO buses and GX card

The DCM provides a GX+ bus that is used to connect to the I/O subsystem. The p5-520 provides two external RIO-2 ports that can operate up to 1 GHz. The RIO-2 ports are used for I/O expansion to external I/O drawers. The only supported I/O drawer that can be connected to the p5-520 is the 7311 Model D20.

2.5 Internal I/O subsystem

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. There is an internal RIO-2 bus 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 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 two PCI-X to PCI-X bridges to the primary PCI-X buses on the PCI-X Host bridge chip.

PCI-X slots 5 and 6 can accept short PCI-X or PCI cards. The remaining PCI-X slots are full length cards. The dual 1 Gb Ethernet adapter is integrated on the system planar.

2.5.1 PCI-X slots and adapters

PCI-X, where the X stands for extended, is an enhanced PCI bus, delivering a bandwidth of up to 1 GB/sec, running a 64-bit bus at 133 MHz. PCI-X is backward compatible, so the p5-520 systems can support existing 3.3 volt PCI adapters.

The PCI-X slots in the p5-520 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 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.

64-bit and 32-bit adapters

IBM offers 64-bit adapter options for the p5-520, 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 p5-520 systems, and for important information regarding adapter placement, see the IBM @server Hardware Information Center. You can find it at:

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

2.5.2 LAN adapters

When a p5-520 is connected to a local area network (LAN), the internal dual port 10/100/1000 Mbps RJ-45 Ethernet controller, integrated on the system planar can be used.

See the Table 2-1 for the list of additional LAN adapters available at the time of writing. IBM supports an installation with NIM using Ethernet and token-ring adapters (CHRP² is the platform type).

Table 2-1 Available LAN adapters

Feature code	Adapter description	Slot	Size	Max
4959	4/16 Token-Ring	32 or 64	short	4
4962	10/100 Ethernet	32 or 64	short	6
5700	Gigabit Ethernet	64	short	4
5701	10/100/1000 Ethernet	64	short	4
5706	2-port 10/100/1000 Ethernet	64	short	4
5707	2-port Gigabit Ethernet - SX	64	short	4
5718	10 Gigabit Ethernet PCI-X	64	short	1

2.5.3 Graphic accelerators

The p5-520 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.

2.5.4 Audio adapter

The p5-520 supports an audio PCI adapter (FC 8244). It is a 3.3 volt, 32-bit PCI adapter that runs at 33 MHz and requires one short 32-bit or 64-bit PCI-X slot. The adapter provides external jacks for headphones, speaker output, line input, microphone input, and an internal connector for CD or DVD drive audio input.

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

2.5.5 SCSI adapters

To connect to external SCSI devices, the following adapters provided in Table 2-2 are available, at the time of writing, to be used in p5-520 system.

Table 2-2 Available SCSI adapters

Feature code	Adapter description	Slot	Size	Max
5703	Ultra320 SCSI RAID (bootable)	64	long	3
5712	Ultra320 SCSI	64	short	4
6204	Ultra SCSI Differential	32	short	2

There is also the option to make the internal Ultra320 SCSI channel externally accessible on the rear side of the system by installing FC 4270. No additional SCSI adapter is required in this case. If FC 4270 is installed, a second 4-pack disk enclosure (FC 6574 or FC 6594) cannot be installed, which limits the maximum number of internal disks to four. FC 4270 also requires one PCI-X slot.

2.6 Internal serial ports

The serial 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 support the attachment of serial console and modem.

If an HMC is connected, a *virtual serial console* is provided by the HMC (logical device `vsa0` under AIX visible with the `lsdev -l vsa0` command). When the HMC is connected, the S1 and S2 ports are not usable by applications, such as for an HACMP heartbeat.

If additional serial port functionality is needed, optional PCI adapters are available (PCI 8-port adapter FC 2943 or PCI 128-port adapter FC 2944).

2.7 Internal storage

There is one dual channel Ultra320 SCSI controller managed by the EADS-X chips, integrated into the system planar, that are used to drive the internal disk drives. The eight internal drives plug into the disk drive backplane, which has two separate SCSI buses with four disk drives per bus.

The internal disk drive can be used in two different modes based on whether the SCSI RAID Enablement Card (FC 5709) is installed (see 2.7.3, "RAID options" on page 28).

The p5-520 supports two 4-pack disk drives using a backplane that 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 media devices

The p5-520 provides two slim-line media bays for optional DVD-ROM (FC 2640) and optional DVD-RAM (FC 5751), and one media bay for a tape drive.

Table 2-3 Available tape drives

Feature code	Description
6258	4-mm 36/72 GB tape (LVD)
6134	8-mm 60/150 GB tape (LVD)
6120	VXA 80/160 GB tape (LVD)

2.7.2 Internal hot swappable SCSI disks

The p5-520 can have up to eight hot-swappable disk drives plugged in the two 4-pack disk drives backplanes. The hot-swap process is controlled by the SCSI enclosure service (SES), which is located in the 4-pack disk drives backplane (AIX assigns the name `ses0` to the first 4-pack, and `ses1` to the second, if present). The two hot-swappable 4-pack disk drives backplanes can accommodate the devices listed in Table 2-4.

Table 2-4 Hot-swappable disk drive options

Feature code	Description
3273	36.4 GB 10,000 RPM Ultra3 SCSI hot-swappable disk drive
3277	36.4 GB 15,000 RPM Ultra3 SCSI hot-swappable disk drive
3274	73.4 GB 10,000 RPM Ultra3 SCSI hot-swappable disk drive
3278	73.4 GB 15,000 RPM Ultra3 SCSI hot-swappable disk drive
3275	146.8 GB 10,000 RPM Ultra3 SCSI hot-swappable disk drive

At the time of writing, if a new order is placed with two 4-pack DASD backplanes (FC 6574 or FC 6595) and more than one disk, the system configuration shipped from manufacturing will balance the total number of SCSI disks between the two 4-pack SCSI backplanes. This is for manufacturing test purposes, and not because of any limitation. Having the disks balanced between the two 4-pack DASD backplanes allows the manufacturing process to systematically test the SCSI paths and devices related to them.

Prior to the hot-swap of a disk 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:

1. Release the tray handle on the disk.
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.
4. Now you can safely remove the disk from the 4-pack 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 issued by the SES to reset all processes when a drive is inserted, and it is not an issue.

Hot-swap disks and Linux

Linux does not support the hot-swap of any disk drive at the time of writing; therefore, the Linux operating system does not support these hot-swappable procedures. A p5-520 system running Linux must be shut down and powered off before you replace any disk drives.

2.7.3 RAID options

Internal hardware RAID is available on the p5-520. Three options are available:

- ▶ Install the Dual Channel SCSI RAID Enablement Card (FC 5709). Install four disk drives in the first 4-pack DASD backplane (FC 6574). This will allow RAID 0, 5, or 10 capabilities within a single 4-pack of DASD with one RAID controller.
- ▶ Install FC 5709. Install a second FC 6574. Install four additional disk drives in the second 4-pack DASD backplane. This will allow RAID 0, 5, or 10 capabilities across two 4-packs of DASD with one RAID controller.
- ▶ Install feature number 5709. Install the Ultra320 SCSI 4-Pack Enclosure for Disk Mirroring (FC 6594). Install the PCI-X Dual Channel Ultra320 SCSI RAID Adapter (FC 5703). Install the SCSI Cable, which connects the PCI Adapter to the second 4-pack DASD backplane (FC 4267). This will allow RAID 0, 5, or 10 capabilities within each 4-pack of DASD with two RAID controllers.

Note: Because the p5-520 has eight disk drive slots, customers performing upgrades must perform appropriate planning to ensure the correct handling of their RAID arrays.

2.8 External I/O subsystem

This section describes the external I/O subsystem, the 7311 D20 I/O drawer that is the only drawer supported on the p5-520 system.

2.8.1 I/O drawers

As described in Chapter 1, “General description” on page 1, the p5-520 system has six internal PCI-X slots, which is enough in many cases. If more PCI-X slots are needed to dedicate more adapters to a partition or to increase the bandwidth of network adapters, up to four 7311 Model D20 I/O drawers can be added to the p5-520 system.

The p5-520 system has a standard RIO-2 bus to connect the internal PCI-X slots through the PCI-X to PCI-X bridges and support up to four external I/O drawers.

The 7311 Model D20 I/O drawer must have the RIO-2 loop adapter (FC 6417) to be connected to the p5-520 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 GB/sec is provided by each of the buses. To avoid overloading an I/O drawer, the recommendation in the IBM *@server* Hardware Information Center should be followed. You can find it at:

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

Figure 2-7 on page 29 shows a conceptual diagram of the 7311 Model D20 I/O drawer subsystem.

