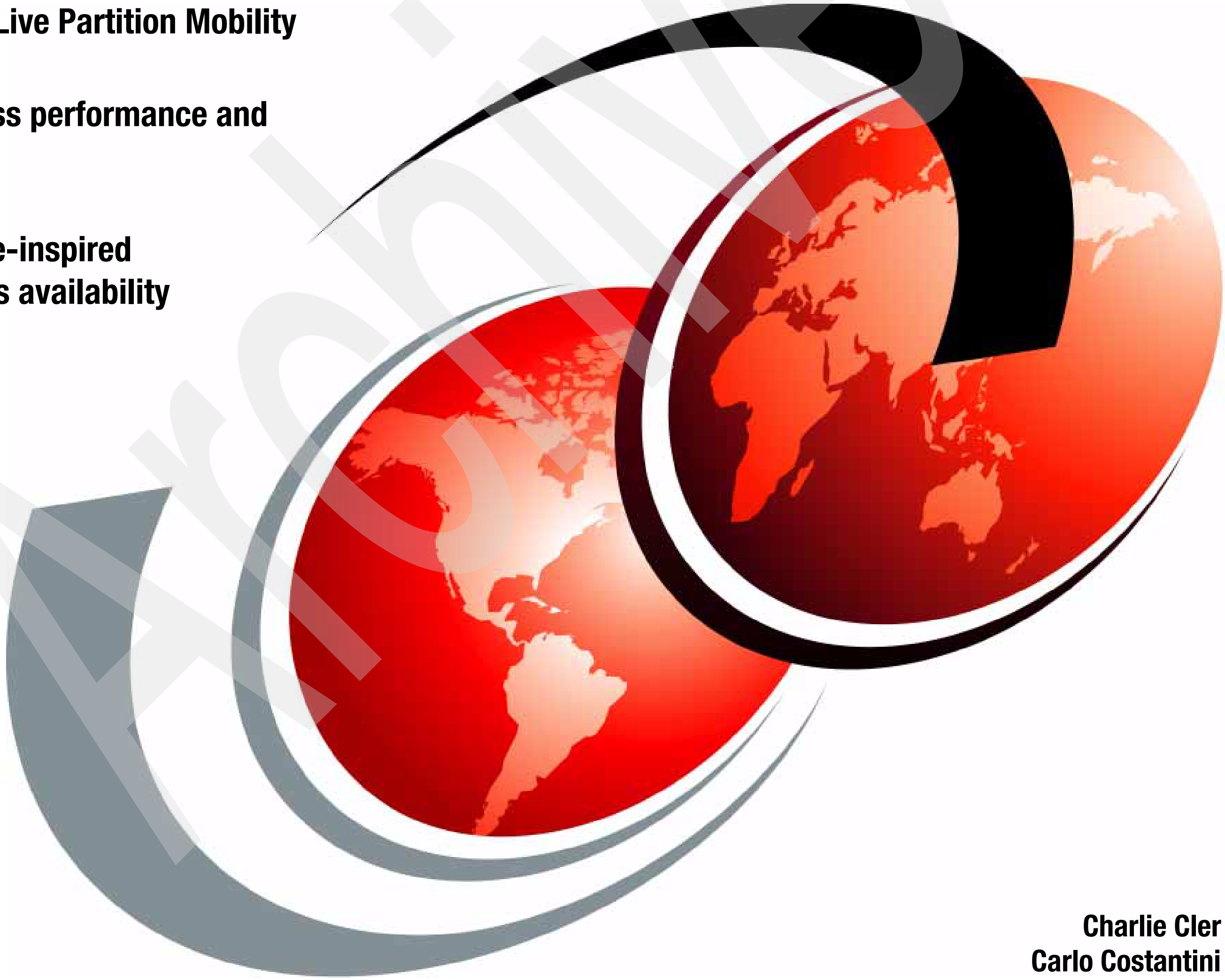
IBM Power 595

Technical Overview and Introduction

PowerVM virtualization technology including Live Partition Mobility

World-class performance and flexibility

Mainframe-inspired continuous availability



Charlie Cler
Carlo Costantini

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM Power 595 Technical Overview and Introduction

August 2008

REDP-4440-00

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

First Edition (August 2008)

This edition applies to the IBM Power Systems 595 (9119-FHA), IBMs most powerful Power Systems offering.

Contents

Notices

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

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

1350™
AIX 5L™
AIX®
BladeCenter®
Chipkill™
DB2®
DS8000™
Electronic Service Agent™
EnergyScale™
eServer™
HACMP™
i5/OS®
IBM®
iSeries®
Micro-Partitioning™
OpenPower®
OS/400®
POWER™
Power Architecture®
POWER Hypervisor™
POWER4™
POWER5™
POWER5+™
POWER6™
PowerHA™
PowerPC®
PowerVM™
Predictive Failure Analysis®
pSeries®
Rational®
Redbooks®
Redbooks (logo) ®
RS/6000®
System i™
System i5™
System p™
System p5™
System Storage™
System x™
System z™
Tivoli®
TotalStorage®
WebSphere®
Workload Partitions Manager™
z/OS®

The following terms are trademarks of other companies:

AMD, the AMD Arrow logo, and combinations thereof, are trademarks of Advanced Micro Devices, Inc.

Novell, SUSE, the Novell logo, and the N logo are registered trademarks of Novell, Inc. in the United States and other countries.

ABAP, SAP NetWeaver, SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Java, JVM, Power Management, Ultra, and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

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

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States, other countries, or both.

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

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

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

Preface

This IBM® Redpaper is a comprehensive guide describing the IBM Power 595 (9119-FHA) enterprise-class IBM Power Systems server. The goal of this paper is to introduce several technical aspects of this innovative server. The major hardware offerings and prominent functions include:

- The POWER6™ processor available at frequencies of 4.2 and 5.0 GHz
- Specialized POWER6 DDR2 memory that provides improved bandwidth, capacity, and reliability
- Support for AIX®, IBM i, and Linux® for Power operating systems.
- EnergyScale™ technology that provides features such as power trending, power-saving, thermal measurement, and processor napping.
- PowerVM™ virtualization
- Mainframe levels of continuous availability.

This Redpaper is intended for professionals who want to acquire a better understanding of Power Systems products, including:

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

This Redpaper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the 595 system.

This Redpaper does not replace the latest marketing materials, tools, and other IBM publications available, for example, at the IBM Systems Hardware Information Center http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp. 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 paper

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

Charlie Cler is an Executive IT Specialist for IBM in the United States. He has worked with IBM Power Systems and related server technology for over 18 years. Charlie’s primary areas of expertise include Power Systems processor virtualization and server consolidation. He holds a masters degree in Mechanical Engineering from Purdue University with specialization in robotics and computer graphics.

Carlo Costantini is a Certified IT Specialist for IBM and has over 30 years of experience with IBM and IBM Business Partners. He currently works in Italy Power Systems Platforms as Presales Field Technical Sales Support for IBM Sales Representatives and IBM Business

Partners. Carlo has broad marketing experience and his current major areas of focus are competition, sales, and technical sales support. He is a certified specialist for Power Systems servers. He holds a masters degree in Electronic Engineering from Rome University.

The project manager that organized the production of this material was:

Scott Vetter, (PMP) is a Certified Executive Project Manager at the International Technical Support Organization, Austin Center. He has enjoyed 23 years of rich and diverse experience working for IBM in a variety of challenging roles. His latest efforts are directed at providing world-class Power Systems Redbooks®, whitepapers, and workshop collateral.

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

Terry Brennan, Tim Damron, George Gaylord, Dan Henderson, Tenley Jackson,
Warren McCracken, Patrick O'Rourke, Paul Robertson, Todd Rosedahl, Scott Smylie,
Randy Swanberg, Doug Szerdi, Dave Williams
IBM Austin

Mark Applegate
Avnet

Become a published author

Join us for a two- to six-week residency program! Help write a book dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You will have the opportunity to team with IBM technical professionals, Business Partners, and Clients.

Your efforts will help increase product acceptance and client satisfaction. As a bonus, you will 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 paper or other IBM Redbooks in one of the following ways:

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

  ibm.com/redbooks

- Send your comments in an e-mail to:

  redbooks@us.ibm.com

- Mail your comments to:

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

1

General description

IBM System i™ and IBM System p™ platforms are unifying the value of their servers into a single and powerful lineup of IBM Power Systems servers based on POWER6-processor technology with support for the IBM i operating system (formerly known as i5/OS®), IBM AIX, and Linux for Power operating systems. This new single portfolio of Power Systems servers offers industry-leading technology, continued IBM innovation, and the flexibility to deploy the operating system that your business requires.

This publication provides overview and introductory-level technical information for the POWER6-based IBM Power 595 server with Machine Type and Model (MTM) 9119-FHA.

The IBM Power 595 server is designed to help enterprises deploy the most cost-effective and flexible IT infrastructure, while achieving the best application performance and increasing the speed of deployment of new applications and services. As the most powerful member of the IBM Power Systems family, the Power 595 server is engineered to deliver exceptional performance, massive scalability and energy-efficient processing for a full range of complex, mission-critical applications with the most demanding computing requirements.

Equipped with ultra-high frequency IBM POWER6 processors in up to 64-core, symmetric multiprocessing (SMP) configurations, the Power 595 server is designed to scale rapidly and seamlessly to address the changing needs of today's data centers. With advanced PowerVM virtualization, EnergyScale technology, and Capacity on Demand (CoD) options, the Power 595 is ready to help businesses take control of their IT infrastructure and confidently consolidate multiple UNIX®-based, IBM i, and Linux application workloads onto a single system.

1.1 Model overview and attributes

The Power 595 server (9119-FHA) offers an expandable, high-end enterprise solution for managing the computing requirements to enable your business to become an On Demand Business. The Power 595 is an 8- to 64-core SMP system packaged in a 20U (EIA-unit) tall central electronics complex (CEC) cage. The CEC is 50 inches tall, and housed in a 24-inch wide rack. Up to 4 TB of memory are supported on the Power 595 server.

The Power 595 (9119-FHA) server consists of the following major components:

- A 42U-tall, 24-inch system rack that houses the CEC, Bulk Power Assemblies (BPA) that are located at the top of the rack, and I/O drawers that are located at the bottom of the rack. A redundant power subsystem is standard. Battery backup is an optional feature. CEC features include:
  - A 20U-tall CEC housing that features the system backplane cooling fans, system electronic components, and mounting slots for up to eight processor books.
  - One to eight POWER6 processor books. Each processor book contains eight, dual-threaded SMP cores that are packaged on four multi-chip modules (MCMs). Each MCM contains one dual-core POWER6 processor supported by 4 MB of on-chip L2 cache (per core) and 32 MB of shared L3 cache. Each processor book also provides:
    - Thirty-two DDR2 memory DIMM slots
    - Support for up to four GX based I/O hub adapter cards (RIO-2 or 12x) for connection to system I/O drawers
    - Two Node Controller (NC) service processors (primary and redundant)
- One or two optional Powered Expansion Racks, each with 32U of rack space for up to eight, 4U I/O Expansion Drawers. Redundant Bulk Power Assemblies (BPA) are located at the top of the Powered Expansion Rack. Optional battery backup capability is available. Each Powered Expansion Rack supports one 42U bolt-on, nonpowered Expansion Rack for support of additional I/O drawers.
- One or two nonpowered Expansion Racks, each supporting up to seven 4U I/O Expansion Drawers.
- One to 30 I/O Expansion Drawers (maximum of 12 RIO-2), each containing 20 PCI-X slots and 16 hot-swap SCSI-3 disk bays.
- In addition to the 24 inch rack-mountable I/O drawers, also available are standard, 2 meters high, 19 inch I/O racks for mounting both SCSI and SAS disk drawers. Each disk drawer is individually powered by redundant, 220 V power supplies. The disk drawers can be configured for either RAID or non-RAID disk storage. A maximum of 40 SCSI drawers (each with 24 disks), and 185 SAS drawers (each with 12 disks), can be mounted in 19-inch racks. The maximum number of disks available in 19 inch racks is 960 hot-swap SCSI disks (288 TB) and 2,220 hot-swap SAS disks (666 TB).

> **Note:** In this publication, the main rack containing the CEC is referred to as the *system rack*. Other IBM documents might use the terms *CEC rack or Primary system rack* to refer to this rack.

Table 1-1 on page 3 lists the major attributes of the Power 595 (9119-FHA) server.

Table 1-1 Attributes of the 9119-FHA

Attribute	**9119-FHA**
SMP processor configurations	8- to 64 core POWER6 using 8-core processor books
8-core processor books	Up to 8
POWER6 processor clock rate	4.2 GHz Standard or 5.0 GHz Turbo
L2 cache	4 MB per core
L3 cache	32 MB per POWER6 processor (shared by two cores)
RAM (memory)	16, 24, or 32 DIMMs configured per processor book Up to 4 TB of 400 MHz DDR2 Up to 1 TB of 533 MHz DDR2 Up to 512 GB of 667 MHz DDR2
Processor packaging	MCM
Maximum memory configuration	4 TB
Rack space	42U 24-inch custom rack
I/O drawers	24": 1 - 30
19" I/O drawers	0 - 96
Internal disk bays	16 maximum per 24" I/O drawer
Internal disk storage	Up to 4.8 TB per 24" I/O drawer
64-bit PCI-X Adapter slots	#5791 RIO-2 drawer: 20 PCI-X (133 MHz), 240 per system #5797 or #5798 drawer: 14 PCI-X 2.0 (266 MHz), 6 PCI-X (133 MHz), 600 per system
I/O ports	4 GX+ adapter ports per processor book, 32 per system
POWER™ Hypervisor	LPAR, Dynamic LPAR, Virtual LAN
PowerVM Standard Edition (optional)	Micro-Partitioning™ with up to 10 micro-partitions per processor (254 maximum); Multiple shared processor pools; Virtual I/O Server; Shared Dedicated Capacity; PowerVM Lx86
PowerVM Enterprise Edition (optional)	PowerVM Standard Edition plus Live Partition Mobility
Capacity on Demand configurations	8 to 64 processor cores in increments of one (using one to eight processor books); 4.2 or 5.0 GHz POWER6 processor cores. [a]
Capacity on Demand (CoD) features (optional)	Processor CoD (in increments of one processor), Memory CoD (in increments of 1 GB), On/Off Processor CoD, On/Off Memory CoD, Trial CoD, Utility CoD
High availability software	PowerHA™ family

Attribute	9119-FHA
RAS features	Processor Instruction Retry Alternate Processor Recovery Selective dynamic firmware updates IBM Chipkill™ ECC, bit-steering memory ECC L2 cache, L3 cache Redundant service processors with automatic failover Redundant system clocks with dynamic failover Hot-swappable disk bays Hot-plug/blind-swap PCI-X slots Hot-add I/O drawers Hot-plug power supplies and cooling fans Dynamic Processor Deallocation Dynamic deallocation of logical partitions and PCI bus slots Extended error handling on PCI-X slots Redundant power supplies and cooling fans Redundant battery backup (optional)
Operating systems	AIX V5.3 or later IBM i V5.4 or later SUSE® Linux Enterprise Server 10 for POWER SP2 or later Red Hat Enterprise Linux 4.7 and 5.2 for POWER or later

a. Minimum requirements include a single 8-core book with three cores active, and for every 8-core book, three cores must be active.

1.2 Installation planning

Complete installation instructions are shipped with each server. The Power 595 server must be installed in a raised floor environment. Comprehensive planning information is available at this address:

http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp

1.2.1 Physical specifications

The key specifications, such as dimensions and weights, are described in this section. Table 1-2 lists the major Power 595 server dimensions.

Table 1-2 Power 595 server dimensions

Dimension	Rack only	Rack with side doors	Slim line		Acoustic	
			1 Frame	2 Frames	1 Frame	2 Frames
Height	201.4 cm (79.3 in)	201.4 cm (79.3 in)	201.4 cm (79.3 in)	201.4 cm (79.3 in)	201.4 cm (79.3 in)	201.4 cm (79.3 in)
Width	74.9 cm (29.5 in)	77.5 cm (30.5 in)	77.5 cm (30.5 in)	156.7 cm (61.7 in)	77.5 cm (30.5 in)	156.7 cm (61.7 in)
Depth	127.3 cm (50.1 in)		148.6 cm (58.5 in)[a] 152.1 cm (61.3 in)[b]		180.6 cm (71.1 in)	180.6 cm (71.1 in)

a. Rack with slim line and side doors, one or two frames

b. Rack with slim line front door and rear door heat exchanger (RDHX), system rack only

Table 1-3 lists the Power 595 server full system weights without the covers.

Table 1-3 Power 595 server full system weights (no covers)

Frame	With integrated battery backup	Without integrated battery backup
A Frame (system rack)	1542 kg (3400 lb)	1451 kg (3200 lb)
A Frame (powered expansion rack)	1452 kg (3200 lb)	1361 kg (3000 lb)
Z Frame (bolt-on expansion rack)	N/A	1157 kg (2559 lb)

Table 1-4 lists the Power 595 cover weights.

Table 1-4 Power 595 cover weights

Covers	Weight
Side covers pair	50 kg (110 lb)
Slim Line doors, single	15 kg (33 lb)
Acoustic doors, single (Expansion frame)	25 kg (56 lb)
Acoustic doors, single (System rack)	20 Kg (46 lb)

Table 1-5 lists the Power 595 shipping crate dimensions.

Table 1-5 Power 595 shipping crate dimensions

Dimension	Weight
Height	231 cm (91 in)
Width	94 cm (37 in)
Depth	162 cm (63.5 in)
Weight	Varies by configuration. Max 1724 kg (3800 lb)

1.2.2 Service clearances

Several possible rack configurations are available for Power 595 systems. Figure 1-1 on page 6 shows the service clearances for a two-rack configuration with acoustical doors.

> **Note:** The Power 595 server must be installed in a raised floor environment.