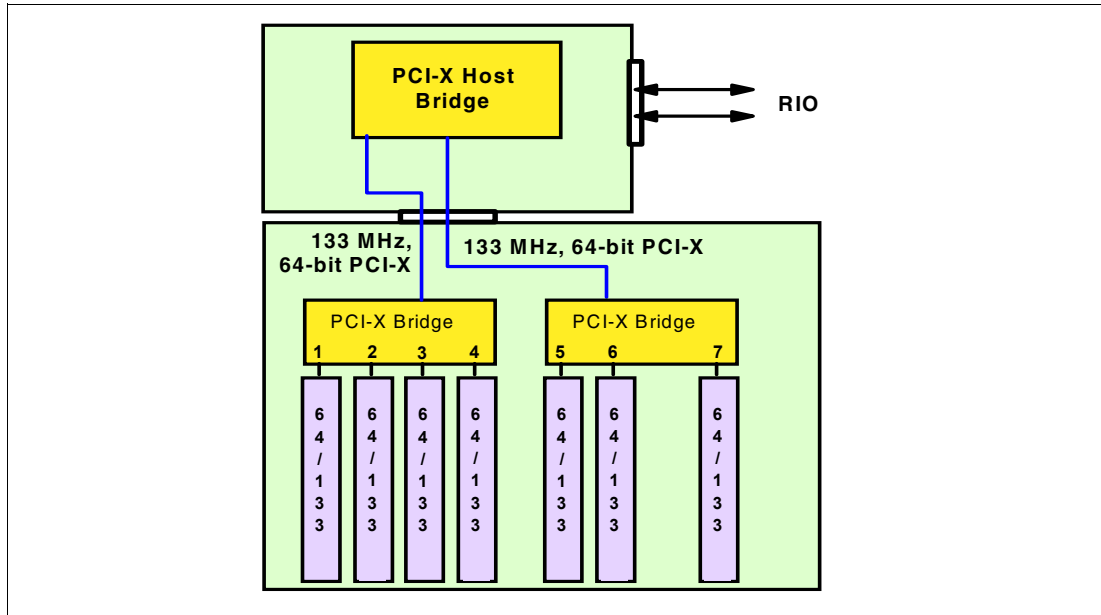


Figure 2-7 Conceptual diagram of the 7311-D20 I/O drawer

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 (that can have 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 2-8).

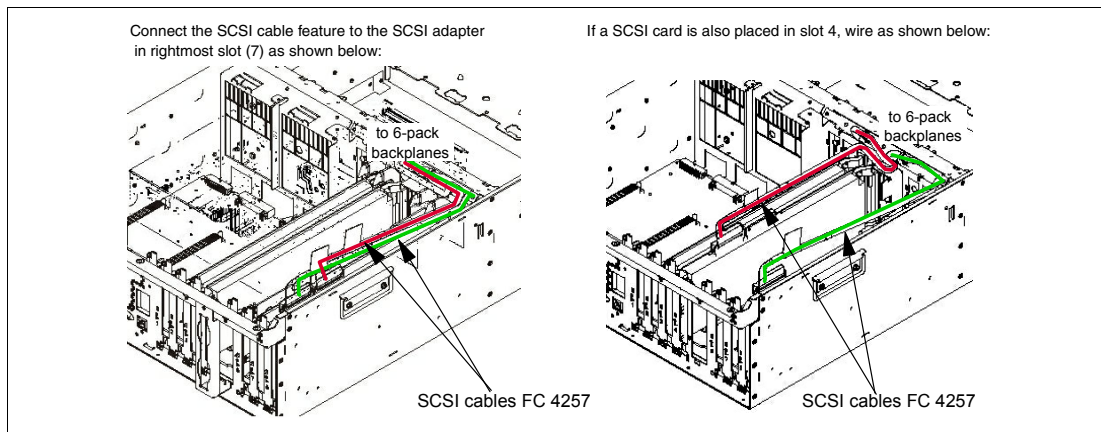


Figure 2-8 7311 Model D20 internal SCSI cabling

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, both 6-packs can be assigned only to the same partition. When the server is configured with the The Advanced POWER Virtualization hardware feature and the Virtual I/O Server is used for virtual SCSI, the disks can be shared between partitions.

2.8.2 7311 I/O drawer RIO-2 cabling

As described in 2.8, “External I/O subsystem” on page 28, you can connect up to four I/O drawers in the same loop to the p5-520 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 GB/s per I/O drawer in double barrel mode.

There is one default primary RIO-2 loop in any p5-520 system. This feature provides two Remote I/O ports for attaching up to four 7311 Model D20 I/O drawers to the system in a single loop. Figure 2-9 shows how you could connect four I/O drawers to one p5-520 system.

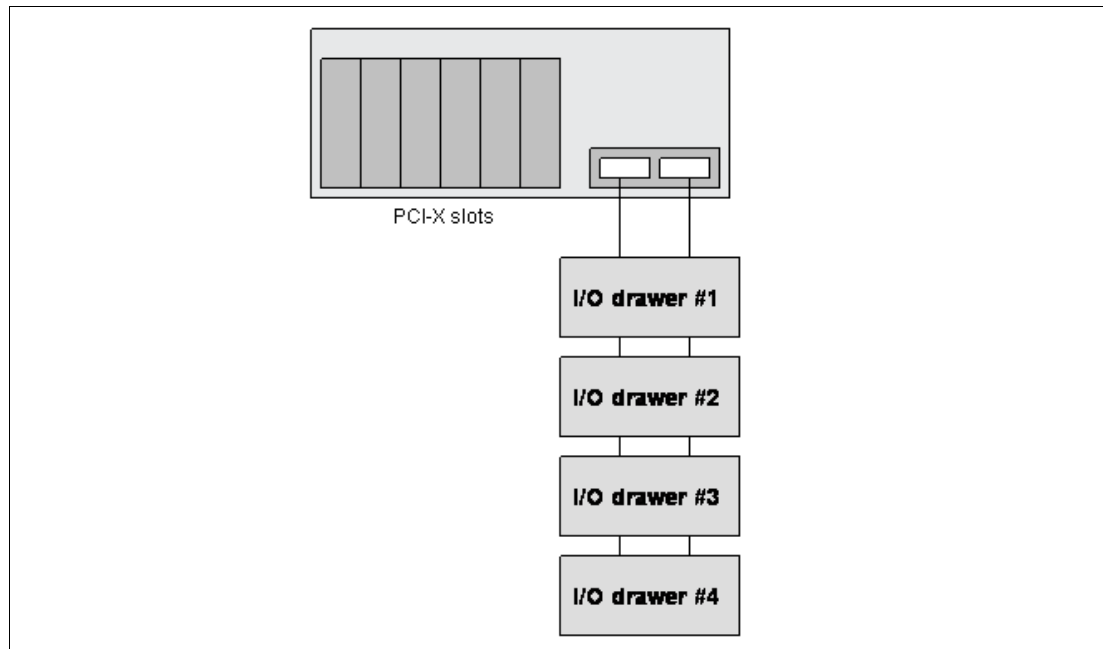


Figure 2-9 RIO-2 connections

The RIO-2 cables used have different lengths to satisfy the different connection requirements:

- ▶ Remote I/O cable, 3.5 m (FC 3147)
- ▶ Remote I/O cable, 10 m (FC 3148)

2.8.3 7311 Model D20 I/O drawer SPCN cabling

The SPCN is used to control and monitor the status of power and cooling within the I/O drawer. The SPCN is a loop, the cabling starts from SPCN port 0 on the p5-520 system 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 the port 1 of p5-520 system. If you have more than one I/O drawer, you continue the loop connecting the following drawer (or drawers) with the same rule. See Figure 2-10 on page 31.

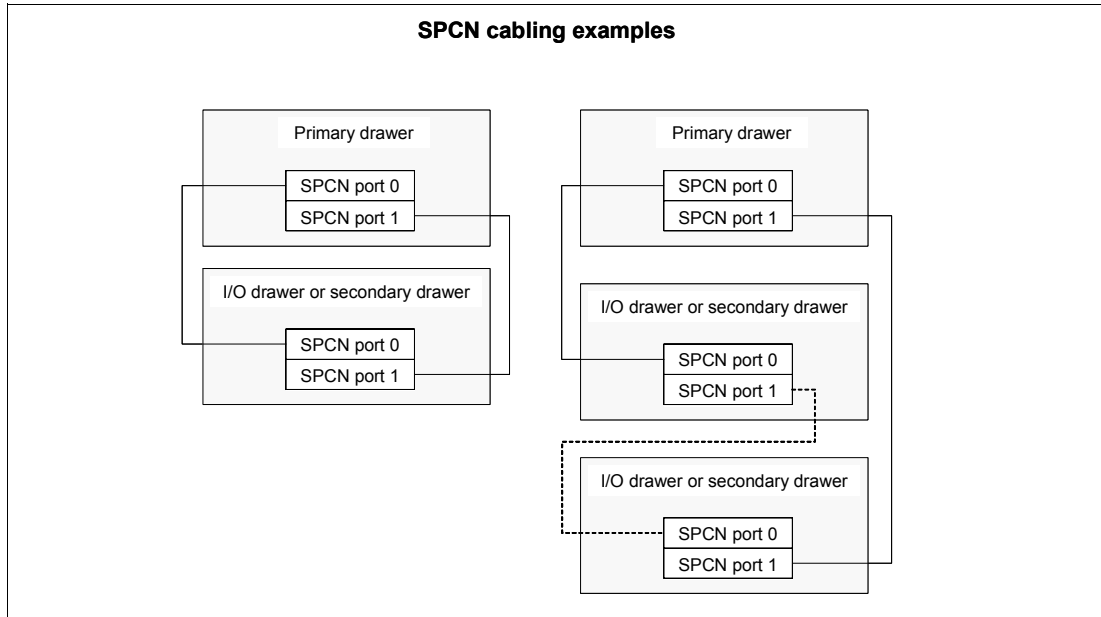


Figure 2-10 SPCN cabling examples

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)

2.8.4 External disk subsystem

The p5-520 system has internal hot-swappable drives. Internal disks are usually used for the AIX rootvg and paging space, or a RAID configuration when the SCSI RAID Enablement Card is featured. Specific customer requirements can be satisfied with the several external disks possibilities that the p5-520 supports.

IBM 2104 Expandable Storage Plus

The IBM 2104 Expandable Storage Plus Model DS4 is a low-cost 3U disk subsystem that supports up to 14 Ultra320 SCSI disks from 18.2 GB up to 146.8 GB, at the time this publication was written. This subsystem can be used in splitbus mode, meaning the bus with 14 disks could be split into two buses with seven disks each. In this configuration, two additional LPARs (using dedicated devices) could be provided with up to seven disks for rootvg by using one Ultra3 SCSI adapter (FC 5712) for each LPAR.

For further information about the IBM 2104 Expandable Storage Plus subsystem, visit the following Web site:

<http://www.storage.ibm.com/hardsoft/products/expplus/expplus.htm>

IBM TotalStorage FAST Storage servers

The IBM® TotalStorage® FAST Storage server family consists of five models: Model 100, 600, 700, and 900. The Model 100 is the smallest model, which scales up to 14 TB, and Model 900 is the largest, which scales up to 32 TB of disk storage, at the time this publication was written. Model 600 provides up to 16 bootable partitions that are attached with the

Gigabit Fibre Channel adapter (FC 5716). Model 700 provides up to 64 bootable partitions. In most cases, both the FAStT Storage server and the p5-520 or the 7311 Model D20 I/O drawers are connected to a storage area network (SAN). If only space for the rootvg is needed, the FAStT Model 100 is a good solution.

For support of additional features and for further information about the FAStT family, refer to the following Web site:

<http://www.storage.ibm.com/hardsoft/disk/fastt/index.html>

IBM TotalStorage Enterprise Storage Server

The IBM TotalStorage Enterprise Storage Server® (ESS) is the high-end premier storage solution for use in storage area networks. The 2105 Model 800 provides from 582 GB up to 55.9 TB of usable disk capacity. An ESS system can also be used to provide disk space for booting LPARs or partitions using Micro-Partitioning technology. An ESS is usually connected to a SAN to which the p5-520 is also connected by using Gigabit Fibre Channel adapters (FC 6239).

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

<http://www.storage.ibm.com/hardsoft/products/ess/index.html>

2.9 Dynamic logical partitioning

Introduced with the POWER4 processor product line and the AIX 5L Version 5.1 operating system, the logical partition (LPAR) became available. This technology offered the capability to divide a pSeries system into separate systems, where each LPAR runs an operating environment on dedicated attached devices, such as processors, memory, and I/O components. The customer requested system flexibility to change the system topology on demand, was achieved by modifying the system layout on the required HMC.

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 V5.2 with all the necessary enhancements to enable dynamic LPAR was introduced in 2002. This requires an attached HMC, with the proper level of software, to control the system resources and an updated system firmware level to electronically isolate systems resources. 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.

Dynamic logical partitioning is supported by the following levels of the AIX and Linux operating systems:

- ▶ AIX 5L for POWER V5.2, or later
- ▶ SUSE LINUX Enterprise Server 9, or later

It is not supported by current version of Red Hat Enterprise Linux AS for POWER Version 3.

USB resources form a group. Slimline devices form another group. A single group must be allocated to a single partition. In a base configuration, each 4-pack is connected to one of the two ports on the integrated SCSI controller. To an LPAR, the entire SCSI controller (including all disks attached to both ports) will be seen as P1-T10, and therefore can only be assigned to one active LPAR at a time. To provide additional internal drives for a second LPAR, either virtual I/O or an optional SCSI PCI adapter feature should be used. If a PCI SCSI adapter is featured, it will be connected to one of the 4-packs allowing it to be assigned to a partition independent of the 4-pack attached to the integrated SCSI adapter.

2.10 Virtualization

On the p5-520 server, logical partitions requiring dedicated resources may now be able to take advantage of a new technology that allows resources to be virtualized, allowing for a better overall balance of global system resources and their effective utilization.

2.10.1 Virtual Ethernet

To enhance intercommunication between partitions, including those using Micro-Partitioning technology, the Virtual Ethernet implementation allows in-memory connections at a high bandwidth from partition to partition. Virtual Ethernet working on LAN technology allows a transmission speed in the range of 1 to 3 GB/sec depending on the MTU³ size. A partition supports 256 Virtual Ethernet connections, where a single Virtual Ethernet resource can be connected to another Virtual Ethernet, a real network adapter, or both in a partition.

2.10.2 Advanced POWER Virtualization feature

The Advanced POWER Virtualization feature is an optional additional cost hardware feature that is available on all IBM *eServer* POWER5 processor-based systems. Each system has a unique feature code for this feature. For the p5-520 server, select FC 7940 to order the Advanced Virtualization feature.

The Advanced POWER Virtualization feature includes:

- ▶ Firmware enablement for Micro-Partitions
- ▶ Installation image for the Virtual I/O Server software that supports:
 - Ethernet adapter sharing
 - Virtual SCSI Server
- ▶ Partition Load Manager:
 - Automated CPU and memory reconfiguration
 - Real-time partition configuration and load statistics
 - Graphical user interface

Micro-Partitioning technology

Based on the partitioning concepts of a stable and well-known mainframe technology and existing LPAR/dynamic LPAR implementation on POWER4 and POWER4+ servers, the POWER5 systems introduce an enhanced partitioning model available as a hardware feature.

IBM Micro-Partitioning technology offers a method of sharing system resources. In POWER5 processor-based systems, physical resources are abstracted into virtual resources that are available to partitions. Resources include processor, Ethernet, and SCSI.

POWER5 Micro-Partitioning technology specifies processor capacity in processing units. One processing unit represents 1% of one physical processor. A partition defined with 220 processing units is equivalent to the power of 2.2 physical processors. Creating a partition, the minimum capacity is 10 processing units, or a 1/10 of a physical processor. A maximum of 10 partitions using IBM Micro-Partitioning technology for each physical processor can be defined. A total of 20 partitions can be created on a p5-520 system, but on a loaded system, the practical limit is less.

³ Maximum transmission unit

Partitions can also be defined with the capped and uncapped concept attributes. A capped partition is not allowed to exceed the defined share, while an uncapped partition is allowed to consume additional capacity with fewer restrictions. Uncapped partitions can be configured to the total idle capacity of the server or a percentage of it. Configuration through the HMC menu sets the allowed share and the capped or uncapped attribute.

The POWER5 processor-based systems use the POWER Hypervisor, which is the new Hypervisor that supports IBM Micro-Partitioning technology. The Hypervisor of existing POWER4 processor-based systems is working on a demand basis, as the result of machine interrupts and callbacks to the operating system. The new Hypervisor is active.

The Advanced POWER Virtualization Feature, described in 2.10.2, “Advanced POWER Virtualization feature” on page 33, facilitates the understanding of all the POWER5 and POWER Hypervisor enhancements to reach the highest level of granularity of installed system resources.

See Figure 2-11 on page 34 for a summary of the Micro-Partitioning LPAR model.