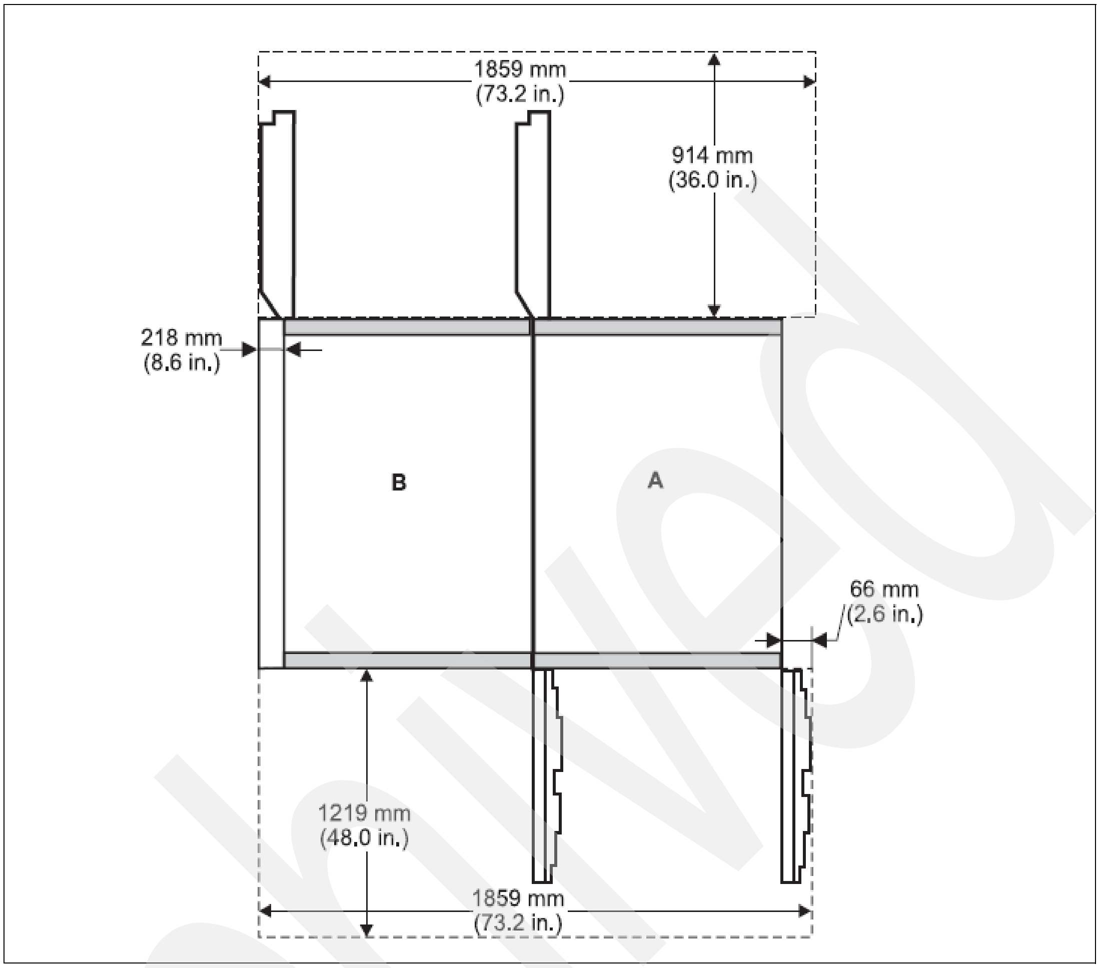


Figure 1-1 Service clearances for a two-rack system configuration with acoustic doors

Service clearances for other configurations can be found at:

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

Important: If the Power 595 server must pass through a doorway opening less than 2.02 meters (79.5 inches), you should order the compact handling option (#7960) which, ships the rack in two parts.

1.2.3 Operating environment

Table 1-6 lists the operating environment specifications for the Power 595 server.

Table 1-6 Power 595 server operating environment specifications

Description	Range
Recommended operating temperature (8-core, 16-core, and 32-core configurations)	10 degrees to 32 degrees C[a] (50 degrees to 89.6 degrees F)
Recommended operating temperature (48-core and 64-core configurations)	10 degrees to 28 degrees C[a] (50 degrees to 82.4 degrees F)
Relative humidity	20% to 80%

Description	Range
Maximum wet bulb	23 degrees C (73 degrees F) (operating)
Sound power	► Declared A-weighted sound power level, per ISO 9296: 9.2 bels (with slim line doors) ► Declared A-weighted sound power level, per ISO 9296: 8.2 bels (with acoustical doors)
Sound pressure	► Declared A-weighted one-meter sound pressure level, per ISO 9296: 79 decibels (with slim line doors) ► Declared A-weighted one-meter sound pressure level, per ISO 9296: 69 decibels (with acoustical doors)

a. The maximum temperatures of 32°C (90°F) and 28°C (82°F) are linearly derated above 1295 m (4250 ft).

1.2.4 Power requirements

All Power 595 configurations are designed with a fully redundant power system. To take full advantage of the power subsystem redundancy and reliability features, each of the two power cords should be connected to different distribution panels.

Table 1-7 lists the electrical and thermal characteristics for the Power 595 server.

Table 1-7 Power 595 electrical and thermal characteristics

Description	Range
Operating voltages	► 3-phase V ac at 50/60 Hz): 200 to 240 V; 380 to 415 V; 480 V ► Rated current (A per phase): 48 A or 63 A or 80 A; 34 A or 43 A; 24 A or 34 A ► Power consumption: 27,500 watts (maximum for full CEC, three I/O drawers) ► Power source loading: 27.7 kVA ► Thermal output: 27,500 joules/sec (93,840 Btu/hr) maximum
Inrush current	134
Power Factor	0.99
Operating frequency	50/60 plus or minus 0.5 Hz
Maximum Power (Fully configured 4.2 GHz system)	23.3 KW
Maximum Power (Fully configured 5.0 GHz system)	28.3 KW
Maximum thermal output (4.2 GHz processor)	74.4 KBTU/hr
Maximum thermal output (5.0 GHz processor)	83.6 KBTU/hr

Table 1-8 on page 8 lists the electrical characteristics for 4.2 GHz and 5.0 GHz Power 595 servers, and the Powered Expansion Rack (U.S., Canada, and Japan).

Table 1-8 Electrical characteristics (U.S., Canada, and Japan)

Description	US, Canada, Japan		US high voltage	
Voltage and Frequency	200-240 V at 50-60 Hz		480 V at 50-60 Hz	
4.2 GHz Server				
System Rating	48 A	63 A	24 A	24 A
Plug rating	60 A	100 A	30 A	30 A
Recommended circuit breaker rating	60 A	80 A	30 A	30 A
Cord size	6 AWG	6 AWG	8 AWG	8 AWG
Recommended receptacle	IEC60309, 60 A, type 460R9W	IEC60309, 100 A, type 4100R9W	IEC60309, 30 A, type 430R7W	IEC60309, 30 A, type 430R7W
5.0 GHz Server				
System Rating	48 A	63 A	24 A	34 A
Plug rating	60 A	100 A	30 A	60 A
Recommend circuit breaker rating	60 A	100 A	30 A	60 A
Cord size	6 AWG	4 AWG	8 AWG	6 AWG
Recommended receptacle	IEC60309, 60 A, type 460R9W	IEC60309, 100 A, type 4100R9W	IEC60309, 30 A, type 430R7W	IEC60309, 30 A, type 430R7W
Powered Expansion Rack				
System rating	48 A	63 A	24 A	24 A
Plug rating	60 A	100 A	30 A	30 A
Recommend circuit breaker rating	60 A	80 A	30 A	30 A
Cord size	6 AWG	6 AWG	8 AWG	8 AWG
Recommended receptacle	IEC60309, 60 A, type 460R9W	IEC60309, 100 A, type 4100R9W	IEC60309, 30 A, type 430R7W	IEC60309, 30 A, type 430R7W

Table 1-9 lists the electrical characteristics for 4.2 GHz and 5.0 GHz Power 595 servers, and the Powered Expansion Rack (World Trade).

Table 1-9 Electrical characteristics (World Trade)

Description	World Trade			
Voltage and frequency	200-240 V at 50-60 Hz		380/415 V at 50-60 Hz	
4.2 GHz server				
System rating	48 A	63 A	34 A	34 A
Plug rating	no plug	no plug	no plug	no plug

Description	World Trade			
Recommend circuit breaker rating	60 A	80 A	40 A	40 A
Cord size	6 AWG	6 AWG	8 AWG	8 AWG
Recommended receptacle	Not specified Electrician installed	Not specified Electrician installed	Not specified Electrician installed	Not specified Electrician installed
5.0 GHz server				
System rating	48 A	80 A	34 A	43 A
Plug rating	no plug	no plug	no plug	no plug
Recommend circuit breaker rating	60 A	100 A	40 A	63 A
Cord size	6 AWG	4 AWG	8 AWG	6 AWG
Recommended receptacle	Not specified Electrician installed	Not specified Electrician installed	Not specified Electrician installed	Not specified Electrician installed
Powered Expansion Rack				
System rating Powered I/O Rack	48 A	63 A	34 A	34 A
Plug rating	no plug	no plug	no plug	no plug
Recommend circuit breaker rating	60 A	80 A	40 A	40 A
Cord size	6 AWG	6 AWG	8 AWG	8 AWG
Recommended receptacle	Not specified Electrician installed	Not specified Electrician installed	Not specified Electrician installed	Not specified Electrician installed

1.3 Minimum configuration requirements

This section discusses the minimum configuration requirements for the Power 595. Also provided are the appropriate feature codes for each system component. The IBM configuration tool also identifies the feature code for each component in your system configuration. Table 1-10 on page 10 identifies the required components for a minimum 9119-FHA configuration.

> **Note:** Throughout this chapter, all feature codes are referenced as #*xxxx*, where *xxxx* is the appropriate feature code number of the particular item.

Table 1-10 Power 595 minimum system configuration

Quantity	Component description	Feature code
1	Power 595	9119-FHA
1	8-core, POWER6 processor book 0-core active	#4694
3	1-core, processor activations	3 x #4754
4	Four identical memory features (0/4 GB or larger)	—
16	1 GB memory activations (16x #5680)	#5680
1	Power Cable Group, first processor book	#6961
4	Bulk power regulators	#6333
2	Power distribution assemblies	#6334
2	Line cords, selected depending on country and voltage	—
1	Pair of doors (front/back), either slim line or acoustic	—
1	Universal lift tool/shelf/stool and adapter	#3759 and #3761
1	Language - specify one	#93xx (country dependent)
1	HMC (7042-COx/CRx) attached with Ethernet cables	—
1 or 1	RIO I/O Loop Adapter 12X I/O Loop Adapter	#1814 #1816
1 or 1	One I/O drawer providing PCI slots attached to the I/O loop As an alternative when the 12X I/O Drawer (#5798) becomes available, located at A05 in the system rack	#5791 (AIX, Linux) #5790 (IBM i) #5798
2	Enhanced 12X I/O Cable 2.5 M	#1831
1	Enhanced 12X I/O Cable, 0.6 m (#1829)	#1829

Prior to the availability of 12X Expansion Drawers (#5797 or #5798), new server shipments will use an RIO I/O Expansion Drawer model dependent on the primary operating system selected. When 12x Expansion Drawers become available, they become the default, base I/O Expansion Drawer for all operating systems.

If AIX or Linux for Power operating system is specified as the primary operating system, see Table 1-11 for a list of the minimum, required features:

Table 1-11 Minimum required features when AIX or Linux for Power is the primary operating system

Quantity	Component description	Feature code
1	Primary Operating System Indicator for AIX or Linux for Power	#2146 #2147
1	PCI-X 2.0 SAS Adapter (#5912 or #5900) or PCI LAN Adapter for attachment of a device to read CD media or attachment to a NIM server	#5912 (#5900 is supported)
2	15,000 RPM, 146.8 GB, SCSI Disk Drives	#3279

Quantity	Component description	Feature code
1	RIO-2 I/O drawer located at location U5 in the system rack prior to the availability of #5798	#5791
2	RIO-2 I/O bus cables, 2.5 m	#3168
1	Remote I/O Cable, 0.6 m	#7924
1	UPIC Cable Group, BPD1 to I/O Drawer at position U5 in the system rack	#6942

If IBM i is specified as the primary operating system, refer to Table 1-12, which lists the minimum required features.

Table 1-12 Minimum required features when IBM i is the primary operating system

Quantity	Component description	Feature code
1	Primary operating system indicator for IBM i	#2145
1	System console specify	—
1 or 1	SAN Load Source Specify: Requires Fibre Channel Adapter Internal Load Source Specify: Requires disk controller and minimum of two disk drives	For example, #5749 For example, #5782, two #4327
1	PCI-X 2.0 SAS Adapter for attachment of a DVD drive	#5912
1	PCI 2-Line WAN Adapter with Modem	#6833
1	RIO-attached PCI Expansion Drawer (prior to feature 5798 availability)	#5790
—	Rack space in a Dual I/O Unit Enclosure	#7307, #7311
1	RIO-2 Bus Adapter	#6438
2	RIO-2 I/O Bus Cables, 8 m	#3170
2	Power cords	#6459
2	Power Control Cable, 6 m SPCN	#6008
1 or 1	Media Drawer, 19-inch (prior to feature 5798/5720 availability). ▸ One DVD drive ▸ Power cords ▸ SAS cable for attachment to #5912 SAS adapter 595 Media Drawer, 24-inch with #5798 availability	#7214-1U2 #5756 For example, #6671 For example, #3684 #5720
1	19-inch rack to hold the #5790 and 7214-1U2	—
1	PDU for power in 19-inch rack	For example, #7188

1.3.1 Minimum required processor card features

The minimum configuration requirement for the Power 595 server is one 4.2 GHz 8-core processor book and three processor core activations, or two 5.0 GHz 8-core processor books and six processor core activations. For a description of available processor features and their associated feature codes, refer to 2.6, “Processor books” on page 70.

1.3.2 Memory features

The Power 595 utilizes DDR2 DRAM memory cards. Each processor book provides 32 memory card slots for a maximum of 256 memory cards per server. The minimum system memory is 16 GB of active memory per processor book.

The Power 595 has the following minimum and maximum configurable memory resource allocation requirements:

- Utilizes DDR2 DRAM memory cards.
- Requires a minimum of 16 GB of configurable system memory.
- Each processor book provides 32 memory card slots for a maximum of 256 memory cards per server. The minimum system memory is 16 GB of active memory per processor book.
- Supports a maximum of 4 TB DDR2 memory.
- Memory must be configured with a minimum of four identical memory features per processor book, excluding feature #5697 (4 DDR2 DIMMs per feature). Feature #5697, 0/64 GB memory must be installed with 8 identical features.
- Different memory features cannot be mixed within a processor book. For example, in a 4.2 GHz processor book (#4694), four 0/4 GB (#5693) features, 100% activated DIMMs are required to satisfy the minimum active system memory of 16 GBs. For two 4.2 GHz or 5.0 GHz processor books (#4694 or #4695), four 0/4 GB (#5693) features, 100% activated in each processor book is required to satisfy the minimum active system memory of 32 GBs. If 0/8 GB (#5694) features are used, then the same minimum system memory requirements can be satisfied with 50% of the DIMMs activated.
- Each processor book has four dual-core MCMs, each of which are serviced by one or two memory features (4 DIMMs per feature). DDR2 memory features must be installed in increments of one per MCM (4 DIMM cards per memory feature), evenly distributing memory throughout the processor books installed. Incremental memory for each processor book must be added in identical feature pairs (8 DIMMs). As a result, each processor book will contain either four, six, or eight identical memory features (two per MCM), which equals a maximum of 32 DDR2 memory DIMM cards.
- Memory features #5694, #5695, and #5696 must be 50% activated as a minimum at the time of order with either feature #5680 or #5681.
- Features #5693 (0/4 GB) and #5697 (0/64 GB) must be 100% activated with either feature #5680 or #5681 at the time of purchase.
- Memory can be activated in increments of 1 GB.
- All bulk order memory features #8201, #8202, #8203, #8204, and #8205 must be activated 100% at the time of order with feature #5681.
- Maximum system memory is 4096 GB and 64 memory features (eight features per processor book or 256 DDR2 DIMMs per system). DDR1 memory is not supported.

For a list of available memory features refer to Table 2-15 on page 80.

1.3.3 System disks and media features

This topic focuses on the I/O device support within the system unit. The Power 595 servers have internal hot-swappable drives supported in I/O drawers. I/O drawers can be allocated in 24-inch or 19-inch rack (IBM i application only). Specific client requirements can be satisfied with several external disk options supported by the Power 595 server.

For further information about IBM disk storage systems, including withdrawn products, visit:

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

Note: External I/O drawers 7311-D11, 7311-D20, and 7314-G30 are not supported on the Power 595 servers.

The Power 595 has the following minimum and maximum configurable I/O device allocation requirements:

- The 595 utilizes 4U-tall remote I/O drawers for directly attached PCI or PCI-X adapters and SCSI disk capabilities. Each I/O drawer is divided into two separate halves. Each half contains 10 blind-swap PCI-X slots for a total of 20 PCI slots and up to 16 hot-swap disk bays per drawer.
- When an AIX operating system is specified as the primary operating system, a minimum of one I/O drawer (#5791) per system is required in the 5U location within the system rack.
- When an IBM i operating system is specified as the primary operating system, a minimum of one PCI-X Expansion Drawer (#5790) per system is required in a 19-inch expansion rack. A RIO-2 Remote I/O Loop Adapter (#6438) is required to communicate with the 595 CEC RIO-G Adapter (#1814).
- When the 12X I/O drawers (#5797. #5798) is available, the above minimum requirement will be replaced by one feature #5797 or #5798 per system in the 5U location within the system rack. All I/O drawer feature #5791, #5797, or #5798 contain 20 PCI-X slots and 16 disk bays.