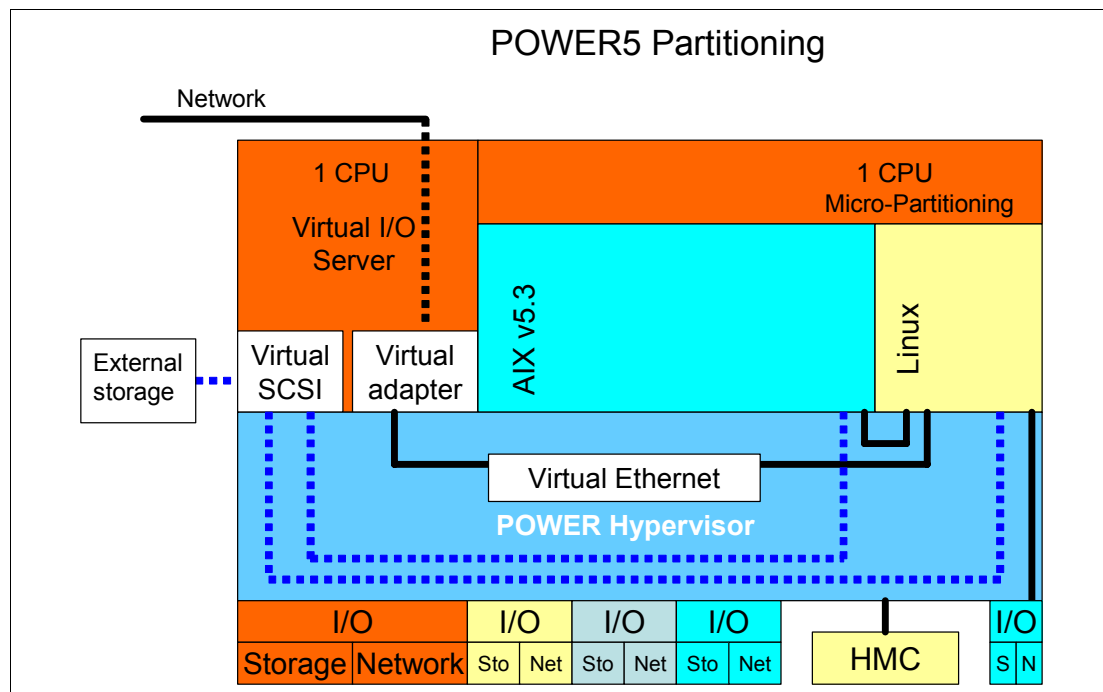


Figure 2-11 Micro-Partitioning LPAR organizational view

Virtual I/O Server

The Virtual I/O Server is a special purpose partition to provide virtual I/O resources to client partitions. The Virtual I/O Server will own the real resources that will be shared with the other clients. The Virtual I/O technology allows a physical adapter assigned to a partition to be shared by one or more partitions, enabling clients to minimize their number of physical adapters. The Virtual I/O Server will be used to reduce costs by eliminating the requirement that each partition has a dedicated network adapter, disk adapter, and disk drive.

It is preferred that you install the Virtual I/O Server in a partition with dedicated resources to help ensure stable performance.

Note: To maximize the performances of I/O intensive applications, dedicated physical adapters should be preferred in dedicated partitions.

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

Shared Ethernet adapter

A shared Ethernet adapter is a new service that acts as a layer 2 network switch to route network traffic from a Virtual Ethernet to a real network adapter. The shared Ethernet adapter must run in a Virtual I/O Server partition.

The advantage of using the Virtual Ethernet services is that partitions can communicate outside the system without having a physical network adapter attached to the partition. At the time of writing, up to 16 Virtual Ethernet x 18 VLANs can be shared on a single network interface. The amount of network traffic will limit the number of client partitions served through a single network interface.

Virtual SCSI

Access to real storage devices is implemented through the Virtual SCSI services, a part of the Virtual I/O Server partition. Logical volumes created and exported on the Virtual I/O Server partition will be shown at the Virtual Storage Client partition as a SCSI disk. All current storage device types, such as SAN, SCSI, and RAID, are supported. iSCSI and SSA are not supported.

The Virtual I/O server supports logical mirroring, and RAID configurations. 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 LVM.

Note: The Shared Ethernet adapter and Virtual SCSI Server functionality is provided in the Virtual I/O Server that is included in the Advanced POWER Virtualization feature.

Partition Load Manager

The Partition Load Manager (PLM) provides automated processor and memory distribution between a dynamic LPAR and Micro-Partitioning 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.

To improve the overall resource utilization of a partitioned system, PLM uses user-defined resource management policies to determine the additional resources, such as processors and memory, for each requesting partition.

PLM uses the Resource Monitoring and Control (RMC) subsystem for network communication, which provides several events on every managed partition node. 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

To help ensure a secure communication between managed partition nodes, OpenSSH and Kerberos V5 are supported in PLM to have a secure communication and an authentication mechanism for administrators. If Kerberos is not installed, PLM uses the next configured authentication method, such as AIX authentication.

2.11 Service processor

The service processor (SP) is an embedded controller based on a PowerPC 405GP processor (PPC405) implementation running the SP internal operating system. The SP operating system contains specific programs and device drivers for the SP hardware.

The p5-520 uses the SP implementation. The key components include a FSP-Base (FSP-B) and an Extender chipset (FSP-E). FSP-B and FSP-E are implemented on a dedicated card. See Figure 2-12.

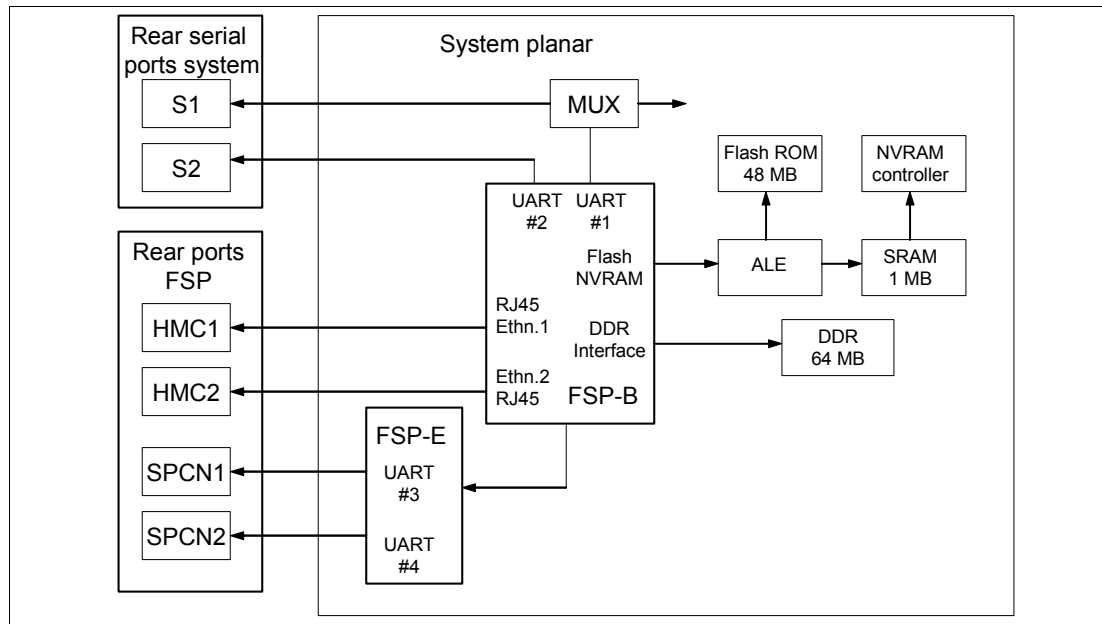


Figure 2-12 Service processor logical diagram

2.11.1 Service processor base

The PPC405 core is 5-stage pipeline instruction processor and contains 32-bit general purpose registers. The Flash ROM contains a compressed image of a software load.

The base unit offers the following connections:

- ▶ Two Ethernet Media Access Controller3 (MAC3) cores, which is a generic implementation of the Ethernet Media Access (MAC) protocol compliant with ANSI/IEEE 802.3, IEEE 802.3u, ISO/IEC 8802.3 CSMA/CD Standard. The Ethernet MAC3 supports both half duplex (CSMA/CD) and full duplex operation at 10/100 Mbps. Both Ethernet ports are visible only to the service processor.
- ▶ Two serial interfaces, accessible only through the serial ports of p5-520 on the rear side. At the time of writing, the System Management Interface (SMI) is usable if a connection is established to serial port 1. Terminals connected to serial port 2 receiving only boot sequence information without manual interaction. When the HMC is connected to the SP, the serial ports are disabled and do not provide any external connection.

2.11.2 Service processor extender

The SP extender unit offers two system power control network (SPCN) ports that are used to control the power of the attached I/O subsystem. The SPCN control software and the service processor software are run on the same PPC405 processor.

2.12 Boot process

From the earlier RS/6000 systems, through the previous pSeries systems, the boot process passed through several enhancements. With the implementation of the POWER5 chip technology in the pSeries platform, the boot process is also based on an enhanced flexibility that the POWER5 processor-based hardware features. Depending on customer demand, a system may or may not require the use of an HMC to manage the system. The boot process, based on the Initial Program Load (IPL) setup, will depend on the hardware setup and from the way you will use the features that POWER5 processor-based systems provide.

The IPL process starts when power is connected to the system. Immediately after, the SP starts an internal self test based on integrated diagnostic programs (Built-In-Self-Test, BIST). Only if all the test units have been successfully passed, the system status changes to standby.

2.12.1 IPL flow without an HMC attached to the system

When the system status is standby, the SP provides a System Management Interface (SMI) that can be accessed by pressing any key on an attached serial console keyboard, or the Advanced System Management Interface (ASMI) using a Web browser⁴ on a client system that is connected to the SP on an Ethernet network.

The SP and the ASMI are standard on all IBM POWER5 processor-based hardware. Both system management interfaces require you to enter the general or admin ID password and allow you to set flags that affect the operation of the system, according to the provided password, such as auto power restart, to view information about the system (such as the error log and VPD), network environment access setup, and to control the system power.