  Note: The 12X I/O drawer (#5798) attaches only to the central electronics complex using 12X cables. The 12X I/O drawer (#5797) comes with a repeater card installed. The repeater card is designed to strengthen signal strength over the longer cables used with the Power Expansion Rack (#6954 or #5792) and nonpowered, bolt-on Expansion Rack (#6983 or #8691). Features #5797 and #5798 will not be supported in p5-595 migrated Expansion Rack.

- 7040-61D I/O drawers are supported with the 9119-FHA.
- A maximum of 12-feature #5791 (or #5807), 5794 (specified as #5808), or 30-feature #5797 I/O drawers can be connected to a 595 server. The total quantity of features (#5791+#5797+#5798+#5807+#5808) must be less than or equal to 30.
- One single-wide, blind-swap cassette (equivalent to those in #4599) is provided in each PCI or PCI-X slot of the I/O drawer. Cassettes not containing an adapter will be shipped with a *dummy* card installed to help ensure proper environmental characteristics for the drawer. If additional single-wide, blind-swap cassettes are needed, feature #4599 should be ordered.
- All 10 PCI-X slots on each I/O drawer planar are capable of supporting either 64-bit or 32-bit PCI or PCI-X adapters. Each I/O drawer planar provides 10 PCI-X slots capable of supporting 3.3-V signaling PCI or PCI-X adapters operating at speeds up to 133 MHz.
- Each I/O drawer planar incorporates two integrated Ultra3 SCSI adapters for direct attachment of the two 4-pack hot-swap backplanes in that half of the drawer. These adapters do not support external SCSI device attachments. Each half of the I/O drawer is powered separately.
- For IBM i applications, if additional external communication and storage devices are required, a 19-inch, 42U high non-powered Expansion Rack can be ordered as feature #0553. Feature #0553 (IBM i) is equivalent to the 7014-T42 rack, which is supported for use with a 9119-FHA server.

- For IBM i applications, a maximum of 96 RIO-2 drawers or 30 12X I/O drawers can be attached to the 595, depending on the server and attachment configuration.The IBM i supported #0595, #0588, #5094/#5294, #5096/#5296 and #5790 all provide PCI slots and are supported when migrated to the Power 595. Up to six I/O drawers/towers per RIO loop are supported. Prior to the 24" 12X drawer's availability, the feature #5790 is also supported for new orders.
- The #5786 EXP24 SCSI Disk Drawer and the #5886 EXP 12S SAS Disk Drawer are 19" drawers which are supported on the Power 595.

For a list of the available Power 595 Expansion Drawers, refer to 2.8.2, “Internal I/O drawers” on page 84.

Note: Also supported for use with the 9119-FHA are items available from a model conversion (all IBM i supported, and AIX and Linux are not supported):

- 7014-T00 and feature 0551 (36U, 1.8 meters)
- 7014-S11 and feature 0551 (11U high)
- 7014-S25 and feature 0551 (25U high)
- In addition to the above supported racks, the following expansion drawers and towers are also supported:
  - PCI-X Expansion Tower/Unit (#5094) (IBM i)
  - PCI-X Expansion Tower (no disk) (#5096, #5088 - no longer available (IBM i)
  - 1.8 m I/O Tower (#5294)
  - 1.8 m I/O Tower (no disk) (#5296)
  - PCI-X Tower Unit in Rack (#0595)
  - PCI Expansion Drawer (#5790)

There is no limit on the number of 7014 racks allowed.

Table 1-13 lists the Power 595 hard disk drive features available for I/O drawers.

Table 1-13 IBM Power 595 hard disk drive feature codes and descriptions

Feature code	Description	Support		
		AIX	IBM i	Linux
#3646	73 GB 15K RPM SAS Disk Drive	✓	—	✓
#3647	146 GB 15K RPM SAS Disk Drive	✓	—	✓
#3648	300 GB 15K RPM SAS Disk Drive	✓	—	✓
#3676	69.7 GB 15K RPM SAS Disk Drive	—	✓	—
#3677	139.5 GB 15K RPM SAS Disk Drive	—	✓	—
#3678	283.7 GB 15K RPM SAS Disk Drive	—	✓	—
#3279	146.8 GB 15K RPM Ultra320 SCSI Disk Drive Assembly	✓	—	✓
#4328	141.12 GB 15K RPM Disk Unit	—	✓	—

The Power 595 server must have access to a device capable of reading CD/DVD media or to a NIM server. The recommended devices for reading CD/DVD media is the Power 595 media drawer (#5720), or and external DVD device (7214-1U2, or 7212-103). Ensure there is a SAS adapter available for the connection.

If an AIX or Linux for Power operating system is specified as the primary operating system, a NIM server can be used. The recommended NIM server connection is a PCI based Ethernet LAN adapter plugged in one of the system I/O drawers.

If an AIX or Linux for Power operating system is specified as the primary operating system, a minimum of two internal SCSI hard disks is required per 595 server. It is recommended that these disks be used as mirrored boot devices. These disks should be mounted in the first I/O drawer whenever possible. This configuration provides service personnel the maximum amount of diagnostic information if the system encounters any errors during in the boot sequence. Boot support is also available from local SCSI and Fibre Channel adapters, or from networks via Ethernet or token-ring adapters.

Placement of the operating systems disks in the first I/O drawer allows the operating system to boot even if other I/O drawers are found offline during boot. If the boot source other than internal disk is configured, the supporting adapter should also be in the first I/O drawer.

Table 1-14 lists the available Power 595 media drawer features.

Table 1-14 IBM Power 595 media drawer features

Feature code	Description	Support		
		AIX	IBM i	Linux
#0274	Media Drawer, 19-inch	✓	✓	—
#4633	DVD RAM	—	✓	—
#5619	80/160 GB DAT160 SAS Tape Drive	✓	—	✓
#5689	DAT160 Data Cartridge	✓	—	✓
#5746	Half High 800 GB/1.6 TB LTO4 SAS Tape Drive	✓	—	✓
#5747	IBM LTO Ultrium 4 800 GB Data Cartridge	✓	—	✓
#5756	IDE Slimline DVD ROM Drive	✓	✓	✓
#5757	IBM 4.7 GB IDE Slimline DVD RAM Drive	✓	✓	✓

1.3.4 I/O Drawers attachment (attaching using RIO-2 or 12x I/O loop adapters)

Existing System i and System p model configurations have a set of I/O enclosures that have been supported on RIO-2 (HSL-2) loops for a number of years.

Most continue to be supported on POWER6 models. This section highlights the newer I/O enclosures that are supported by the POWER6 models that are actively marketed on new orders. See 2.8, “Internal I/O subsystem” on page 82 for additional information about supported I/O hardware.

- System I/O drawers are always connected using RIO-2 loops or 12X HCA loops to the GX I/O hub adapters located on the front of the processor books. Drawer connections are always made in loops to help protect against a single point-of-failure resulting from an open, missing, or disconnected cable.

- Systems with non-looped configurations could experience degraded performance and serviceability.
- RIO-2 loop connections operate bidirectional at 1 GBps (2 GBps aggregate). RIO-2 loops connect to the system CEC using RIO-2 loop attachment adapters (#1814). Each adapter has two ports and can support one RIO-2 loop. Up to four of the adapters can be installed in each 8-core processor book.
- 12X HCA loop connections operate bidirectional at 3 GBps (6 GBps aggregate). 12X loops connect to the system CEC using 12X HCA loop attachment adapters (#1816). For AIX applications up to 12 RIO-2 drawers or 30 12X I/O drawers can be attached to the 595, depending on the server and attachment configuration.
- The total quantity of features #5791+#5797+#5798+#5807+#5808 must be less than or equal to 30.

Slot plugging rules are complex, and depend on the number of processor books present. Generally, the guidelines are:

- Slots are populated from the top down.
- #1816 adapters are placed first and #1814 are placed second.
- A maximum of 32 GX adapters per server are allowed.

Refer to 2.8.1, “Connection technology” on page 83 for a list of available GX adapter types and their feature numbers.

I/O drawers can be connected to the CEC in either single-loop or dual-loop mode. In some situations, dual-loop mode might be recommended because it provides the maximum bandwidth between the I/O drawer and the CEC. Single-loop mode connects an entire I/O drawer to the CEC through one loop (two ports). The two I/O planars in the I/O drawer are connected with a short jumper cable. Single-loop connection requires one loop (two ports) per I/O drawer.

Dual-loop mode connects each I/O planar in the drawer to the CEC separately. Each I/O planar is connected to the CEC through a separate loop. Dual-loop connection requires two loops (four ports) per I/O drawer.

Refer to Table 2-22 on page 89 for information about the number of single-looped and double-looped I/O drawers that can be connected to a 595 server based on the number of processor books installed.

> **Note:** On initial Power 595 server orders, IBM manufacturing places dual-loop connected I/O drawers as the lowest numerically designated drawers followed by any single-looped I/O drawers.

1.3.5 IBM i, AIX, Linux for Power I/O considerations

As we indicated previously, some operating system-specific requirements, and current System i and System p client environments dictate differences, which are documented where appropriate in this publication.

Examples of unique AIX I/O features include graphic adapters, specific WAN/LAN adapters, SAS disk/tape controllers, iSCSI adapters, and specific Fibre Channel adapters.

Examples of unique IBM i I/O features include the #5094/#5294/#5088/#0588/#0595 I/O drawers/towers (I/O enclosures), I/O Processors (IOPs), IOP-based PCI adapter cards, very large write cache disk controllers, specific Fibre Channel adapters, iSCSI adapters, and specific WAN/LAN adapters.

System i hardware technologies and the IBM i operating system (OS/400®, i5/OS, and so forth) have a long history of supporting I/O adapters (IOAs, also commonly referred to as *controllers*) that also required a controlling I/O Processor (IOP) card. A single IOP might support multiple IOAs. The IOP card originally had a faster processor technology than its attached IOAs. Thus, microcode was placed in the IOP to deliver the fastest possible performance expected by clients.

IOAs introduced over the last two to three years (since the time of writing), have very fast processors and do not require a supporting IOP. Among the System i community, these adapters are sometimes referred to as *smart IOAs* that can operate with or without an IOP. Sometimes, these IOAs are also referred to as a *dual mode IOA*. There are also IOAs that do not run with an IOP. These are sometimes referred to as an *IOP-less* IOA (or one that does not run with an IOP).

AIX or Linux client partitions hosted by an IBM i partition are independent of any unique IBM i I/O hardware requirements.

For new system orders, IOP-less IOAs are what AIX or Linux users consider as the normal I/O environment. New orders for IBM i, AIX, and Linux operating systems should specify the smart or IOP-less IOAs.

However, many System i technology clients who are considering moving to the POWER6 models should determine how to handle any existing IOP-IOA configurations they might have. Older technology IOAs and IOPs should be replaced or I/O enclosures supporting IOPs should be used.

The POWER6 system unit does not support IOPs and thus IOAs that require an IOP are not supported. IOPs can be used in supported I/O enclosures attached to the system by using a RIO-2 loop connection. RIO-2 is the System p technology term used in this publication to also represent the I/O loop technology typically referred to as HSL-2 by System i clients.

Later in this publication, we discuss the PCI technologies that can be placed within the processor enclosure. For complete PCI card placement guidance in a POWER6 configuration, including the system unit and I/O enclosures attached to loops, refer to the documents available at the IBM Systems Hardware Information Center at the following location (the documents are in the Power Systems information category):

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp

PCI placement information for the Power 595 server can be found in the *Power Systems PCI Adapter Placement Guide for Machine Type 820x and 91xx,* SA76-0090.

1.3.6 Hardware Management Console models

Each Power 595 server must be connected to an Hardware Management Console (HMC) for system control, LPAR, Capacity on Demand, and service functions. It is highly recommended that each 595 server is connected to two HMCs for redundancy. The Power 595 is connected to the HMC through Ethernet connections to the front and rear Bulk Power Hub (BPH) in the CEC Bulk Power Assembly.

Several HMC models are available for POWER6-based systems at the time of writing. See 2.13, “Hardware Management Console (HMC)” on page 101 for details about specific HMC models. HMCs are preloaded with the required Licensed Machine Code Version 7 (#0962) to support POWER6 systems, in addition to POWER5™ and POWER5+™ systems.

Existing HMC models 7310 can be upgraded to Licensed Machine Code Version 7 to support environments that can include POWER5, POWER5+, and POWER6 processor-based servers. Version 7 is not available for the 7315 HMCs. Licensed Machine Code Version 6 (#0961) is not available for 7042 HMCs, and Licensed Machine Code Version 7 (#0962) is not available on new 7310 HMC orders.

1.3.7 Model conversion

Clients with installed p5-590, p5-595, i5-595, and i5-570 servers can increase their computing power by ordering a model conversion to the 595 server. Table 1-15 lists the available model conversions.

Table 1-15 Available model conversions

From type-model	To type-model
9119 590	9119 FHA
9119 595	9119 FHA
9406 570	9119 FHA
9406 595	9119 FHA

Due to the size and complex nature of the miscellaneous equipment specification (MES) model upgrades into the Power 595 server, a two-step MES process is required. The two MESs are configured in a single eConfig session (the ordering tool used by sales and technical professionals) and contained within the same eConfig Proposed Report. These MESs are processed in sequence.

The initial MES is a Record Purpose Only (RPO) activity that positions the inventory record and conceptually redefines the installed product with Power Systems feature nomenclature. This MES contains a series of RPO feature additions and removals within the installed machine type and model, and adds specify code #0396. This RPO MES serves several purposes. It keeps your maintenance billing whole throughout the upgrade process, reduces the number of feature conversions on the normal MES, and reduces the overall size of the normal MES. This RPO MES should be stored in a separate eConfig file for reference prior to order forward.

The second MES is a normal machine/model upgrade MES from 9406/9119 to 9119-FHA with all the appropriate model/feature conversions and subject to the usual scheduling, manufacturing, and installation rules and processes. Care must be taken that both MESs are processed completely through installation prior to configuration/placement of any subsequent MES orders.

Note: In the event that the RPO MES is reported as installed and the normal MES is cancelled, a sales representative must submit an additional RPO MES reversing the transactions of the initial RPO MES to return your inventory record to its original state. Failure to do so prevents future MES activity for your machine and could corrupt your maintenance billing. The saved eConfig Proposed Report can be used as the basis for configuring this reversal RPO MES.

Ordering a model conversion provides:

- Change in model designation from p5-590, p5-595, i5-595, and i5-570 to 595 (9119-FHA)
- Power 595 labels with the same serial number as the existing server
- Any model-specific system documentation that ships with a new 595 server

Each model conversion order also requires feature conversions to:

- Update machine configuration records.
- Ship system components as necessary.

The model conversion requires an order of a set of feature conversions.

- The existing components, which are replaced during a model or feature conversion, become the property of IBM and must be returned.
- In general, feature conversions are always implemented on a quantity-of-one for quantity-of-one basis. However, this is not true for 16-core processor books.
- Each p590, p595, or i595, 16-core processor book conversion to the Power 595 server actually results in two, 8-core 595 processor books. The duplicate 8-core book is implemented by using eConfig in a unique two-step, MES order process as part of a model conversion described in section1.3.7, “Model conversion” on page 18.
- Excluding p590, p595, and i595 processor books, single existing features might not be converted to multiple new features. Also, multiple existing features might not be converted to a single new feature.
- DDR1 memory is not supported.
- DDR2 memory card (#7814) 9119-590, 9119-595, 4 GB, 533 MHz is not supported.
- Migrated DDR2 memory cards from p590, p595 and i595 donor servers are supported in a 595 server. These are the #4501, #4502, and #4503 memory features.
- If migrating DDR2 memory, each migrated DDR2 memory feature requires an interposer feature. Each memory size (0/8, 0/16, and 0/32 GB) has its own individual interposer: one feature 5605 per 0/8 GB feature, one feature 5611 per 0/16 GB feature, and one feature #5584 per 0/32 GB feature. Each interposer feature is comprised of four interposer cards.
- DDR2 migrated memory features must be migrated in pairs. Four interposers are required for each migrated DDR2 feature (4 DIMMs/feature). Interposer cards must be used in increments of two within the same processor book.
- Each Power 595 processor book can contain a maximum of 32 interposer cards. Within a server, individual processor books can contain memory different from that contained in another processor book. However, within a processor book, all memory must be comprised of identical memory features. This means that, within a 595 processor book, migrated interposer memory cannot be mixed with 595 memory features, even if they are the same size. Within a 595 server, it is recommended that mixed memory should not be different by more than 2x in size. That is, a mix of 8 GB and 16 GB features is acceptable, but a mix of 4 GB and 16 GB is not recommended within a server
- When migrated, the Powered Expansion Rack (#5792) and non-powered Bolt-on Expansion Rack (#8691) are available as expansion racks. When you order features #5792 and #8691 with battery backup, both primary (#6200) and redundant (#6201) battery backup units can be ordered.
- At the time of writing, a conversion of a 9119-590 or 595 to a 9119-FHA requires the purchase of a new 24 inch I/O drawer.

Feature codes for model conversions

The following tables list, for every model conversion, the feature codes involved (processor, memory, RIO adapters, racks, memory, Capacity on Demand, and others).

From type-model 9119-590

Table 1-16 details newly announced features to support an upgrade process.

Table 1-16 Feature conversions for 9119-590 to 9119-FHA

Description	Feature code
Model Upgrade Carry-Over Indicator for converted #4643 with DCA	#5809
Migrated Bolt-on rack	#5881
Migrated Self-Powered rack	#5882
1 GB Carry-Over Activation	#5883
256 GB Carry-Over Activation	#5884
Base 1 GB DDR2 Memory Act	#8494
Base 256 GB DDR2 Memory Act	#8495

For lists of features involved in the 9119-590 to 9119-FHA model conversion, see Table 1-17 (processor) and Table 1-18 (adapter).