You can start and shut down the system in addition to setting IPL options. This server has a permanent firmware boot side, or A side, and a temporary firmware boot side, or B side. New levels of firmware should be installed on the temporary side first in order to test the update's compatibility with your applications. When the new level of firmware has been approved, it can be copied to the permanent side.

In the SMI and ASMI, you can view and change IPL settings:

- ▶ System boot speed.
Fast or Slow. Fast boot results in skipped diagnostic tests and shorter memory tests during the boot.
- ▶ Firmware boot side for next boot.
Permanent or Temporary. Firmware updates should be tested by booting from the temporary side before being copied into the permanent side.
- ▶ System operating mode.
Manual or Normal. Manual mode overrides various automatic power-on functions, such as auto-power restart, and enables the power switch button.
- ▶ AIX/Linux partition mode boot, available only if the system is not managed by the HMC:
 - Service mode boot from saved list. This is the preferred way to run concurrent AIX diagnostics
 - Service mode boot from default list. This is the preferred way to run stand-alone AIX diagnostics

⁴ Supported browsers are Netscape (Version 7.1), Microsoft® Internet Explorer (Version 6.0), and Opera (Version 7.23). At the time of writing, previous versions of these browsers are not supported. JavaScript™ and cookies must be enabled.

- Boot to open firmware prompt
- Boot to System Management Service (SMS) to further select the boot devices or network boot options.
- ▶ Boot to server firmware:
 - Select the state for the server firmware: Standby or Running.
 - When the server is in the server firmware standby state, partitions can be set up and activated.

2.12.2 Hardware Management Console

Depending from the model, the HMC provides a number of native serial ports and Ethernet ports. One serial port can be used to attach a modem for the Service Agent. The Service Agent Connection Manager can be used instead if the HMC has a TCP/IP port 80 connection to the Internet. The HMC provides an Ethernet port (or ports) to connect to partitions on its POWER5 processor-based managed systems. The network connection is mandatory for the HMC to p5 systems, and highly recommended between the HMC and partitions. It supports the following functions:

- ▶ Logical partition configuration and management
- ▶ Dynamic logical partitioning
- ▶ Capacity and resource management
- ▶ System status
- ▶ HMC management
- ▶ Service functions (for example, Microcode Updates and Service Focal Point)
- ▶ Remote HMC interface

Note: The same HMC cannot be attached to POWER4 and POWER5 processor-based systems simultaneously, but, for redundancy purposes, one POWER5 processor-based server can be attached to two HMCs.

All the managed servers must be authenticated from the HMC. If a new attached system is discovered, the HMC will prompt you to set two passwords using the HMC interface:

- ▶ Advanced System Management general user ID password
- ▶ Advanced System Management admin ID password
- ▶ HMC access password

2.12.3 IPL flow with an HMC attached to the system

When the system status is standby, you can use the HMC to open a virtual terminal and access the SMI, or launch a Web browser to access the ASMI.

Using the SMI or the ASMI, you can view or modify the proper IPL settings in order to set the boot mode to partition standby and then turn the system on. However, the HMC can be also used to power on the managed system (and is highly recommended). Using the HMC to turn the system on requires you to select one of the following choices:

- ▶ Partition Standby
 - The Partition Standby power-on mode allows you to create and activate logical partitions.

- When the Partition Standby power-on is completed, the operator panel on the managed system displays *LPAR...*, indicating the managed system is ready for you to use the HMC to partition its resources and, possibly, activate them.
- When a partition is activated, the HMC requires you to select the boot mode of the single partition.
- ▶ **System Profile**
The System Profile option powers on the system according to a predefined set of profiles. The profiles are activated in the order in which they are shown in the system profile.
- ▶ **Partition autostart**
This option powers on the managed system to partition standby mode and then activates all partitions that have been designated autostart.

After the system boots with any of the above choices, the HMC can be used to manage the system, such as continuing to boot from the operating system or manage the logical partitions. See 2.12.2, “Hardware Management Console” on page 38.

2.12.4 Definitions of partitions

Describing the detailed process to work with the HMC and the management tasks to create and manage a logical partition, LPAR or dynamic LPAR, is not the intention of this documentation. The following section describes the additional functionality used to create partitions that are using fractional elements of available system resources using Micro-Partitioning technology.

For a better understanding of the partitioning concept, this section contains an overview of common terminology. On top of the partitioning concept, there are two components:

- ▶ **Managed systems**
- ▶ **Profiles**

Managed systems

Managed systems are physical systems that are managed by the HMC, whereby one HMC can manage more managed systems at a time.

Profiles

A profile defines the configuration of a logical partition or managed system. There are three types of profiles that can be used to create multiple profiles for each logical partition or managed system:

- ▶ **Partition profile**
 - A partition profile includes the collection of resource specifications, such as processing units, memory, and I/O resources, because a logical partition is not aware of a resource until it is activated.
 - A logical partition can have more than one partition profile, but at least one is a minimum requirement.
- ▶ **All resources dedicated partition profile**
A partition profile that contains the entire resource list of the machine, using all physical resources working as one system.

- ▶ System profile
 - A system profile is an ordered list of partition profiles. When you activate a system profile, the managed system will attempt to activate the partition profiles in the defined order.
 - To enhance the flexibility to use the system within several different logical configuration, a system profile can be defined to collect more than one partition profile to provide the requested system behavior.

2.12.5 Hardware requirements for partitioning

To implement Micro-Partitioning on a POWER5 processor-based system, resource planning is important to have a base configuration and enough flexibility to make desirable changes to the running logical partitions. To configure a partition, the minimum requirements needed are processors, memory, and possibly an expansion unit to define more partitions than possible in a single system.

Processors

Within POWER5 technology and depending on performance requirements, a logical partition can be created by using a shared processor pool or a dedicated processor.

Shared processors can be defined by a fractional number of 1/10 as a minimum requirement of a real processor. To calculate the required processor power, a real processor is divided into 100 processing units, and 1/10 of a processor is equal with 10 processing units.

Dedicated processors are entire processors that can be assigned to a single logical partition without the capability to share free capacity to other logical partitions.

Memory

Depending on given application and performance requirements, a logical partition requests memory to execute the installed operating system and application. To create partitions, the minimum memory requirement is 128 MB per logical partition and dynamically increased by increments of 16 MB from the overall memory available in the system.

Expansion unit

Expansion units extend the flexibility of the server system to enlarge the number of possible logical partitions by adding additional hardware, such as storage or network devices.

2.12.6 Specific partition definitions used for Micro-Partitioning

In addition to the base definition for a partition using Micro-Partitioning technology, new parameters must be defined to receive more flexibility and capacity usage of logical partitions included in POWER5 technology.

Capped and uncapped partition

A capped partition indicates that the local partition will never exceed its assigned capacity. An uncapped partition indicates that if the capacity entitlement is reached, additional capacity from the shared pool can be used if available.

To manipulate the behavior of uncapped partitions, the parameter uncapped weight, in the range from 0 through 255, must be defined. To prevent an uncapped partition from receiving extra capacity, the uncapped weight parameter should be 0.

The default uncapped weight is 128.

2.12.7 System Management Services

Either booting up a full partition system or a logical partition to System Management Services (SMS), the ASCII⁵ interface or the GUI are identical in contents and functionality.

The p5-520 (or the logical partition) must be equipped with either a graphic adapter connected to a graphics display, keyboard, and mouse device, or an ASCII display terminal connected to one of the native serial ports, or the attached HMC to use the SMS menus. It is possible to view information about the system (or the single logical partition) and perform tasks such as setting a password, changing the boot list, and setting the network parameters.

If the system or the partition has been activated without flagging the option to stop to the SMS, there is the option to press the 1 key on the terminal, or in the graphic window, after the word keyboard appears and before the word speaker appears. In the terminal, or in the GUI, the system or the partitions will require you to enter the password defined for admin or general access. After the text-based SMS starts (either for terminal or graphic window), a screen similar to the one shown in Figure 2-13 on page 41 opens.

```
Version SF220_004
SMS 1.5 (c) Copyright IBM Corp. 2000,2003 All right reserved
-----
Main Menu
 1. Select Language
 2. Setup Remote IPL (Initial Program Load)
 3. Change SCSI Settings
 4. Select Console
 5. Select Boot Options

-----
Navigation Keys:

                                     X = eXit System Management Services
-----
Type the number of the menu item and press Enter or select Navigation Key:
```

Figure 2-13 System Management Services main menu

Note: The version of system firmware currently installed in your system is displayed at the top of each screen. Processor and other device upgrades might require a specific version of firmware to be installed in your system.

On each menu screen, you are given the option of choosing a menu item and pressing Enter (if applicable), or selecting a navigation key. You can use the different options to review or set the boot list information, or to set up the network environment parameters if you want the system boots from a NIM server.

⁵ American Standard Code for Information Interchange: This is the world-wide standard for the code numbers used by computers to represent all the uppercase and lowercase Latin letters, numbers, punctuation, and so forth.

2.12.8 Boot options

The p5-520 handles the boot process in a way that is similar to other pSeries servers.

The initial stage of the boot process is to establish that the machine has powered up correctly and the memory and CPUs are functioning correctly. After the machine or the logical partition reaches the SMS menus, all of the necessary tests have been performed and the machine is scanning the bus for a boot source.

Most pSeries system backplanes are designed such that the drive in the first slot spins up immediately after power-on, and other drives will wait for the operating system to send a command before spinning up. Disk drive bays 4 and 8 are hardwired to spin-up immediately. The left-most slot of the 4-pack disk backplanes (SCSI ID 8, boot, autostart) is set to spin up immediately after power-on. The power-on delay sequence is performed to prevent power supply overloading. This behavior makes the disk in the first slot of the first 4-pack DASD backplane the preferred boot device. See Figure 2-14 on page 42 to locate all of the disk bays and the SCSI enclosure services (SES) ID.

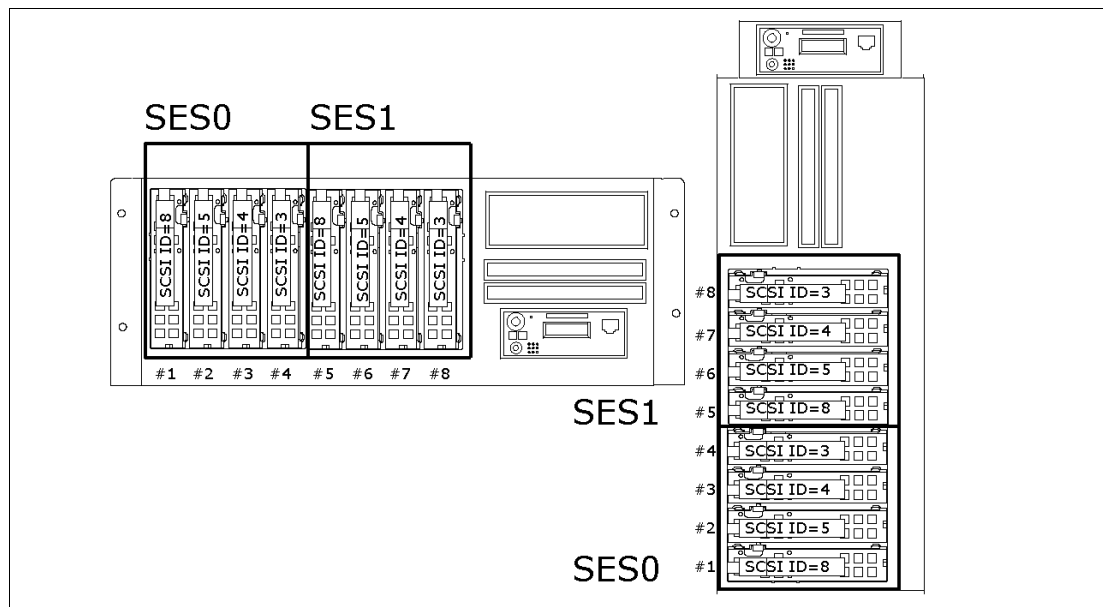


Figure 2-14 Disk bays and SCSI addresses within an p5-520

When SMS menus are available, the Select Boot Options menu can be used to view and set various options regarding the installation devices and boot devices:

1. Select Install or Boot a Device
Enables you to select a device to boot from or install the operating system from. This selection is for the current boot only.
2. Select Boot Devices
Enables you to set the boot list.
3. Multiboot Startup
Toggles the multiboot startup flag, which controls whether the multiboot menu is invoked automatically on startup.

2.12.9 Additional boot options

Instead of booting from the preferred boot device, or from any other internal disks, there are a number of other possibilities:

DVD-ROM, DVD-RAM

These devices can be used to boot the system, or a logical partition (if the resource is available to the specific partition), so that a system can be loaded, system maintenance performed, or stand-alone diagnostics performed.

Internal or external tape drives

The media bay tape drive or any externally attached tape drive can be used to boot the system, or a logical partition (if the resource is available to the specific partition) using `mksysb`, for example.

SCSI disk, and Virtual SCSI disk

The more common method of booting the system is to use a disk situated in one of the hot-swap bays in the front of the machine. However, any external SCSI-attached disk could be used if required. As described in previous sections, Virtual SCSI devices are also available to a logical partition.

SAN boot

It is possible to boot the p5-520 system from a SAN using a 2 GB Fibre Channel Adapter (FC 6239), or it is possible to boot one partition using the dedicated 2 GB Fibre Channel Adapter or the Virtual SCSI device related to this adapter. The IBM 2105 Enterprise Storage Server (ESS) is an example of a SAN-attached device that can provide a boot medium.

LAN boot

Network boot and NIM installs can be used if required. Logical partitions can use both a dedicated Ethernet adapter or Virtual Ethernet to accomplish that.

2.12.10 Security

The p5-520 system allows you to set two different types of passwords to limit the access to these systems. These are defined in the ASMI menus. This password is usually used by the system administrator. The *general ID password* provides limited access to the system functions and is usually available to all users who are allowed to power on the server, especially remotely.

2.13 Operating system requirements

All new POWER5 servers are capable of running IBM AIX 5L for POWER and support appropriate versions of Linux. AIX 5L has been specifically developed and enhanced to exploit and support the extensive RAS features on IBM @server pSeries systems.

2.13.1 AIX 5L

The p5-520 requires AIX 5L Version 5.3 or AIX 5L Version 5.2 Maintenance Package 5200-04 (IY56722) or later. The use of the Advanced POWER Virtualization feature requires AIX 5L Version 5.3.

The system requires the following media:

- ▶ AIX 5L for POWER Version 5.2 5765-E62, dated 08/2004 (CD# LCD4-1133-04) or later
- ▶ AIX 5L for POWER Version 5.3 5765-G03, dated 08/2004 (CD# LCD4-7463-00) or later

IBM periodically releases maintenance packages for the AIX 5L operating system. These packages are available on CD-ROM (FC 0907), or they can be downloaded from the Internet at:

<http://techsupport.services.ibm.com/server/fixes>

You can also get individual operating system fixes and information about obtaining AIX 5L service at this site. In AIX 5L Version 5.3, there is also the `suma` command available that helps the administrator to automate the task of checking and downloading operating system downloads. For more information about the `suma` command functionality, see 3.2.4, “Service Update Management Assistant” on page 53.

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

2.13.2 Linux

For the p5-520, Linux distributions are available through SUSE and Red Hat at the time this publication was written. The p5-520 requires the following version of Linux distributions:

- ▶ SUSE LINUX Enterprise Server 9 for POWER systems, or later
- ▶ Red Hat Enterprise LINUX AS for POWER Version 3

The Advanced POWER Virtualization feature, DLPAR, and other features require SUSE SLES 9. Red Hat Enterprise LINUX supports the Advanced POWER Virtualization feature.

In Japan, Turbolinux is also available. In the Latin America sales region, Conectiva is also available. For related information and an overview, see:

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

Find full information about SUSE LINUX Enterprise Server 9 for POWER at:

http://www.suse.com/us/business/products/server/sles/i_pseries.html

For information about Red Hat Enterprise Linux AS for pSeries from Red Hat, see:

<http://www.redhat.com/software/rhel/as>

For information about UnitedLinux for pSeries from Turbolinux, see:

<http://www.turbolinux.co.jp>

For the latest in IBM Linux news, subscribe to the Linux Line. See:

<https://www6.software.ibm.com/reg/linux/linuxline-i>

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

http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux_pseries.html

Linux support

IBM only supports the Linux systems of customers with a SupportLine contract covering Linux. Otherwise, the Linux distributor should be contacted for support.



RAS and manageability

The following sections provide more detailed information about IBM @server p5 design features that will help lower the total cost of ownership (TCO). This section includes several features based on the benefits available when using AIX 5L. Support of these features using Linux can vary.

3.1 Reliability, availability, and serviceability

Excellent quality and reliability are inherent in all aspects of the IBM @server p5 processor 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 p5-520 as a POWER5 server enhances the RAS capabilities implemented in POWER4-based systems. RAS enhancements available on POWER5 servers 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 2 to 10 for better self-healing capability.

The following sections describe the concepts that form the basis of leadership RAS features of IBM @server p5 systems in more detail.

3.1.1 Fault avoidance

p5 systems are built on a quality-based design to keep errors from ever 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 p5-520 incorporates advanced capability in start-up diagnostics and in run-time First Failure Data Capture (FDDC) based on strategic error checkers built into the chips.

Any errors detected by the pervasive error checkers are captured into Fault Isolation Registers (FIRs), which can be interrogated by the service processor (SP). The SP in the p5-520 has the capability to access system components using special purpose service processor ports or by access to the error registers. Figure 3-1 on page 47 shows a schematic of a Fault Register Implementation.