Table 1-17 Feature conversions for 9119-590 to 9119-FHA processor

From feature code	To feature code
#7981 - 16-core POWER5 Standard CUoD Processor Book, 0-core Active	#1630 - Transition Feature from 9119-590-7981 to 9119-FHA-4694/4695
#8967 - 16-core POWER5+ 2.1 GHz Standard CUoD Processor Book, 0-core Active	#1633 - Transition Feature from 9119-590-8967 to 9119-FHA-4694/4695
#7667 - Activation, #8967 #7704 CoD Processor Book, One Processor	#4754 - Processor Activation #4754
#7925 - Activation, #7981 or #7730 CoD Processor Book, One Processor	#4754 - Processor Activation #4754
#7667 - Activation, #8967 #7704 CUoD Processor Book, One Processor	#4755 - Processor Activation #4755
#7925 - Activation, #7981 or #7730 CUoD Processor Book, One Processor	#4755 - Processor Activation #4755

Table 1-18 Feature conversions for 9119-590 to 9119-FHA adapters

From feature code	To feature code
#7818 - Remote I/O-2 (RIO-2) Loop Adapter, Two Port	#1814 - Remote I/O-2 (RIO-2) Loop Adapter, Two Port

From feature code	To feature code
#7820 - GX Dual-port 12x HCA	#1816 - GX Dual-port 12x HCA

Table 1-19, Table 1-20, and Table 1-21 list features involved in the 9119-590 to 9119-FHA model conversion (rack-related, the specify-codes, and memory).

Table 1-19 Feature conversions for 9119-590 to 9119-FHA rack-related

From feature code	To feature code
#5794 - I/O Drawer, 20 Slots, 8 Disk Bays	#5797 - 12X I/O Drawer PCI-X, with repeater
#5794 - I/O Drawer, 20 Slots, 8 Disk Bays	#5798 - 12X I/O Drawer PCI-X, no repeater

Table 1-20 Feature conversions for 9119-590 to 9119-FHA specify codes feature

From feature code	To feature code
#4643 - 7040-61D I/O Drawer Attachment Indicator	#5809 - Model Upgrade Carry-Over Indicator for converted #4643 with DCA

Table 1-21 Feature conversions for 9119-590 to 9119-FHA memory

From feature code	To feature code
#7669 - 1 GB Memory Activation for #4500, #4501, #4502 and #4503 Memory Cards	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#7970 - 1 GB Activation #7816 & #7835 Memory Features	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#8471 - 1 GB Base Memory Activations for #4500, #4501, #4502 and #4503	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#7280 - 256 GB Memory Activations for #4500, #4501, #4502 and #4503 Memory Cards	#5681 - Activation of 256 GB DDR2 POWER6 Memory
#8472 - 256 GB Base Memory Activations for #4500, #4501, #4502 and #4503 Memory Cards	#5681 - Activation of 256 GB DDR2 POWER6 Memory
#8493 - 256 GB Memory Activations for #8151, #8153 and #8200 Memory Packages	#5681 - Activation of 256 GB DDR2 POWER6 Memory
#4500 - 0/4 GB 533 MHz DDR2 CoD Memory Card	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS- 667 MHz- POWER6 CoD Memory
#7814 - 4 GB DDR2 Memory Card, 533 MHz	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS- 667 MHz- POWER6 CoD Memory
#7816 - 4 GB CUoD Memory Card 2 GB Active, DDR1	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4500 - 0/4 GB 533 MHz DDR2 CoD Memory Card	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4501 - 0/8 GB 533 MHz DDR2 CoD Memory Card	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory

From feature code	To feature code
#7814 - 4 GB DDR2 Memory Card, 533 MHz	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz- POWER6 CoD Memory
#7816 - 4 GB CUoD Memory Card 2 GB Active, DDR1	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#7835 - 8 GB CUoD Memory Card 4 GB Active, DDR1	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4501 - 0/8 GB 533 MHz DDR2 CoD Memory Card	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4502 - 0/16 GB 533 MHz DDR2 CoD Memory Card	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#7828 - 16 GB DDR1 Memory Card, 266 MHz	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#7835 - 8 GB CUoD Memory Card 4 GB Active, DDR1	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4502 - 0/16 GB 533 MHz DDR2 CoD Memory Card	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4503 - 0/32 GB 400 MHz DDR2 CoD Memory Card	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#7828 - 16 GB DDR1 Memory Card, 266 MHz	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#7829 - 32 GB DDR1 Memory Card, 200 MHz	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4503 - 0/32 GB 400 MHz DDR2 CoD Memory Card	#5697 - 0/64 GB DDR2 Memory(4X16 GB) DIMMS, 400 MHz, POWER6 CoD Memory
#7829 - 32 GB DDR1 Memory Card, 200 MHz	#5697 - 0/64 GB DDR2 Memory(4X16 GB) DIMMS, 400 MHz, POWER6 CoD Memory
#8195 - 256 GB DDR1 Memory32 X 8 GB)	#8201 - 0/256 GB 533 MHz DDR2 Memory Package (32x#5694)
#8153 - 0/256 GB 533 MHz DDR2 Memory Package	#8202 - 0/256 GB 533 MHz DDR2 Memory Package (16x#5695)
#8151 - 0/512 GB 533 MHz DDR2 Memory Package	#8203 - 0/512 GB 533 MHz DDR2 Memory Package (32x#5695)
#8197 - 512 GB DDR1 Memory (32 X 16 GB Cards)	#8203 - 0/512 GB 533 MHz DDR2 Memory Package (32x#5695)
#8198 - 512 GB DDR1 Memory(16 X 32 GB Cards)	#8204 - 0/512 GB 400 MHz DDR2 Memory Package (16x#5696)
#8200 - 512 GB DDR2 Memory (16 X 32 GB Cards)	#8204 - 0/512 GB 400 MHz DDR2 Memory Package (16x#5696)

From type-model 9119-595

Table 1-22 on page 23, Table 1-23 on page 23, and Table 1-24 on page 23 list features in a 9119-595 to 9119-FHA model conversion.

Table 1-22 Processor feature conversions for 9119-595 to 9119-FHA

From feature code	To feature code
#7988 - 16-core POWER5 Standard CoD Processor Book, 0-core Active	#1631 - Transition Feature from 9119-595 #7988 to 9119-FHA #4694/#4695
#7813 - 16-core POWER5 Turbo CoD Processor Book, 0-core Active	#1632 - Transition Feature from 9119-595 #7813 to 9119-FHA #4694/#4695
#8970 - 16-core POWER5+ 2.1 GHz Standard CoD Processor Book, 0-core Active	#1634 - Transition Feature from 9119-595 #8970 to 9119-FHA #4694/#4695
#8968 - 16-core POWER5+ 2.3 GHz Turbo CoD Processor Book, 0-core Active	#1635 - Transition Feature from 9119-595 #8968 to 9119-FHA #4694/#4695
#8969 - New 16-core POWER5 Turbo CoD Processor Book, 0-core Active	#1636 - Transition Feature from 9119-595 #8969 to 9119-FHA #4694/#4695
#7668 - Activation, #8968 or #7705 CoD Processor Book, One Processor	#4754 - Processor Activation #4754
#7693 - Activation, #8970 or #7587 CoD Processor Book, One Processor	#4754 - Processor Activation #4754
#7815 - Activation #7813, #7731, #7586, or #8969 CoD Processor Books, One Processor	#4754 - Processor Activation #4754
#7990 - Activation, #7988 or #7732 CoD Processor Book, One Processor	#4754 - Processor Activation #4754
#7668 - Activation, #8968 or #7705 CoD Processor Book, One Processor	#4755 - Processor Activation #4755
#7693 - Activation, #8970 or #7587 CoD Processor Book, One Processor	#4755 - Processor Activation #4755
#7815 - Activation #7813, #7731, #7586, or #8969 CoD Processor Books, One Processor	#4755 - Processor Activation #4755
#7990 - Activation, #7988 or #7732 CoD Processor Book, One Processor	#4755 - Processor Activation #4755

Table 1-23 Adapter feature conversions for 9119-595 to 9119-FHA

From feature code	To feature code
#7818 - Remote I/O-2 (RIO-2) Loop Adapter, Two Port	#1814 - Remote I/O-2 (RIO-2) Loop Adapter, Two Port
#7820 - GX Dual-port 12x HCA	#1816 - GX Dual-port 12x HCA

Note: Table 1-24 lists just one feature because all other features are the same as in Table 1-21.

Table 1-24 Additional memory feature conversions for 9119-595 to 9119-FHA

From Feature Code	To Feature Code
#7799 - 256 1 GB Memory Activations for #7835 Memory Cards	#5681 - Activation of 256 GB DDR2 POWER6 Memory

Table 1-25 and Table 1-26 list features involved in the 9119-595 to 9119-FHA model conversion (rack-related, specify codes).

Table 1-25 Feature conversions for 9119-595 to 9119-FHA rack-related

From feature code	To feature code
#5794 - I/O Drawer, 20 Slots, 8 Disk Bays	#5797 - 12X I/O Drawer PCI-X, with repeater
#5794 - I/O Drawer, 20 Slots, 8 Disk Bays	#5798 - 12X I/O Drawer PCI-X, no repeater

Table 1-26 Feature conversions for 9119-595 to 9119-FHA specify codes

From feature code	To feature code
#4643 - 7040-61D I/O Drawer Attachment Indicator	#5809 - Model Upgrade Carry-Over Indicator for converted #4643 with DCA

Conversion within 9119-FHA

Table 1-27, Table 1-28, and Table 1-29 list features involved in model conversion within 9119-FHA.

Table 1-27 Feature conversions for 9119-FHA adapters (within 9119-FHA)

From feature code	To feature code
#1814 - Remote I/O-2 (RIO-2) Loop Adapter, Two Port	#1816 - GX Dual-port 12x HCA
#5778 - PCI-X EXP24 Ctl - 1.5 GB No IOP	#5780 - PCI-X EXP24 Ctl-1.5 GB No IOP
#5778 - PCI-X EXP24 Ctl - 1.5 GB No IOP	#5782 - PCI-X EXP24 Ctl-1.5 GB No IOP

Table 1-28 Processor feature conversions for 9119-FHA (within 9119-FHA)

From feature code	To feature code
#4694 - 0/8-core POWER6 4.2 GHz CoD, 0-core Active Processor Book	#4695 - 0/8-core POWER6 5.0 GHz CoD, 0-core Active Processor Book
#4754 - Processor Activation #4754	#4755 - Processor Activation #4755

Table 1-29 I/O drawer feature conversions for 9119-FHA

From feature code	To feature code
#5791 - I/O Drawer, 20 Slots, 16 Disk Bays	#5797 - 12X I/O Drawer PCI-X, with repeater
#5791 - I/O Drawer, 20 Slots, 16 Disk Bays	#5798 - 12X I/O Drawer PCI-X, no repeater

From type-model 9406-570

Table 1-30, Table 1-31 on page 25, Table 1-32 on page 25, Table 1-32 on page 25, and Table 1-33 on page 26 list the feature codes in a 9406-570 to 9119 FHA model conversion.

Table 1-30 Processor feature conversions for 9406-570 to 9119-FHA processor

From feature code	To feature code
#7618 - 570 One Processor Activation	#4754 - Processor Activation #4754
#7738 - 570 Base Processor Activation	#4754 - Processor Activation #4754

From feature code	To feature code
#7618 - 570 One Processor Activation	#4755 - Processor Activation #4755
#7738 - 570 Base Processor Activation	#4755 - Processor Activation #4755
#7260 - 570 Enterprise Enablement	#4995 - Single #5250 Enterprise Enablement
#7577 - 570 Enterprise Enablement	#4995 - Single #5250 Enterprise Enablement
#9286 - Base Enterprise Enablement	#4995 - Single #5250 Enterprise Enablement
#9299 - Base 5250 Enterprise Enable	#4995 - Single #5250 Enterprise Enablement
#7597 - 570 Full Enterprise Enable	#4996 - Full #5250 Enterprise Enablement
#9298 - Full 5250 Enterprise Enable	#4996 - Full #5250 Enterprise Enablement
#7897 - 570 CUoD Processor Activation	#4754 - Processor Activation #4754
#8452 - 570 One Processor Activation	#4754 - Processor Activation #4754
#7897 - 570 CUoD Processor Activation	#4755 - Processor Activation #4755
#8452 - 570 One Processor Activation	#4755 - Processor Activation #4755

Table 1-31 Administrative feature conversions for 9406-570 to 9119-FHA

From feature code	To feature code
#1654 - 2.2 GHz Processor	#4694 - 0/8-core POWER6 4.2 GHz CoD, 0-core Active Processor Book
#1655 - 2.2 GHz Processor	#4694 - 0/8-core POWER6 4.2 GHz CoD, 0-core Active Processor Book
#1656 - 2.2 GHz Processor	#4694 - 0/8-core POWER6 4.2 GHz CoD, 0-core Active Processor Book
#1654 - 2.2 GHz Processor	#4695 - 0/8-core POWER6 5.0 GHz CoD, 0-core Active Processor Book
#1655 - 2.2 GHz Processor	#4695 - 0/8-core POWER6 5.0 GHz CoD, 0-core Active Processor Book
#1656 - 2.2 GHz Processor	#4695 - 0/8-core POWER6 5.0 GHz CoD, 0-core Active Processor Book

Table 1-32 Capacity on Demand feature conversions for 9406-570 to 9119-FHA

From feature code	To feature code
#7950 - 570 1 GB CoD Memory Activation	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#8470 - 570 Base 1 GB Memory Activation	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#7663 - 570 1 GB Memory Activation	#5680 - Activation of 1 GB DDR2 POWER6 Memory

Table 1-33 Memory feature conversions for 9406-570 to 9119-FHA

From feature code	To feature code
#4452 - 2 GB DDR-1 Main Storage	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS-667 MHz- POWER6 CoD Memory
#4453 - 4 GB DDR Main Storage	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS-667 MHz- POWER6 CoD Memory
#4490 - 4 GB DDR-1 Main Storage	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS-667 MHz- POWER6 CoD Memory
#4453 - 4 GB DDR Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS-667 MHz- POWER6 CoD Memory
#4454 - 8 GB DDR-1 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS-667 MHz- POWER6 CoD Memory
#4490 - 4 GB DDR-1 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS-667 MHz- POWER6 CoD Memory
#7890 - 4/8 GB DDR-1 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS-667 MHz- POWER6 CoD Memory
#4454 - 8 GB DDR-1 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4491 - 16 GB DDR-1 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4494 - 16 GB DDR-1 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#7890 - 4/8 GB DDR-1 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4491 - 16 GB DDR-1 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4492 - 32 GB DDR-1 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4494 - 16 GB DDR-1 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4492 - 32 GB DDR-1 Main Storage	#5697 - 0/64 GB DDR2 Memory(4X16 GB) DIMMS, 400 MHz, POWER6 CoD Memory
#7892 - 2 GB DDR2 Main Storage	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#7893 - 4 GB DDR2 Main Storage	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4495 - 4/8 GB DDR2 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#7893 - 4 GB DDR2 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#7894 - 8 GB DDR2 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4495 - 4/8 GB DDR2 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory

From feature code	To feature code
#4496 - 8/16 GB DDR2 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4497 - 16 GB DDR2 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4499 - 16 GB DDR2 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#7894 - 8 GB DDR2 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4496 - 8/16 GB DDR2 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4497 - 16 GB DDR2 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4498 - 32 GB DDR2 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4499 - 16 GB DDR2 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4498 - 32 GB DDR2 Main Storage	#5697 - 0/64 GB DDR2 Memory(4X16 GB) DIMMS, 400 MHz, POWER6 CoD Memory

From type-model 9406-595

Table 1-34, Table 1-35 on page 28, Table 1-36 on page 28, Table 1-37 on page 28, Table 1-38 on page 29, and Table 1-39 on page 29 list the feature codes involved in 9406-595 to 9119-FHA model conversion.

Table 1-34 Feature conversions for 9406-595 processor features

From feature code	To feature code
#7668 - 595 One Processor Activation	#4754 - Processor Activation #4754
#7815 - 595 One Processor Activation	#4754 - Processor Activation #4754
#7925 - 595 One Processor Activation	#4754 - Processor Activation #4754
#7668 - 595 One Processor Activation	#4755 - Processor Activation FC4755
#7815 - 595 One Processor Activation	#4755 - Processor Activation FC4755
#7925 - 595 One Processor Activation	#4755 - Processor Activation FC4755
#7261 - 595 Enterprise Enablement	#4995 - Single #5250 Enterprise Enablement
#7579 - 595 Enterprise Enablement	#4995 - Single #5250 Enterprise Enablement
#9286 - Base Enterprise Enablement	#4995 - Single #5250 Enterprise Enablement
#9299 - Base 5250 Enterprise Enable	#4995 - Single #5250 Enterprise Enablement
#7259 - 595 Full Enterprise Enable	#4996 - Full #5250 Enterprise Enablement
#7598 - 595 Full Enterprise Enable	#4996 - Full #5250 Enterprise Enablement
#9298 - Full 5250 Enterprise Enable	#4996 - Full #5250 Enterprise Enablement

Table 1-35 Feature conversions for 9406-595 adapters

From feature code	To feature code
#7818 - HSL-2/RIO-G 2-Ports Copper	#1814 - Remote I/O-2 (RIO-2) Loop Adapter, Two Port

Table 1-36 Feature conversions for 9406-595 to 9119-FHA Capacity on Demand

From feature code	To feature code
#7669 - 1 GB DDR2 Memory Activation	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#7280 - 256 GB DDR2 Memory Activation	#5681 - Activation of 256 GB DDR2 POWER6 Memory
#7970 - 595 1 GB CUoD Memory Activation	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#8460 - 595 Base 1 GB Memory Activation	#5680 - Activation of 1 GB DDR2 POWER6 Memory
#7663 - 595 256 GB Memory Activation	#5681 - Activation of 256 GB DDR2 POWER6 Memory

Table 1-37 Feature conversions for 9406-595 to 9119-FHA memory features

From feature code	To feature code
#4500 - 0/4 GB DDR2 Main Storage	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4500 - 0/4 GB DDR2 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4501 - 0/8 GB DDR2 Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz-POWER6 CoD Memory
#4501 - 0/8 GB DDR2 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz- POWER6 CoD Memory
#4502 - 0/16 GB DDR2 Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#4502 - 0/16 GB DDR2 Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4503 - 0/32 GB DDR2 Main Storage	#5696 - 0/32 GB DDR2 Memory(4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#4503 - 0/32 GB DDR2 Main Storage	#5697 - 0/64 GB DDR2 Memory(4X16 GB) DIMMS, 400 MHz, POWER6 CoD Memory
#7816 - 2/4 GB CUoD Main Storage	#5693 - 0/4 GB DDR2 Memory (4X1 GB) DIMMS-667 MHz- POWER6 CoD Memory
#7816 - 2/4 GB CUoD Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz- POWER6 CoD Memory
#7835 - 4/8 GB CUoD Main Storage	#5694 - 0/8 GB DDR2 Memory (4X2 GB) DIMMS- 667 MHz- POWER6 CoD Memory
#7828 - 16 GB Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory

From feature code	To feature code
#7835 - 4/8 GB CUoD Main Storage	#5695 - 0/16 GB DDR2 Memory (4X4 GB) DIMMS- 533 MHz-POWER6 CoD Memory
#7828 - 16 GB Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#7829 - 32 GB Main Storage	#5696 - 0/32 GB DDR2 Memory (4X8 GB) DIMMS- 400 MHz-POWER6 CoD Memory
#7829 - 32 GB Main Storage	#5697 - 0/64 GB DDR2 Memory(4X16 GB) DIMMS, 400 MHz, POWER6 CoD Memory

Table 1-38 Feature conversions for 9406-595 to 9119-FHA miscellaneous

From feature code	To feature code
#8195 - 256 GB Main Storage (32x8)	#8201 - 0/256 GB 533 MHz DDR2 Memory Package (32x#5694)
#8197 - 512 GB Main Storage (32x16)	#8203 - 0/512 GB 533 MHz DDR2 Memory Package (32x#5695)
#8198 - 512 GB Main Storage (16x32)	#8204 - 0/512 GB 400 MHz DDR2 Memory Package (16x#5696)

Table 1-39 Feature conversions for 9406-595to 9119-FHA specify codes

From feature code	To feature code
#4643 - 7040-61D I/O Drawer Attachment Indicator	#5809 - Model Upgrade Carry-Over Indicator for converted #4643 with DCA

1.4 Racks power and cooling

A Power 595 system uses racks to house its components:

- The 24-inch rack system rack includes an integrated power subsystem called the Bulk Power Assemblies (BPA) that is located at the top of the rack on both the front and rear sides. The system rack provides a total of 42U of rack-mounting space and also houses the CEC and its components.
- A powered Expansion Rack (#6954) is available for larger system configurations that require additional 24-inch I/O Expansion Drawers beyond the three (without battery backup) that are available in the system rack. It provides an identical redundant power subsystem as that available in the system rack. The PCI Expansion Drawers (#5797, #5791, and #5792) can be used with rack feature 6954. The 12X PCI drawer #5798 is supported only in the system rack.
- An nonpowered Expansion Rack (#6953) is available if additional 24-inch rack space is required. To install the Expansion Rack feature, the side cover of the powered Expansion Rack is removed, the Expansion Rack (#6953) is bolted to the side, and the side cover is placed on the exposed side of the Expansion Rack (#6953). Power for components in the Expansion Rack is provided from the bulk power assemblies in the powered Expansion Rack.

> **Note:** One nonpowered Expansion Rack can be attached to each Powered Expansion rack. The nonpowered Expansion Rack (#6953) cannot be attached to the 595 system rack.

Additional requirements are as follows:

- The 12X I/O drawer (#5797) can be used for additional I/O capacity in the 595 system rack and both of the 595 Expansion Racks (#6953 and #6954). The 12X I/O drawer (#5798) can only be used in the system rack.
- The 9119-590/595 PCI Expansion Drawers (#5791 and #5794) can be used with the 595 system rack and Expansion Racks (#6953 and #6954).
- Although not available for new orders, the 9119-595 powered Expansion Rack (#5792) can also be used for additional 24-inch I/O drawer expansion. The powered Expansion Rack (#5792) only supports the RIO-2 I/O Drawers (#5791 and #5794). It does not support the 12X I/O Drawers (#5797 nor #5798). When the 9119-595 powered Expansion Rack (#5792) is used, the nonpowered Expansion Rack (#8691) can be used for additional I/O expansion. The feature 8691 rack is bolted onto the powered Expansion Rack (#5792).
- The 9119-595 Expansion Racks (#5792) do not support additional I/O expansion using the 12X PCI Drawers (#5797 and #5798).
- The 9119-595 Expansion Racks (#5792 and #8691) only support the RIO-2 I/O Drawers (#5791 and #5794).
- All 24-inch, 595 racks and expansion feature racks must have door assemblies installed. Door kits containing front and rear doors are available in either slimline, acoustic, rear heat exchanger, or acoustic rear heat exchanger styles.

Additional disk expansion for IBM i partitions is available in a 42U high, 19-inch Expansion Rack (#0553). Both the feature #5786 SCSI (4U) and feature #5886 SAS drawers can be mounted in this rack. Also available is the PCI-X Expansion Drawer (#5790). A maximum of four I/O bus adapters (#1814) are available in each CEC processor book for the PCI-X Expansion Drawer (#5790). The Expansion Drawer (#5790) must include a #6438, dual-port RIO-G adapter which is placed into a PCI slot.

> **Note:** The height of the system rack or Expansion Rack features (42U) might require special handling when shipping by air or when moving under a low doorway.

1.4.1 Door kit

The slimline door kit provides a smaller footprint alternative to the acoustical doors for those clients who might be more concerned with floor space than noise levels. The doors are slimmer because they do not contain acoustical treatment to attenuate the noise.

The acoustical door kit provides specially designed front and rear acoustical doors that greatly reduce the noise emissions from the system and thereby lower the noise levels in the data center. The doors include acoustically absorptive foam and unique air inlet and exhaust ducts to attenuate the noise. This is the default door option.

The non-acoustical front door and rear door heat exchanger kit provides additional cooling to reduce environmental cooling requirements for the 595 server installation. This feature provides a smaller footprint alternative to the acoustical doors along with a Rear Door Heat Exchanger for those clients who might be more concerned with floor space than noise levels and also want to reduce the environmental cooling requirements for the 595 server.

The acoustical front door and rear door heat exchanger kit provides both additional cooling and acoustical noise reduction for use where a quieter environment is desired along with additional environmental cooling. This feature provides a specially designed front acoustical door and an acoustical attachment to the Rear Door Heat Exchanger door that reduce the noise emissions from the system and thereby lower the noise levels in the data center. Acoustically absorptive foam and unique air inlet and exhaust ducts are employed to attenuate the noise.

Note: Many of our clients prefer the reduction of ambient noise through the use of the acoustic door kit.

1.4.2 Rear door heat exchanger

The Power 595 systems support the rear door heat exchanger (#6859) similar to the one used in POWER5 based 590/595 powered system racks. The rear door heat exchanger is a water-cooled device that mounts on IBM 24-inch racks. By circulating cooled water in sealed tubes, the heat exchanger cools air that has been heated and exhausted by devices inside the rack. This cooling action occurs before the air leaves the rack unit, thus limiting the level of heat emitted into the room. The heat exchanger can remove up to 15 kW (approximately 50,000 BTU/hr) of heat from the air exiting the back of a fully populated rack. This allows a data center room to be more fully populated without increasing the room's cooling requirements. The rear door heat exchanger is an optional feature.

1.4.3 Power subsystem

The Power 595 uses redundant power throughout its design. The power subsystem in the system rack is capable of supporting 595 servers configured with one to eight processor books, a media drawer, and up to three I/O drawers.

The system rack and powered Expansion Rack always incorporate two bulk power Assemblies (BPA) for redundancy. These provide 350 V dc power for devices located in those racks and associated nonpowered Expansion Racks. These bulk power assemblies are mounted in front and rear positions and occupy the top 8U of the rack. To help provide optimum system availability, these bulk power assemblies should be powered from separate power sources with separate line cords.

Redundant Bulk Power Regulators (BPR #6333) interface to the bulk power assemblies to help ensure proper power is supplied to the system components. Bulk power regulators are always installed in pairs in the front and rear bulk power assemblies to provide redundancy. The number of bulk power regulators required is configuration-dependent based on the number of processor MCMs and I/O drawers installed.

A Bulk Power Hub (BPH) is contained in each of the two BPAs. Each BPH contains 24 redundant Ethernet ports. The following items are connected to the BPH:

- Two (redundant) System Controller service processors (SC),
- One HMC (an additional connection port is provided for a redundant HMC)
- Bulk Power Controllers
- Two (redundant) Node Controller (NC) service processors for each processor book

The 595 power subsystem implements redundant bulk power assemblies (BPA), Bulk Power Regulators (BPR, #6333), Power controllers, Power distribution assemblies, dc power converters, and associated cabling. Power for the 595 CEC is supplied from dc bulk power assemblies in the system rack. The bulk power is converted to the power levels required for the CEC using dc to dc power converters (DCAs).

Additional Power Regulators (#6186) are used with the p5 Powered Expansion Rack (#5792), when needed. Redundant Bulk Power Distribution (BPD) Assemblies (#6334) provide additional power connections to support the system cooling fans dc power converters contained in the CEC and the I/O drawers. Ten connector locations are provided by each power distribution assembly. Additional BPD Assemblies (#7837) are provided with the p5 Powered Expansion Rack (#5792), when needed.

An optional Integrated Battery Backup feature (IBF) is available. The battery backup feature is designed to protect against power line disturbances and provide sufficient, redundant power to allow an orderly system shutdown in the event of a power failure. Each IBF unit occupies both front and rear positions in the rack. The front position provides primary battery backup; the rear positions provides redundant battery backup. These units are directly attached to the system bulk power regulators. Each IBF is 2U high and will be located in the front and rear of all powered racks (system and Powered Expansion). The IBF units displace an I/O drawer at location U9 in each of these racks.

1.5 Operating system support

The Power 595 supports the following levels of IBM AIX, IBM i, and Linux operating systems:

- AIX 5.3 with the 5300-06 Technology Level and Service Pack 7, or later
- AIX 5.3 with the 5300-07 Technology Level and Service Pack 4, or later
- AIX 5.3 with the 5300-08 Technology Level, or later
- AIX 6.1 with the 6100-00 Technology Level and Service Pack 5, or later
- AIX 6.1 with the 6100-01 Technology Level, or later
- IBM i 5.4 (formerly known as i5/OS V5R4), or later
- IBM i 6.1 (formerly known as i5/OS V6R1), or later
- Novell® SUSE Linux Enterprise Server 10 Service Pack 2 for POWER, or later
- Red Hat Enterprise Linux 4.7 for POWER, or later
- Red Hat Enterprise Linux 5.2 for POWER, or later

Note: Planned availability for IBM i is September 9, 2008. Planned availability for SUSE Linux Enterprise Server 10 for POWER and Red Hat Enterprise Linux for POWER is October 24, 2008.

For the IBM i operating system, a console choice must be specified which can be one of the following:

- Operations console attached via Ethernet port (LAN console) or WAN port (*ops console*)
- Hardware Management Console (HMC)

IBM periodically releases fixes, group fixes and cumulative fix packages for IBM AIX and IBM i operating systems. These packages can be ordered on CD-ROM or downloaded from:

http://www.ibm.com/eserver/support/fixes/fixcentral

Select a product (hardware) family. For the 595 server, select Power.

A sequence of selection fields is available with each entry you select. Selection fields include an operating system (for example IBM i, AIX, Linux) and other software categories that include microcode, firmware, and others. For most options you must select a release level.

The Fix Central Web site provides information about how to obtain the software using the media (for example, the CD-ROM).

You can also use the Fix Central Web site to search for and download individual operating system fixes licensed program fixes, and additional information.

Part of the fix processing includes Fix Central dialoguing with your IBM i or AIX operating system to identify fixes already installed, and whether additional fixes are required.

1.5.1 IBM AIX 5.3

When installing AIX 5L™ 5.3 on the 595 server, the following minimum requirements must be met:

- AIX 5.3 with the 5300-06 Technology Level and Service Pack 7, or later
- AIX 5.3 with the 5300-07 Technology Level and Service Pack 4, or later
- AIX 5.3 with the 5300-08 Technology Level, or later

IBM periodically releases maintenance packages (service packs or technology levels) for the AIX 5L operating system. These packages can be ordered on CD-ROM or downloaded from:

http://www.ibm.com/eserver/support/fixes/fixcentral/main/pseries/aix

The Fix Central Web site also provides information about how to obtain the software via the media (for example, the CD-ROM).

You can also get individual operating system fixes and information about obtaining AIX 5L service at this site. From AIX 5L V5.3 the Service Update Management Assistant (SUMA), which helps the administrator to automate the task of checking and downloading operating system downloads, is part of the base operating system. For more information about the **suma** command, refer to:

http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.cmds/doc/aixcmds5/suma.htm

AIX 5L is supported on the System p servers in partitions with dedicated processors (LPARs), and shared-processor partitions (SPLARs). When combined with one of the PowerVM features, AIX 5L Version 5.3 or later can make use of all the existing and new virtualization features such as micro-partitioning technology, virtual I/O, virtual LAN, and PowerVM Live Partition Mobility.

1.5.2 IBM AIX V6.1

IBM AIX 6.1 is the most recent version of AIX and includes significant new capabilities for virtualization, security features, continuous availability features, and manageability. The system must meet the following minimum requirements before you install AIX 6.1 on the 595 server:

- AIX 6.1 with the 6100-00 Technology Level and Service Pack 5, or later
- AIX 6.1 with the 6100-01 Technology Level, or later

AIX V6.1 features include support for:

- PowerVM AIX 6 Workload Partitions (WPAR) - software based virtualization
- Live Application Mobility - with the IBM PowerVM AIX 6 Workload Partitions Manager™ for AIX (5765-WPM)
- 64-bit Kernel for higher scalability and performance
- Dynamic logical partitioning and Micro-Partitioning support
- Support for Multiple Shared-Processor Pools
- Trusted AIX - MultiLevel, compartmentalized security
- Integrated Role Based Access Control
- Encrypting JFS2 file system
- Kernel exploitation of POWER6 Storage Keys for greater reliability
- Robust journaled file system and Logical Volume Manager (LVM) software including integrated file system snapshot
- Tools for managing the systems environment:
  - System Management Interface Tool (SMIT)
  - IBM Systems Director Console for AIX

1.5.3 IBM i 5.4 (formerly IBM i5/OS V5R4)

IBM i 5.4 contains a wide range of medium to small enhancements and new functions built on top of IBM i's integrated work management, performance management, database (DB2® for i5/OS), security, backup and recovery functions and System i Navigator graphical interface to these functions. When installing IBM i 5.4 on the 595 server, the following minimum requirements must be met:

- IBM i 5.4 (formerly known as i5/OS V5R4), or later

IBM i 5.4 enhancements include:

- Support of POWER6 processor technology models
- Support of large write cache disk controllers (IOAs)
- Expanded support of IOAs that do not require IOPs
- More flexible back up and recovery options and extended support in local remote journaling and cross-site mirroring and clustering
- Expanded DB2 and SQL functions and graphical management
- IBM Control Language (CL) extensions
- Initial release support of IBM Technology for Java™ 32-bit JVM™
- Support for IBM Express Runtime Web Environments for i5/OS, which contains a wide range of capabilities intended to help someone new to or just beginning to use the Web get started and running in a Web application serving environment.
- Expanded handling of 5250 workstation applications running in a Web environment via the WebFacing and HATS components
- Licensed program enhancements include Backup Recovery and Media Services, and application development enhancements including RPG, COBOL, and C/C++.

Note: IBM i 5.4 has a planned availability date of September 9, 2008.

1.5.4 IBM i 6.1

Before you install IBM i 6.1 on the 595 server, your system must meet the following minimum requirements:

- IBM i 6.1 (formerly known as i5/OS V6R1), or later

As with previous releases, 6.1 builds on top of the IBM i integrated capabilities with enhancement primarily in the following areas:

- IBM i security, including greatly expanded data encryption/decryption and network intrusion detection
- Support for the IBM PCI-X (#5749) and PCIe Fibre Channel (#5774) IOP-less adapters and a new performance improved code path for attached IBM System Storage™ DS8000™ configurations
- Expanded base save/restore, journaling and clustering support
- New IBM high availability products that take advantage of the expanded 6.1 save/restore, journaling and clustering support
- System i PowerHA for i (formerly known as High Availability Solutions Manager (HASM)) and IBM iCluster for i
- Logical partitioning extensions including support of multiple shared processor pools and IBM i 6.1 as a client partition to another 6.1 server partition or a server IBM Virtual I/O Server partition. The VIOS partition can be on a POWER6 server or a POWER6 IBM Blade JS22 or JS12.
- Expanded DB2 and SQL functions, graphical management of the database, and generally improved performance
- Integrated Web application server and Web Services server (for those getting started with Web services
- Integrated browser-based IBM Systems Director Navigator for i5/OS that includes a new Investigate Performance Data graphically function
- Initial release support of IBM Technology for Java 64-bit JVM
- RPG COBOL, and C/C++ enhancements as well as new packaging of the application development tools: the WebSphere® Development Studio and Rational® Developer suite of tools

Note: IBM i 6.1 has a planned availability date of September 9, 2008.

1.5.5 Linux for Power Systems summary

Linux is an open source operating system that runs on numerous platforms from embedded systems to mainframe computers. It provides a UNIX-like implementation across many computer architectures.