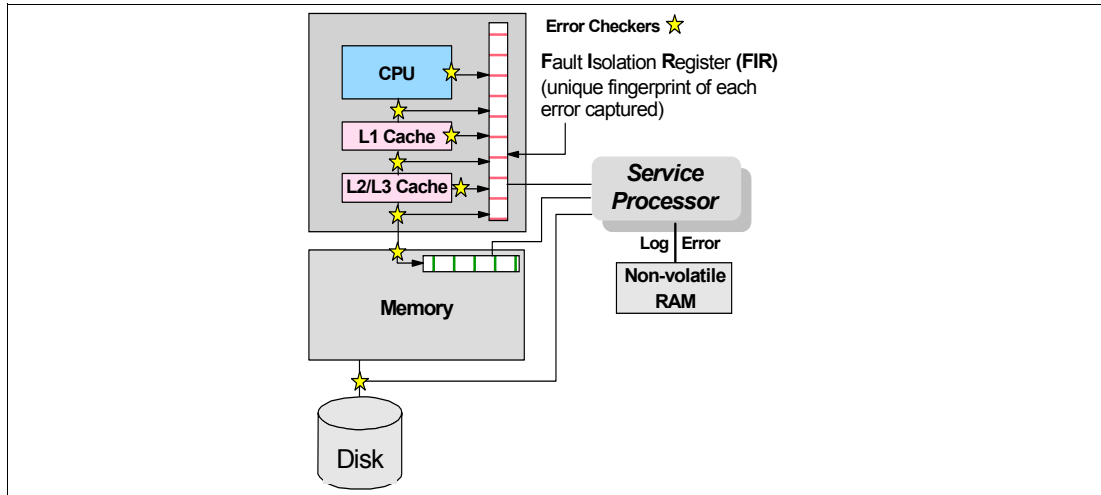


Figure 3-1 Schematic of Fault Isolation Register implementation

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 SP event history log, along with a notification of the event to AIX 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 49.

3.1.3 Permanent monitoring

The SP included in the p5-520 provides a way to monitor the system even when the main processor is inoperable. See the next subsection for a more detailed description of monitoring functions in p5-520.

Mutual surveillance

The SP 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, Class 5 AC power loss) 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 SP.

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)
- ▶ Single bit error correction using ECC without reaching error thresholds for main, L2, and L3 cache memory
- ▶ 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 p5-520 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 p5-520 to remain operational with full resources:

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

3.1.6 Fault masking

If corrections and retries succeed and do not exceed threshold limits, the system remains operational with full resources, and no client or IBM customer engineer intervention is required:

- ▶ 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 run-time deconfiguration is provided if a threshold of L1 or L2 recovered errors is exceeded.

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

PCI 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. Run-time 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 p5-520 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 p5-520 system, hardware, system firmware, and AIX 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 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. p5-520 improves service productivity by providing the functions described in the following subsections.

Error indication and LEDs indicators

The p5-520 is designed for customer setup of the machine and for the subsequent addition of most hardware features. The p5-520 also allows customers to replace service parts (Customer Replaceable Unit). To accomplish this, the p5-520 provides internal 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 Field Replaceable Unit.

For more information about Customer 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

3.2 Manageability

The functions and tools provided for IBM *eServer* p5 systems are described in the next sections.

3.2.1 Service processor

With system in power standby mode, or with an operating system in control of the machine, or controlling the related partition, the SP is working and checking the system for errors, ensuring the connection to the HMC for manageability purposes. With the system up and running, the SP provides the possibility to view and change the Power-On settings, using the Advanced System Management Interface (ASMI). Also, the surveillance function of the SP is monitoring the operating system to check that it is still running and has not stalled.

See Figure 3-2 on page 52 for an example of the ASMI accessed from a Web browser.

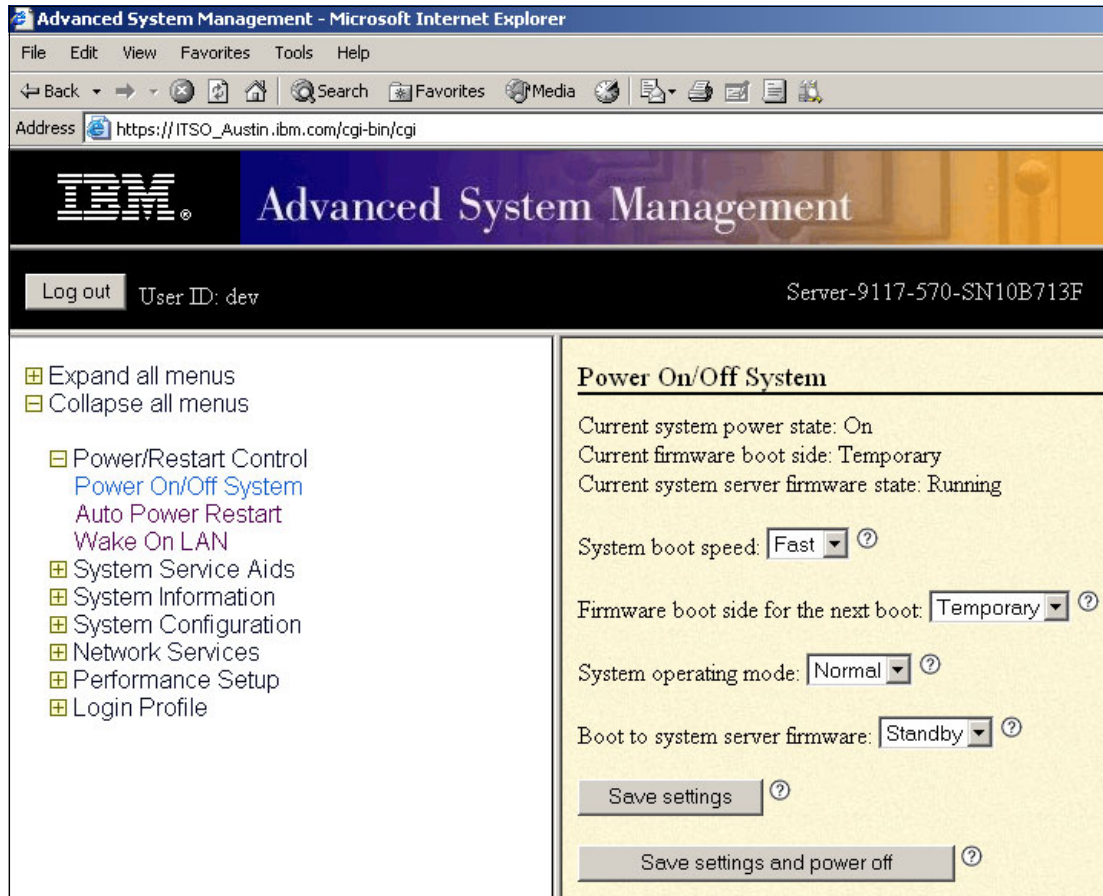


Figure 3-2 Advanced System Management main menu

3.2.2 Service Agent

Service Agent is an application program that operates on an IBM @server p5, pSeries, or IBM RS/6000 computer and monitors them for hardware errors. It reports detected errors, assuming they meet certain criteria for criticality, to IBM for service with no customer intervention. It is an enhanced version of Service Director™ with a graphical user interface.

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

- ▶ Automatic problem analysis
- ▶ Problem-definable threshold levels for error reporting
- ▶ Automatic problem reporting; service calls placed to IBM without intervention
- ▶ Automatic customer 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.
- ▶ VPD data can be sent to IBM using Performance Management.

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 customer 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 Focal Point

Service Focal Point is used by service technicians to start and end their service calls. It provides service representatives with event, Vital Product Data (VPD), and diagnostic information. The HMC can also notify service representatives of hardware failures automatically by using the Service Agent features. You can configure the HMC to use the Service Agent call-home feature to send IBM event information. This information is stored, analyzed, and then acted upon by the service representative. Some parts of Service Focal Point need to be configured so that the proper information is sent to IBM.

You can download the latest version of Service Agent at:

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

3.2.3 IBM eServer p5 Customer-Managed Microcode

The pSeries and RS/6000 Customer-Managed Microcode is a methodology that enables you to manage and install microcode updates on p5, pSeries, and RS/6000 systems and associated I/O adapters. The IBM pSeries Microcode Update Web site can be found at:

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

IBM provides service tools that can assist you in determining microcode levels and updating systems with the latest available microcode. To determine which tool to use in a specific environment, visit:

<http://techsupport.services.ibm.com/server/mdownload/mcodetools.html>

3.2.4 Service Update Management Assistant

The Service Update Management Assistant (SUMA) helps system administrators retrieve maintenance updates from the Web. SUMA offers flexible options that let customers set up policies to automate the download of fixes to their systems. SUMA policies can be scheduled to periodically check the availability of specific new fixes (microcode, APAR, PTF, or fileset), critical or security fixes, or an entire maintenance level. A notification e-mail can be sent detailing updates that are needed when comparing available fixes to installed software, a fix repository, or a maintenance level.