The supported versions of Linux for Power systems include the following brands to be run in partitions:

- Novell SUSE Linux Enterprise Server 10 Service Pack 2 for POWER, or later
- Red Hat Enterprise Linux 4.7 for POWER, or later
- Red Hat Enterprise Linux 5.2 for POWER, or later

The PowerVM features are supported in Version 2.6.9 and above of the Linux kernel. The commercially available latest distributions from Red Hat Enterprise (RHEL AS 5) and Novell SUSE Linux (SLES 10) support the IBM System p 64-bit architectures and are based on this 2.6 kernel series.

Clients who want to configure Linux partitions in virtualized System p systems should consider the following:

- Not all devices and features supported by the AIX operating system are supported in logical partitions running the Linux operating system.
- Linux operating system licenses are ordered separately from the hardware. Clients can acquire Linux operating system licenses from IBM, to be included with their 595 server or from other Linux distributors.

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

http://www-03.ibm.com/systems/p/os/linux/index.html

For information about SUSE Linux Enterprise Server 10, refer to:

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

For information about Red Hat Enterprise Linux Advanced Server 5, refer to:

http://www.redhat.com/rhel/features

Supported virtualization features

SLES 10, RHEL AS 4.5 and RHEL AS 5 support the following virtualization features:

- Virtual SCSI, including for the boot device
- Shared-processor partitions and virtual processors, capped and uncapped
- Dedicated-processor partitions
- Dynamic reconfiguration of processors
- Virtual Ethernet, including connections through the Shared Ethernet Adapter in the Virtual I/O Server to a physical Ethernet connection
- Simultaneous multithreading

SLES 10, RHEL AS 4.5, and RHEL AS 5 do not support the following:

- Dynamic reconfiguration of memory
- Dynamic reconfiguration of I/O slot

Note: SUSE Linux Enterprise Server 10 for POWER, or later, and Red Hat Linux operating system support has a planned availability date of October 24, 2008. IBM only supports the Linux systems of clients with a SupportLine contract covering Linux. Otherwise, contact the Linux distributor for support.

2

Architectural and technical overview

This chapter discusses the overall system architecture and technical aspects of the Power 595 server. The 595 is based on a modular design, where all components are mounted in 24-inch racks. Figure 2-1 represents the processor book architecture. The following sections describe the major components of this diagram. 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.

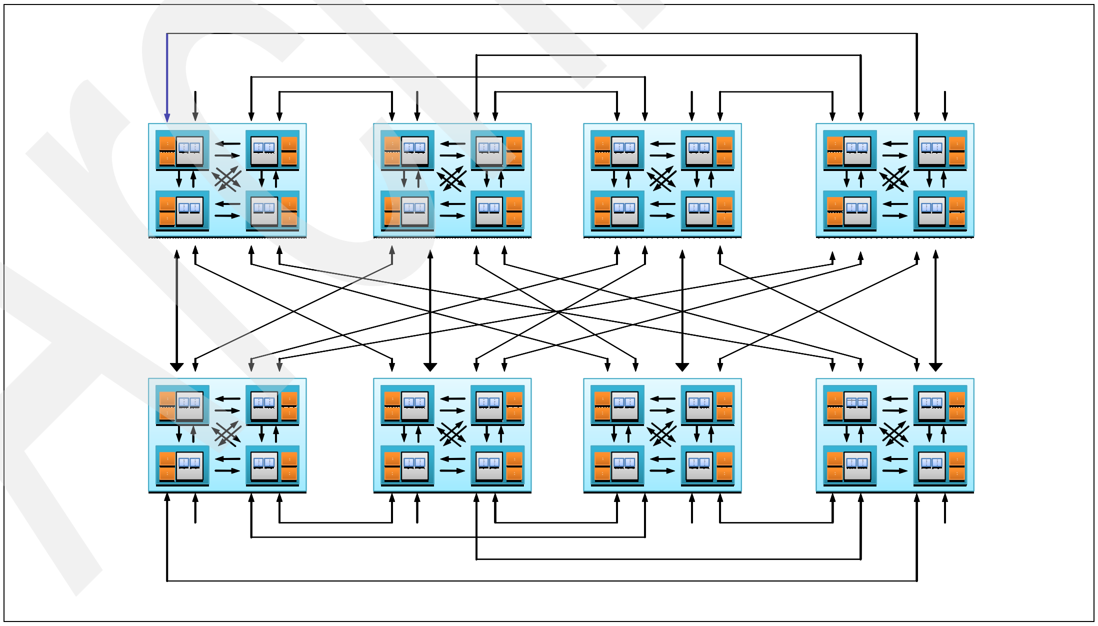


Figure 2-1 IBM Power 595 processor book architecture

2.1 System design

The IBM Power 595 Enterprise Class structure is the result of the continuous evolution of 595 through pSeries®. Its structure and design have been continuously improved, adding more capacity, performance, functionality, and connectivity, while considering the balanced system approach for memory sizes, internal bandwidth, processing capacity, and connectivity. The objective of the 595 system structure and design is to offer a flexible and robust infrastructure to accommodate a wide range of operating systems and applications, either traditional or based on WebSphere, Java, and Linux for integration and deployment in heterogeneous business solutions. The Power 595 is based on a modular design, in which all components are mounted in 24-inch racks. Inside this rack, all the server components are placed in specific positions. This design and mechanical organization offer advantages in optimization of floor space usage.

2.1.1 Design highlights

The IBM Power 595 (9119-FHA) is a high-end POWER6 processor-based symmetric multiprocessing (SMP) system. To avoid any possible confusion with earlier POWER5 model 595 systems, we will hereafter refer to the current system as the Power 595. The Power 595 is based on a modular design, where all components are mounted in 24-inch racks. Inside this rack, all the server components are placed in specific positions. This design and mechanical organization offer advantages in optimization of floor space usage.

Conceptually, the Power 595 is similar to the IBM eServer™ p5 595 and i5 595, which use POWER5 technology, and can be configured in one (primary) or multiple racks (primary plus expansions). The primary Power 595 frame is a 42U, 24-inch based primary rack, containing major subsystems, which as shown in Figure 2-2 on page 39, from top to bottom, include:

- A 42U-tall, 24-inch system rack (primary) houses the major subsystems.
- A redundant power subsystem housed in the Bulk Power Assemblies (BPA's) located in front and rear at the top 8U of the system rack, has optional battery backup capability.
- A 20U-tall Central Electronics Complex (CEC) houses the system backplane cooling fans and system electronic components.
- One to eight, 8-core, POWER6 based processor books are mounted in the CEC. Each processor book incorporates 32 memory dual in-line memory module (DIMM) slots.
- Integrated battery feature (IBF) for backup is optional.
- Media drawer is optional.
- One to 30 I/O drawers each contains 20 PCI-X slots and 16 hot-swap SCSI-3 disk bays.

In addition, depending on the configuration, it is possible to have:

- One or two powered expansion racks, each with 32U worth of rack space for up to eight 4U I/O drawers. Each powered expansion rack supports a 42U bolt-on, nonpowered expansion rack for mounting additional I/O drawers as supported by the 595 I/O expansion rules.
- One to two nonpowered expansion racks, each supporting up to seven I/O drawers, 4Us high.

> **Note:** Nonpowered expansion racks must be attached to a powered expansion rack. Maximum configuration can be up to five racks: One primary rack, two powered expansion racks, and two nonpowered expansion racks

Figure 2-2 details primary rack major subsystems.

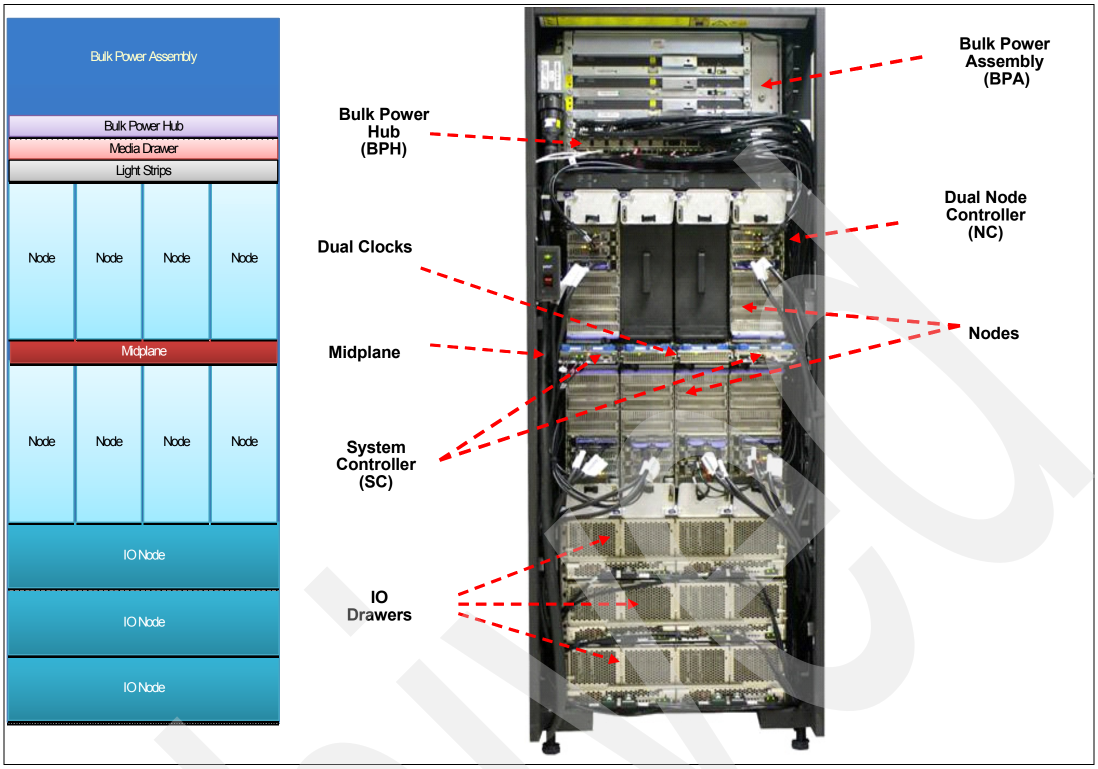


Figure 2-2 Power 595 front view

Figure 2-3 on page 40 shows the four air-moving devices at the rear of the Power 595. Figure 2-25 on page 69 shows the air flow.

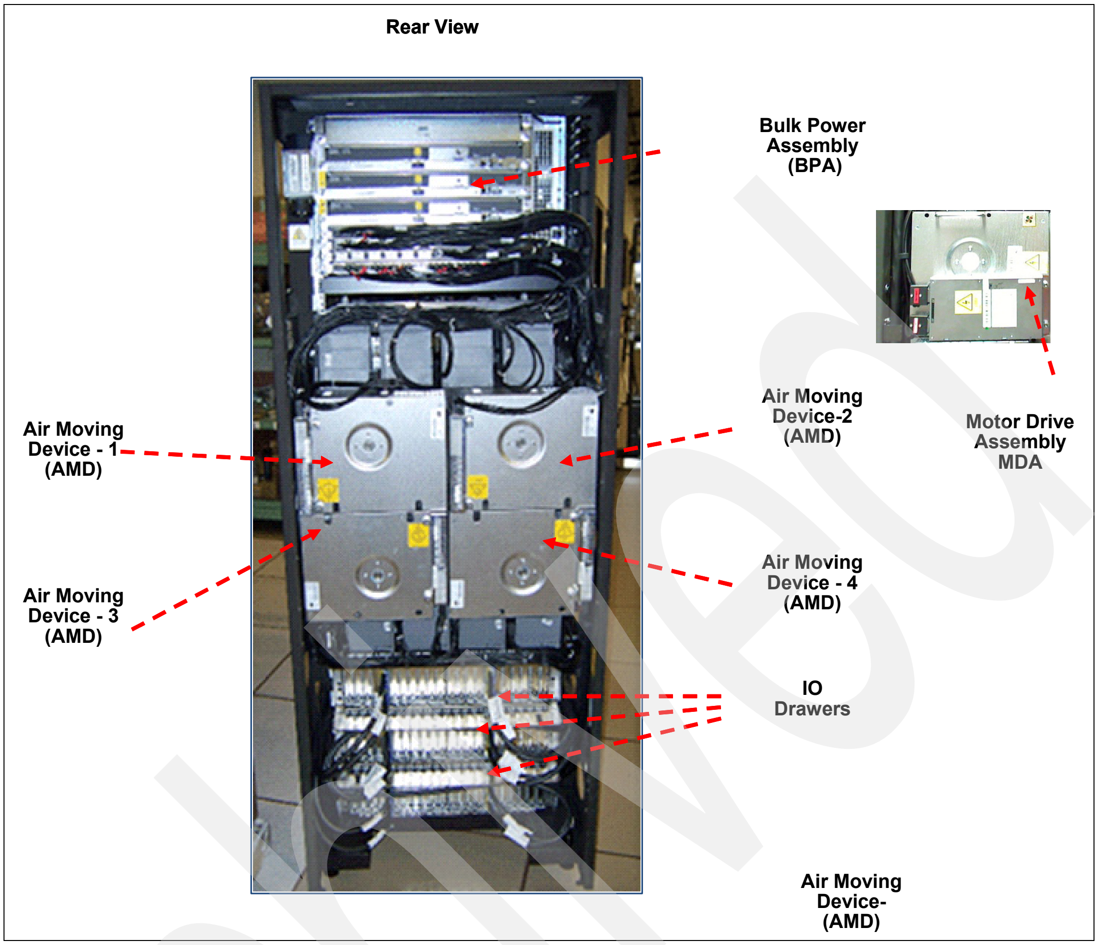


Figure 2-3 Power 595 rear view

2.1.2 Center electronic complex (CEC)

The Power 595 CEC is a 20U tall, 24-inch rack-mounted device. It houses the system processors, memory, redundant system service processors, I/O drawer connection capability, and associated components. The CEC is installed directly below the power subsystem.

The CEC features a packaging concept based on books. The books contain processors memory, and connectors to I/O drawers and other servers. These books will hereafter be referred to as *processor books*. The processor books are located in the CEC, which is mounted in the primary rack.

Each Processor book assembly contains many components, some of which include:

- The processor book planar provides support for four multichip modules (MCM), 32 memory DIMM slots, a connector to the mid-plane, four I/O adapter connectors, two node controller connectors, and one VPD card connector.
- A node power distribution planar provides support for all DCA connectors, memory, and air temperature sensors.

- The processor book VPD card holds the VPD card and SVPD (CoD) for processor book information. Routes sense and control signals pass between the DCAs and processor book planar (DIMM LED Control). The VPS card plugs into processor book planar and the node power distribution planar
- Two Distributed Converter Assemblies (DCAs) located at the back of each processor book.
- Four RIO-2 or 12x I/O hub adapter slots (two wide and two narrow) are located at the front of the processor book.
- Two embedded Node Controller service processor cards (NC) are located in the front of the processor book. The node controller cards communicate to the HMC via the Bulk Power Hub (BPH) and are connected to both front and rear Ethernet ports on the BPH.

The layout of a processor book and its components is shown in Figure 2-4.

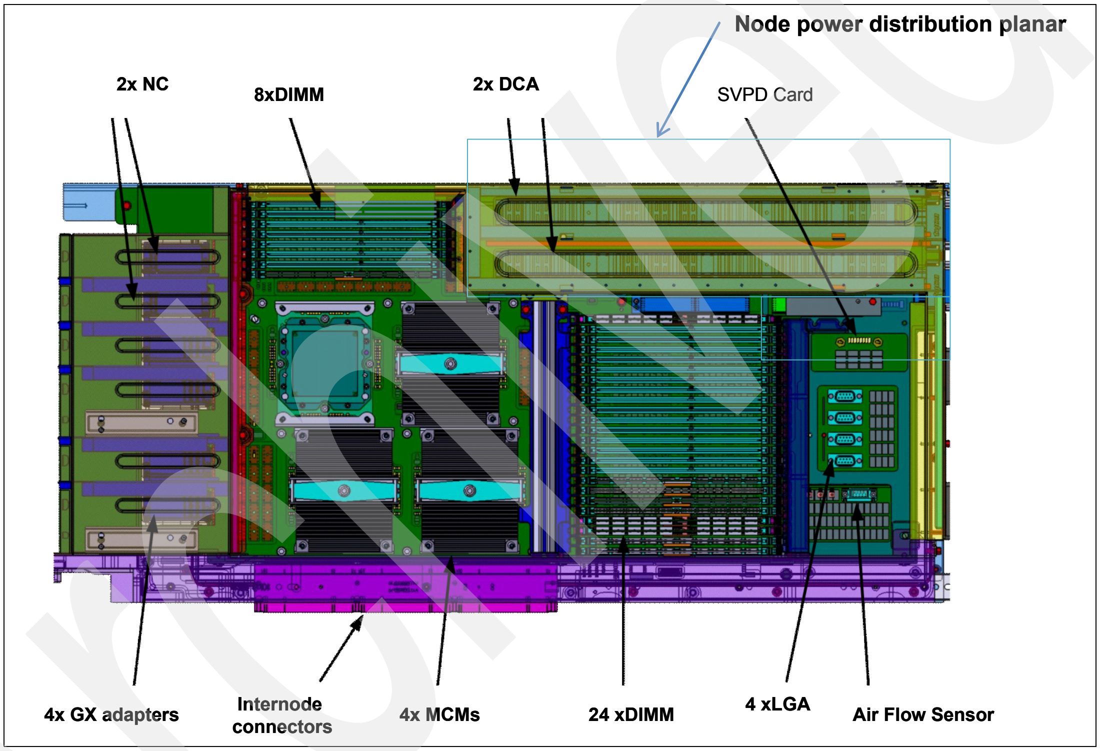


Figure 2-4 processor book cards layout

Processor book placement

Up to eight processor books can reside in the CEC cage. The processor books slide into the mid-plane card, which is located in the middle of the CEC cage. Support is provided for up to four books on top and four books on the bottom of the mid-plane. The processor books are installed in a specific sequence as listed in Table 2-1 on page 42.

Two oscillator (system clock) cards are also connected to the mid-plane. One oscillator card operates as the primary and the other as a backup. In case the primary oscillator would fail, the backup card detects the failure and continues to provide the clock signal so that no outage occurs due to an oscillator failure.

Table 2-1 Processor book installation sequence

Plug sequence	PU book	Location code	Orientation
1	Book 1	U*n*-P9	Bottom
2	Book 2	U*n*-P5	Top
3	Book 3	U*n*-P6	Bottom
4	Book 4	U*n*-P2	Top
5	Book 5	U*n*-P7	Bottom
6	Book 6	U*n*-P8	Bottom
7	Book 7	U*n*-P3	Top
8	Book 8	U*n*-P4	Top

Figure 2-5 details the processor book installation sequence.

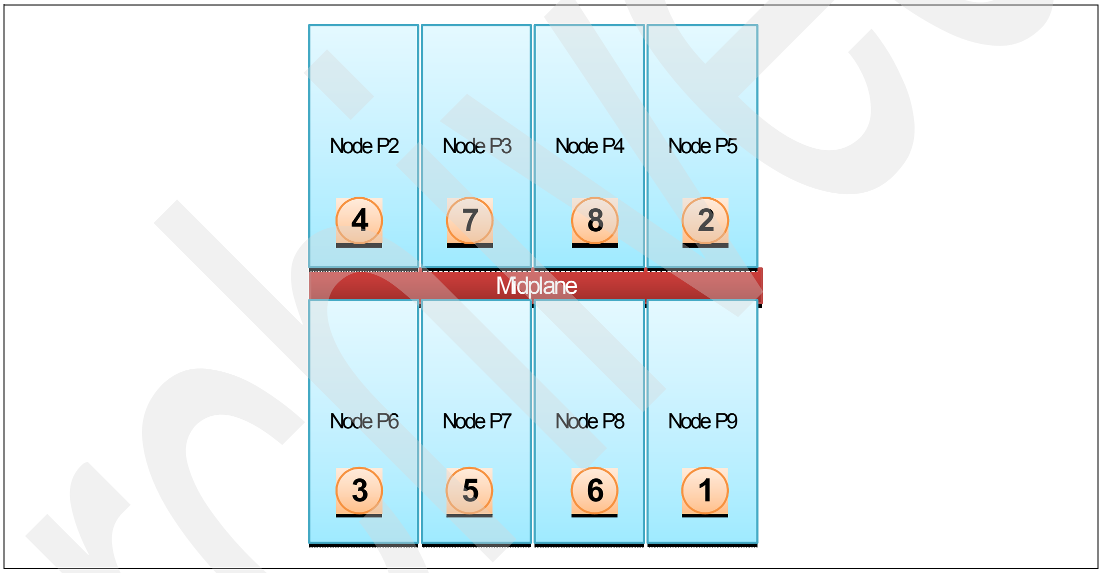


Figure 2-5 Processor Book layout

2.1.3 CEC midplane

The POWER6 595 CEC midplane holds the processor books in an ingenious way: four attach to the top of the midplane and four to the bottom. A Node Actualization Mechanism (NAM) raises or lowers processing unit (PU) Books into position. After a processor book is aligned correctly and fully seated, the Node Locking Mechanism (NLM) secures the book in place. The midplane assembly contains:

- Eight processor node slots (4 Upper / 4 Lower)
- Two system controllers (SC)
- One VPD-card, which contains the CEC cage VPD information
- One VPD-anchor (SVPD) card, which contains the anchor point VPD data. Dual Smartchips

Figure 2-6 on page 43 shows the CEC midplane layout.

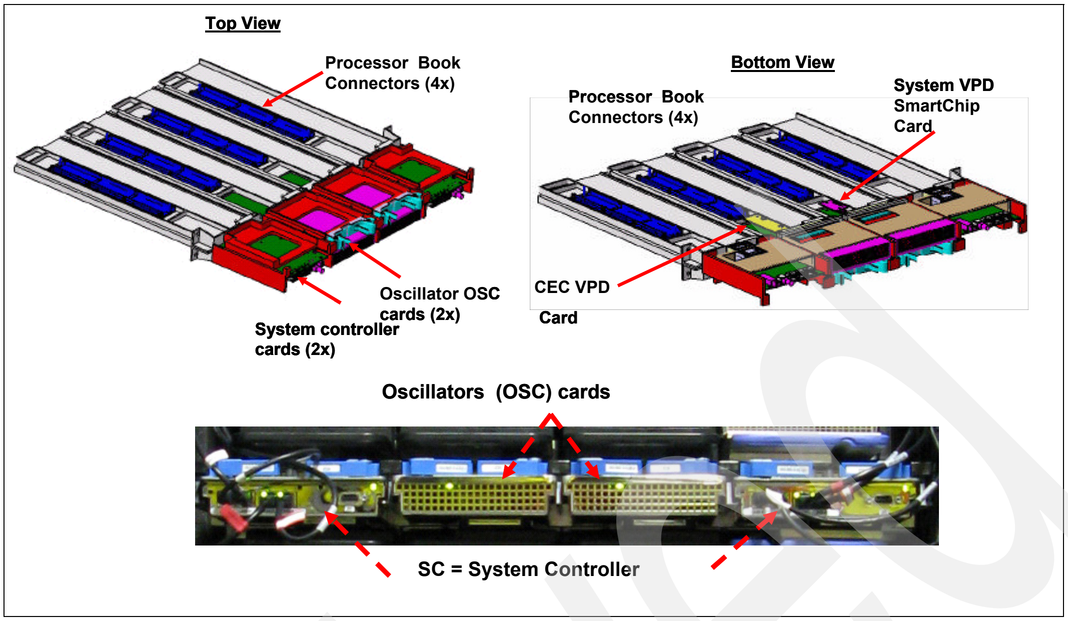

Top View
Processor Book Connectors (4x)
Oscillator OSC cards (2x)
System controller cards (2x)
Bottom View
Processor Book Connectors (4x)
System VPD SmartChip Card
CEC VPD Card
Oscillators (OSC) cards
SC = System Controller

Figure 2-6 CEC midplane

Table 2-2 lists the CEC midplane component locations codes.

Table 2-2 CEC location codes

Location code	Component
U*n*-P1	CEC Midplane
U*n*-P1-C1	System VPD anchor card
U*n*-P1-C2	CEC System Controller SC card 0
U*n*-P1-C3	CEC Oscillator card 0
U*n*-P1-C4	CEC Oscillator card 1
U*n*-P1-C5	System Controller SC card 1

2.1.4 System control structure (SCS)

The Power 595 has the ability to run multiple different operating systems on a single server. Therefore, a single instance of an operating system is no longer in full control of the underlying hardware. As a result, a system control task running on an operating system that does not have exclusive control of the server hardware can no longer perform operations that were previously possible. For example, what would happen if a control task, in the course of recovering from an I/O error, were to decide to reset the disk subsystem? Data integrity might no longer be guaranteed for applications running on another operating system on the same hardware.

As a solution, system-control operations for large systems must be moved away from the resident operating systems and be integrated into the server at levels where full control over the server remains possible. System control is therefore increasingly delegated to a set of other helpers in the system outside the scope of the operating systems. This method of host operating system-independent system management is often referred to as *out-of-band control*, or out-of-band system management.

The term system control structure (SCS) describes an area within the scope of hardware platform management. It addresses the lower levels of platform management, for example the levels that directly deal with accessing and controlling the server hardware. The SCS implementation is also called the *out-of band service subsystem*.

The SCS can be seen as key infrastructure for delivering mainframe RAS characteristics (for example CoD support functions, on-chip array sparing) and error detection, isolation, and reporting functions (Instant failure detection, Isolation of failed parts, continued system operation, deferred maintenance, Call-home providing detailed problem analysis, pointing to FRU to be replaced. The SCS provides system initialization and error reporting and facilitates service. Embedded service processor-based control cards reside in the CEC cage (redundant System Controllers-SC), node (redundant Node Controllers-NC), and in the BPC.

Figure 2-7 shows a high-level view of a Power 595, together with its associated control structure. The system depicted to the right is composed of CEC with many processor books and I/O drawers.

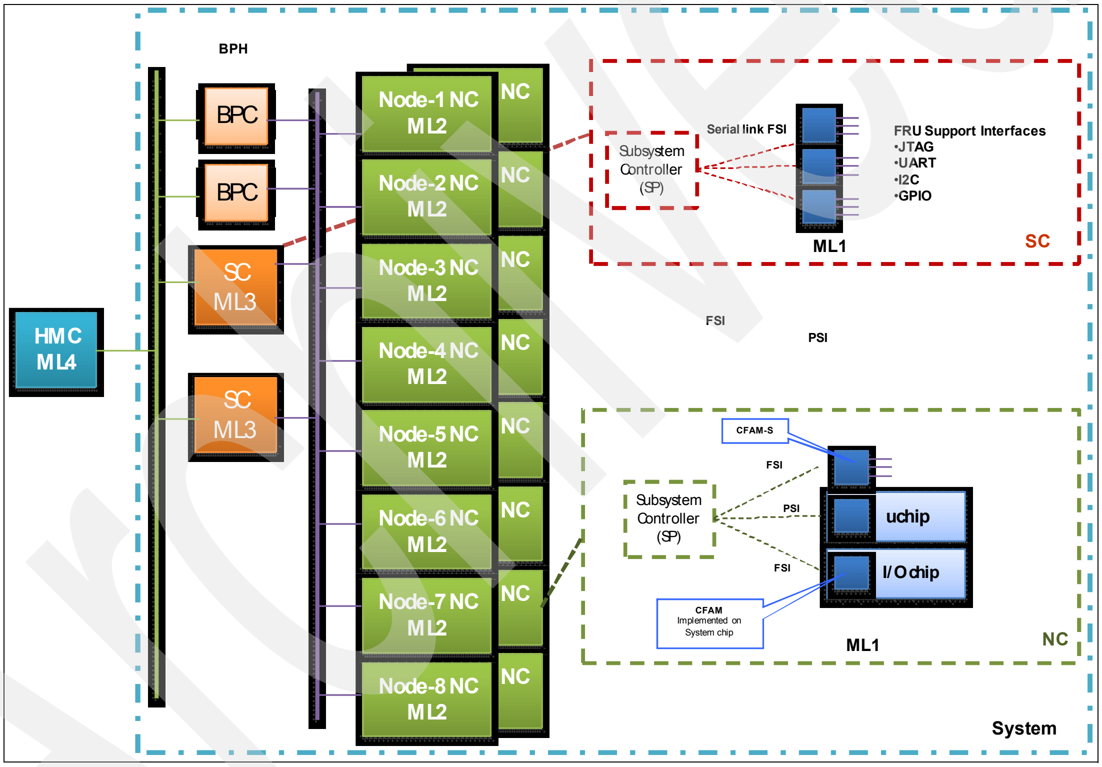


Figure 2-7 System Control Structure (SCS) architecture

The System Controllers's scope is to manage one system consisting of one or more subsystems such as processor books (called *nodes*), I/O drawers and the power control subsystem.

In addition to the functional structure, Figure 2-7 also shows a system-control infrastructure that is orthogonal to the functional structure.

To support management of a truly modular server hardware environment, the management model must have similar modular characteristics, with pluggable, standardized interfaces. his required the development of a rigorously modular management architecture, which organizes the management model in the following ways:

- Groups together the management of closely related subsets of hardware elements and logical resources.
- Divides the management into multiple layers, with operations of increasing breadth and scope.
- Implements the management layering concepts consistently throughout the distributed control structures of the system (rather than viewing management as something that is added on top of the control structures).
- Establishes durable interfaces, open interfaces within and between the layers

The Figure 2-7 on page 44 also shows that SCS is divided into management domains or management levels, as follows:

- Management Level 1 domain (ML1): This layer refers to hardware logic and chips present on the circuit boards (actuators and sensors used to perform node-control operations.
- Management Level 2 domain (ML2): This is management of a single subsystem (for example processor book) or node instance within a system.The ML2 layer controls the devices of such a subsystem through device interfaces (for example, FSI, PSI) other than network services. The devices are physical entities attached to the node. Controlling a node has the following considerations:
  - Is limited to strict intra-node scope.
  - Is not aware of anything about the existence of a neighbor node.
  - Is required to maintain steady-state operation of the node.
  - Does not maintain persistent state information.
- Management Level 3 domain (ML3): Platform management of a single system instance, comprises all functions within a system scope. Logical unit is responsible for a system and controlling all ML2s through network interfaces; acts as state aggregation point for the super-set of the individual ML2 controller states. Managing a system (local to the system) requires the following considerations:
  - Controls a system.
  - Is the service focal point for the system being controlled.
  - Aggregates multiple nodes to form a system.
  - Exports manageability to management consoles.
  - Implements the firewall between corporate intranet and private service network.
  - Facilitates persistency for:
    - Firmware code loads.
    - Configuration data.
    - Capturing of error data.
- Management Level 4 domain (ML4): Set of functions that can manage multiple systems; can be located apart from the system to be controlled. (HMC level)

The Power System HMC implements ML4 functionalities.

The relationship between the ML2 layer (NC) and ML3 layer (SC) is such that the ML3 layer's function set controls a system, which consists of one or more ML2 layers. The ML3 layer's function set exists once per system, while there is one ML2 layer instantiation per node. The ML2 layer operates under the guidance of the ML3 layer, for example, the ML3 layer is the manager and ML2 layer is the agent or the manageable unit. ML3 layer functions submit transactions that are executed by the ML2 layer.

The System Controllers's scope is, as reported before, to manage one system consisting of one or more subsystems such as processor books (called *nodes*), I/O drawers and the power control subsystem. The system control structure (SCS) is implemented with complete redundancy. This includes the service processors, interfaces and VPD and smart chip modules.

The SC communicates exclusively via TCP/IP over Ethernet through the Bulk Power Hub (BPH), is implemented as a service processor embedded controller. Upstream it communicates with the HMC, downstream it communicates with the processor book subsystem controllers called Node Controllers (NC). The NC is also implemented as a service processor embedded controller.

Each processor book cage contains two embedded controllers called Node Controllers (NCs), which interface with all of the logic in the corresponding book. Two NCs are used for each processor book to avoid any single point of failure. The controllers operate in master and subordinate configuration. At any given time, one controller performs the master role while the other controller operates in standby mode, ready to take over the master's responsibilities if the master fails. Node controller boot over the network from System Controller.

Referring again to Figure 2-7 on page 44, in addition to its intra-cage control scope, the NC interfaces with a higher-level system-control entity as the system controller (SC). The SC operates in the ML3 domain of the system and is the point of system aggregation for the multiple processor books. The SCS provides system initialization and error reporting and facilitates service. The design goal for the Power systems function split is that every Node Controller (ML2 controller) controls its node as self-contained as possible, such as initializes all HW components within the node in an autonomic way.

The SC (ML3 controller) is then responsible for all system-wide tasks, including NC node management, and the communication with HMC and hypervisor. This design approach yields maximum parallelism of node specific functions and optimizes performance of critical system functions, such as system IPL and system dump, while minimizing external impacts to HMC and hypervisor.

Further to the right in Figure 2-7 on page 44, the downstream fan-out into the sensors and effectors is shown. A serial link, the FRU Support interface (FSI), is used to reach the endpoint controls.

The endpoints are called Common FRU Access Macro (CFAM). CFAMs are integrated in the microprocessors and all I/O ASICs. CFAMs support a variety of control interfaces such as JTAG, UART, I2C, and GPIO.

Also what it shown is a link called the Processor Support Interface (PSI). This interface is new in Power Systems. It is used for high-speed communication between the service subsystem and the host firmware subsystem. Each CEC node has 4 PSI links associated with it, one from each processor chip.

The BPH is a VLAN capable switch, that is part of the BPA. All SCs and NCs and the HMC plug into that switch. The VLAN capability allows a single physical wire to act as separate virtual LAN connections. The SC and BPC will make use of this functionality.

The switch is controlled (programmed) by the BPC firmware.

2.1.5 System controller (SC) card

Two service processor cards are on the CEC midplane. These cards are referred to as system controllers (SC). Figure 2-8 shows a system controller.

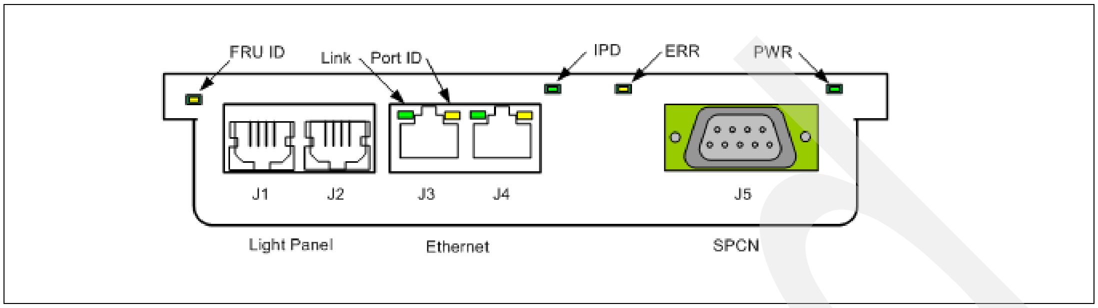


Figure 2-8 System controller card

The SC card provides connectors for:

- Two Ethernet ports (J3, J4)
- Two Lightstrips port - one for the front lightstrip (J1) and one for the rear lightstrip (J2)
- One System Power Control Network (SPCN) port (J5)

SPCN Control network

The System Power Control Network (SPCN) control software and the system controller software run on the embedded system controller service processor (SC).

SPCN is a serial communication network that interconnects the operating system and power components of all IBM Power Systems.It provides the ability to report power failures in connected components directly to the operating system.It plays a vital role in system VPD along with helping map logical to physical relationships.SPCN also provides selective operating system control of power to support concurrent system upgrade and repair.