Benefits provided by SUMA:

- ▶ Moves administrators away from the task of manually retrieving maintenance updates from the Web.
- ▶ Policy can be scheduled to run periodically, for example, to download the latest critical fixes weekly.
- ▶ Can compare fixes needed against software inventory, fix repository, or a maintenance level.
- ▶ Receive mail notification after a fileset preview or download operation.
- ▶ Allows for FTP, HTTP, or secure HTTPS transfers

- ▶ Provides the same requisite checking as the IBM fix distribution Web site
- ▶ Available through SMIT menus (smitty suma) or a command line interface

3.3 IBM eServer Cluster 1600

Today's IT infrastructure requires that systems 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.

IBM @server Cluster 1600 is a POWER-based AIX 5L and Linux Cluster targeting scientific and technical computing, large-scale databases, and workload consolidation

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

The p5-520 is supported with the IBM @server Cluster 1600 running CSM for AIX, V1.3.1. To attach a p5-520 to a Cluster 1600, an HMC is required. One HMC can also control several p5-520s that are part of the cluster. If a p5-520 configured in partition mode (with physical or virtual resources) is part of the cluster, all partitions must be part of the cluster.

It is not possible to use selected partitions as part of the cluster and use others for non-cluster use. The HMC uses a dedicated connection to the p5-520 to provide the functions needed to control the server, such as powering the system on and off. The HMC must have an Ethernet connection to the Control Work Station (CWS). Each partition in p5-520 must have an Ethernet adapter to connect to the CWS *trusted* LAN.

Information regarding HMC control, cluster building block servers, and cluster software available can be found at:

<http://www.ibm.com/servers/eserver/clusters/>

The benefits of a clustered environment based on logical partitions

The evolution of processor and storage technologies has a great impact on the architecture of IT infrastructures. This was the most significant challenge for the infrastructure in the past and will also be in the future. During the first half of the 1990s, one single central instance of an application per node was suitable, moreover, most productive systems needed additionally associated nodes, so-called application servers.

Increasing performance and reliability by simply replicating application server nodes led to complex environments that often resulted in poor system management. The reason for these complex constructions was the limited computing power of a single node. This limitation was softened during the second half of the 1990s.

Big symmetric multiprocessor (SMP) nodes with higher clock rates and increased memory provided the possibility to install more than one system on a node. This had some side effects regarding systems operations: Release planning processes had to pay attention to different databases or application versions, or both, to avoid unresolved conflicts.

In 2000, Workload Manager for AIX (WLM) was announced. Multiple application instance installations became more and more popular because of the permanently increasing number of systems dedicated to applications at customer sites. The general availability of this functionality of AIX to separate the workloads of dedicated systems eliminated the last obstacle for consolidating several systems in one node.

Some customers expanded the usage of their dedicated systems and consequently model more business processes. This often caused an increased number of dedicated systems used and a stronger demand on flexibility. In addition, the life cycle of these systems differed extremely. Renaming, removal, or deletion became more and more common system administration tasks.

In 2001, the pSeries hardware technology with logical partitioning was generally available. Logical partitioning creates the possibility to define the logical partitions (LPARs) that are adapted to customer needs regarding the number of processors, assigned memory, and I/O adapters, meaning no waste of resources, but the flexibility to assign the right power at the right moment. The p5-520 offers the flexibility to increase the usage of the resources even more and reduce the total cost of ownership (TCO).

Partitions with associated physical resources or virtual resources are not different from a collection of stand-alone nodes.

Today, server consolidation is a must for many IT sites. Minimized TCO and complexity, with the maximum amount of flexibility, is a crucial goal of nearly all customers. LPARs allow a flexible distribution of resources with LPAR boundaries. Each logical partition can be configured according to the specific needs of the occupant application. LPARs provide a protection boundary between the systems. More test and development systems can exist on the same server in separate partitions.

The CSM value points

The CSM allows the management of different hardware platforms from one single point of control and it has consistent interfaces to manage systems and logical partitions running both AIX and Linux. The management is achieved across multiple switch and interconnect topologies. PSSP forced system administrators to do some things a certain way (such as NIM, and SP user management). The CSM provides assistance on setting these things up, but allows the system administrator to tailor their system to their own needs, and it has the ability to manage systems across different geographical sites.

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

The CSM ships with AIX 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.

CSM Version 1.4 on AIX and Linux (planned 4Q04)

CSM V1.4 on AIX and Linux introduces an optional IBM CSM High Availability Management Server feature, designed to allow automated failover of the CSM management server to a backup management server. In addition, sample scripts for setting up NTP, and network tuning (AIX 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.

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 59. Note that some of the documents referenced here may be available in softcopy only.

- ▶ *IBM @server pSeries 670 and pSeries 690 System Handbook*, SG24-7040
- ▶ *The Complete Partitioning Guide for IBM pSeries Servers*, SG24-7039
- ▶ *Managing AIX Server Farms*, SG24-6606
- ▶ *Practical Guide for SAN with pSeries*, SG24-6050
- ▶ *Problem Solving and Troubleshooting in AIX 5L*, SG24-5496
- ▶ *Understanding IBM @server pSeries Performance and Sizing*, SG24-4810

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.
- ▶ *7316-TF3 17-Inch Flat Panel Rack-Mounted Monitor and Keyboard Installation and Maintenance Guide*, SA38-0643, contains information regarding the 7316-TF3 Flat Panel Display, which can be installed in your rack to manage your system units.
- ▶ *IBM @server Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590, provides information to operators and system administrators on how to use a IBM Hardware Management Console for pSeries (HMC) to manage a system. It also discusses the issues associated with logical partitioning planning and implementation.
- ▶ *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 @server pSeries Adapters, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516, contains information about adapters, devices, and cables for your system. This manual is intended to supplement the service information found in the *Diagnostic Information for Multiple Bus Systems* documentation.
- ▶ *RS/6000 and @server pSeries Diagnostics Information for Multiple Bus Systems*, SA38-0509, contains diagnostic information, service request numbers (SRNs), and failing function codes (FFCs).
- ▶ *RS/6000 and pSeries PCI Adapter Placement Reference*, 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/pseries/aixfixes.html>
- ▶ Autonomic computing on IBM @server pSeries servers
<http://www.ibm.com/autonomic/index.shtml>
- ▶ Ceramic Column Grid Array (CCGA), see IBM Chip Packaging
<http://www.ibm.com/chips/micronews>
- ▶ Copper circuitry
<http://www.ibm.com/chips/technology/technologies/copper/>
- ▶ Frequently asked SSA-related questions
<http://www.storage.ibm.com/hardsoft/products/ssa/faq.html>
- ▶ Hardware documentation
http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/
- ▶ IBM @server Information Center
<http://publib.boulder.ibm.com/eserver/>
- ▶ IBM @server pSeries and RS/6000 microcode update
<http://techsupport.services.ibm.com/server/mdownload2/download.html>
- ▶ IBM @server pSeries support
<http://www.ibm.com/servers/eserver/support/pseries/index.html>
- ▶ IBM @server support: Tips for AIX administrators
<http://techsupport.services.ibm.com/server/aix.srchBroker>
- ▶ IBM Linux news: Subscribe to the Linux Line
<https://www6.software.ibm.com/reg/linux/linuxline-i>
- ▶ Information about UnitedLinux for pSeries from Turbolinux
<http://www.turbolinux.co.jp>
- ▶ IBM online sales manual
<http://www.ibm1ink.ibm.com>
- ▶ Linux for IBM @server pSeries
<http://www.ibm.com/servers/eserver/pseries/linux/>
- ▶ Microcode Discovery Service
<http://techsupport.services.ibm.com/server/aix.invscountMDS>
- ▶ POWER4 system micro architecture, comprehensively described in the *IBM Journal of Research and Development*, Vol 46 No.1 January 2002
<http://www.research.ibm.com/journal/rd46-1.html>
- ▶ SCSI T10 Technical Committee
<http://www.t10.org>
- ▶ Silicon-on-insulator (SOI) technology
<http://www.ibm.com/chips/technology/technologies/soi/>

- ▶ SSA boot FAQ
<http://www.storage.ibm.com/hardsoft/products/ssa/faq.html#microcode>
- ▶ SUSE LINUX Enterprise Server 8 for pSeries information
http://www.suse.de/us/business/products/server/sles/i_pseries.html
- ▶ The LVT is a PC based tool intended assist you in logical partitioning
<http://www-1.ibm.com/servers/eserver/series/lpar/systemdesign.htm>

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 *@*server p5 520 Technical Overview and Introduction



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

This document is a comprehensive guide covering the IBM *@*server p5 520 UNIX servers. We introduce major hardware offerings and discuss their prominent functions. Professionals wishing to acquire a better understanding of IBM *@*server p5 products should consider reading this document. The intended audience includes:

Outstanding performance based on POWER5 processor technology

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

From Web servers to integrated cluster solutions

This document expands the current set of IBM *@*server documentation by providing a desktop reference that offers a detailed technical description of the p5-520 system. This publication does not replace the latest pSeries marketing materials and tools. It is intended as an additional source of information that, together with existing sources, can 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**