The SCs implement an SPCN serial link. A 9-pin D-shell connector on each SC implements each half of the SPCN serial loop. A switch on each SC allows the SC in control to access its own 9-pin D-shell and the 9-pin D-shell on the other SC.

Each service processor inside SC provides an SPCN port and is used to control the power of the attached I/O subsystems.

The SPCN ports are RS485 serial interface and uses standard RS485 9-pin female connector (DB9).

Figure 2-9 on page 48 details the SPCN control network.

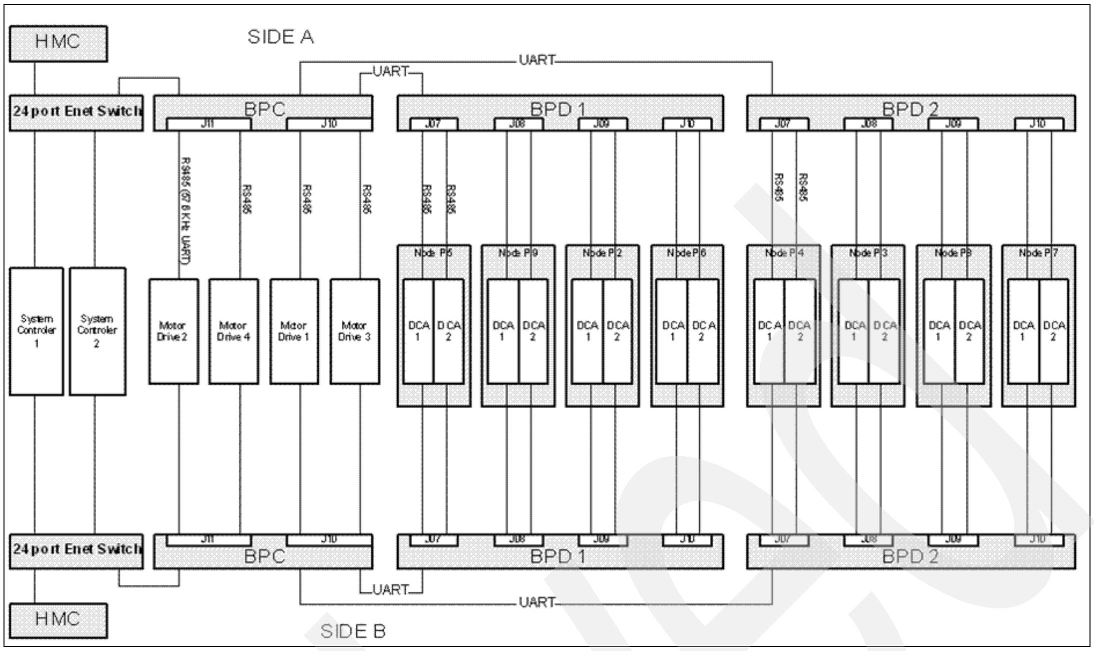


Figure 2-9 SPCN control network

2.1.6 System VPD cards

Two types of Vital Product Data (VPD) cards are available: VPD and smartchip VPD (SVPD). VPD for all field replaceable unit (FRUs) are stored in Serial EPROM (SEEPROM). VPD SEEPROM modules are provided on daughter cards on the midplane (see Figure 2-6 on page 43) and on a VPD daughter card part of processor book assembly, Both are redundant. Both SEEPROMs on the midplane daughter card will be accessible from both SC cards. Both SEEPROMs on the processor book card will be accessible from both Node Controller cards. VPD daughter cards on midplane and processor book planar are not FRUs and are not replaced if one SEEPROM module fails.

SVPD functions are provided on daughter cards on the midplane (see Figure 2-6 on page 43) and on a VPD card part of the processor book assembly; both are redundant. These SVPD cards are available for Capacity Upgrade on Demand (CUoD) functions. The midplane SVPD daughter card also serves as the anchor point for system VPD collection. SVPD function on both the midplane board and the processor book board will be redundant. Both SVPD functions on the midplane board must be accessible from both SC cards. Both SVPD functions on the processor book planar board must be accessible from both NC cards. Note that individual SVPD cards are not implemented for each processor module but just at processor book level. MCM level SVPD is not necessary for the following reasons:

- All four processors in a book are always populated (CoD).
- All processors within a system must run at the same speed (dictated by the slowest module) and that speed can be securely stored in the anchor card or book SVPD modules.
- The MCM serial number is stored in the SEEPROMs on the MCM.

Figure 2-6 on page 43 shows the VPD cards location in midplane.

2.1.7 Oscillator card

Two (redundant) oscillator cards are on the CEC midplane. These oscillator cards are sometimes referred to as *clock cards*. An oscillator card provides clock signals to the entire system. Although the card is actively redundant, only one is active at a time. In the event of a clock failure, the system dynamically switches to the redundant oscillator card. System clocks are initialized based on data in the PU Book VPD. Both oscillators must be initialized so that the standby oscillator can dynamically switch if the primary oscillator fails.

The system oscillators support spread spectrum for reduction of radiated noise. Firmware must ensure that spread spectrum is enabled in the oscillator. A system oscillator card is shown in Figure 2-10.

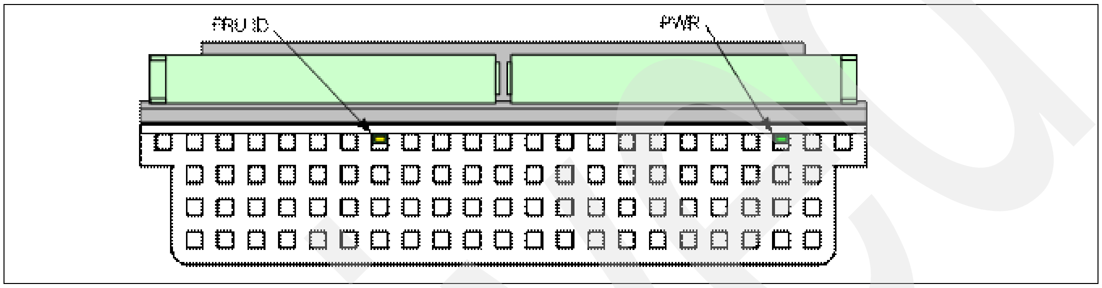


Figure 2-10 Oscillator card

2.1.8 Node controller card

Two embedded node controller (NC) service processor cards are on every processor book. They plug on the processor book planar.

The NC card provides connectors for two Ethernet ports (J01, J02) to BPH.

An NC card is shown in Figure 2-11.

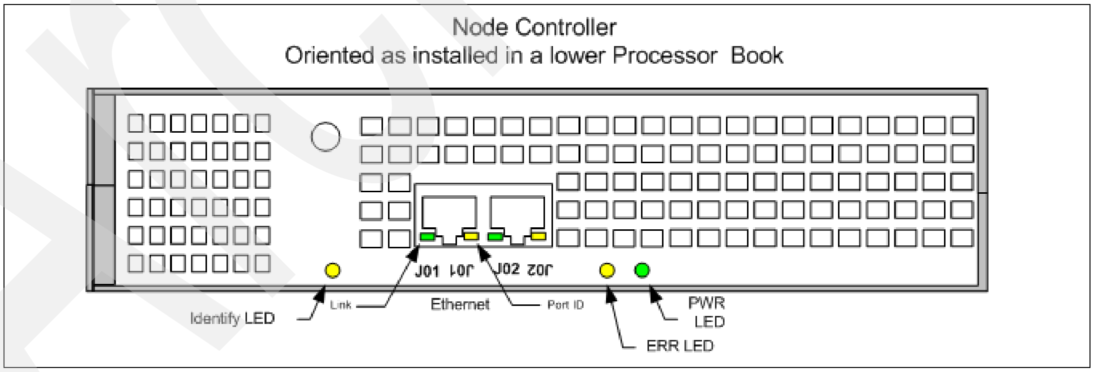


Figure 2-11 Node Controller card

There is a full duplex serial link between each node controller NC and each DCA within a processor book. This link is intended primarily for the relaying of BPC-ip address and MTMS information to the System Power Control Network (SPCN), but can be used for other purposes. The DCA asynchronously forwards this information to the NC without command input from SPCN.

2.1.9 DC converter assembly (DCA)

For the CEC, dc power is generated by redundant, concurrently maintainable dc to dc converter assemblies (DCAs). The DCAs convert main isolated 350VDC to voltage levels appropriate for the processors, memory and CEC contained I/O hub cards. Industry standard dc-dc voltage regulator module (VRM) technology is used.

The DCA does not support multiple core voltage domains per processor. The processor book planar is wired to support a core voltage/nest domain and a cache array voltage domain for each of the four MCMs. A common I/O voltage domain is shared among all CEC logic.

Figure 2-12 shows the DCA assembly on the processor book.

Note: A special tool is required to install and remove the DCA (worm-screw mechanism).

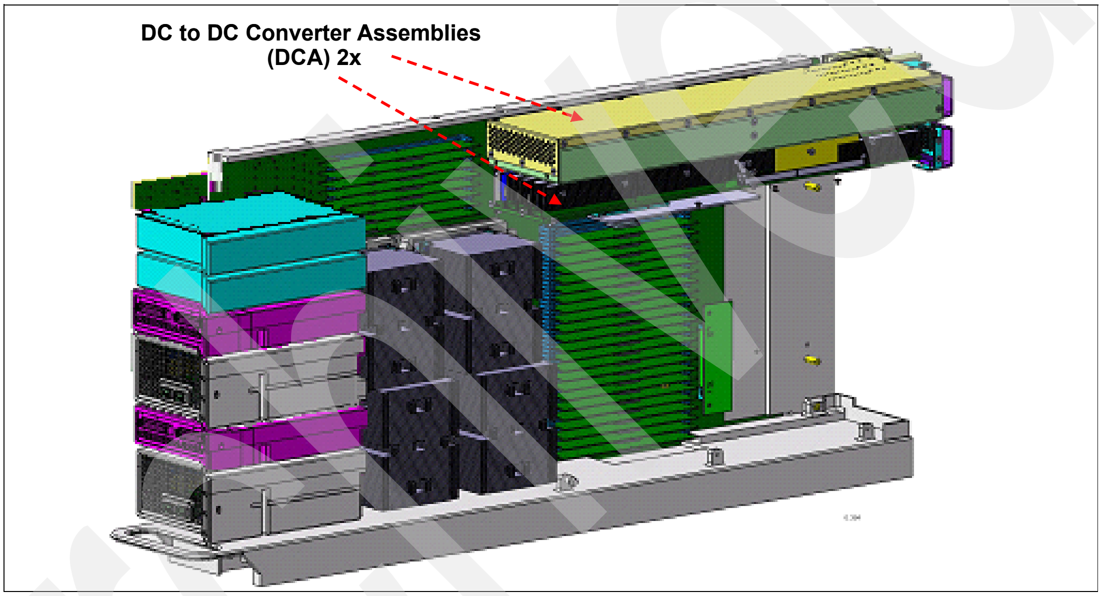


Figure 2-12 DC converter assembly (DCA)

When both DCAs are operational, adequate power is available for all operating conditions. For some technical workloads, the processors can draw more current than a single DCA can supply. In the event of a DCA failure, the remaining operational DCA can supply the needed current for only a brief period before overheating. To prevent overheating when load is excessive, the remaining DCA (through processor services) can reduce the processor load by *throttling* processors or reducing processor clock frequency. Consider the following notes:

- Reducing processor clock frequency must be done by the system controller and can take a long time to accomplish. Throttling can be accomplished much more quickly.
- The DCA uses an I2C connection to each processor to accomplish throttling within 10 ms. Throttling causes the processors to slow down instruction dispatch to reduce power draw. Throttling can affect performance significantly (by approximately 90% or more).
- To regain performance after throttling, the system controller slows down the system clocks and reduces core voltage, and then unthrottles the processor or processors. This is effective because slowing the processor clocks and reducing voltage is in turn much more effective at reducing load than throttling. After a failing DCA is replaced, the system clocks are returned to normal if needed.

2.2 System buses

The POWER6 processor interfaces can be divided into three categories:

- SMP interconnect: These interfaces connect the POWER6 processors to each other. These links form a coherent *fabric* for system requests in addition to a data routing network. The links are multiplexed—the same wires are time-sliced among address, data, and control information.
- Local interconnect: Local interfaces communicate the memory structures associated with a specific POWER6 technology-based chip.
- External interconnect: Interfaces provide for communication with I/O devices outside the central system.

This section discusses the SMP and external interconnects.

2.2.1 System interconnects

The Power 595 uses point-to-point SMP fabric interfaces between processor node books. Each processor book holds a processor node consisting of four dual-core processors designated S, T, U and V.

The bus topology is no longer ring-based as in POWER5, but rather a multi-tier, fully-connected topology in order to reduce latency, increase redundancy, and improve concurrent maintenance. Reliability is improved with error correcting code (ECC) on the external I/Os, and ECC and parity on the internal chip wires.

Books are interconnected by a point-to-point connection topology, allowing every book to communicate with every other book. Data transfer never has to go through another books read cache to address the requested data or control information. Inter-book communication takes place at the Level 2 (L2) cache.

The POWER6 fabric bus controller (FBC) is the framework for creating a cache-coherent multiprocessor system. The FBC provides all of the interfaces, buffering, and sequencing of address and data operations within the storage subsystem. The FBC is integrated on the POWER6 processor. The POWER6 processor has five fabric ports that can be used to connect to other POWER6 processors.

- Three for intranode bus interconnections. They are designated as X, Y, and Z and are used to fully connect the POWER6 processor on a node.
- Two for internode bus connections. They are designated as A and B ports and are used to fully-connect nodes in multi-node systems.

Physically, the fabric bus is an 8-, 4-, or 2-byte wide, split-transaction, multiplexed address. For Power 595 the bus is 8 bytes and operates at half the processor core frequency

From a fabric perspective, a node (processor book node) is one to four processors fully connected with XYZ busses. AB busses are used to connect various fabric nodes together. The one to four processor on a node work together to broadcast address requests to other nodes in the system. Each node can have up to 8 AB links (two for processor, four processor per node). Figure 2-13 on page 52 illustrates the internode bus interconnections.

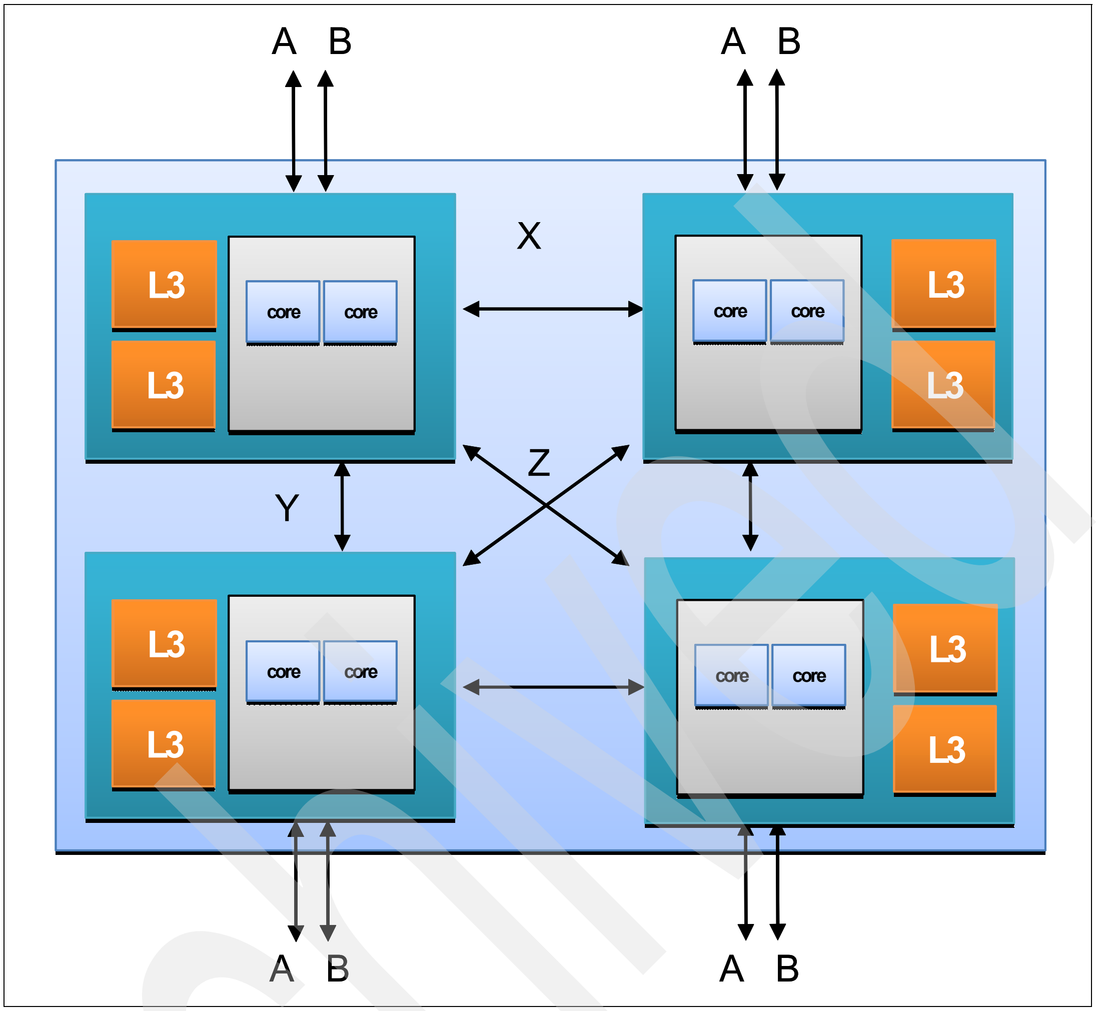


Figure 2-13 FBC bus connections

The topologies that are illustrated in Figure 2-14 on page 54 are described in Table 2-3.

Table 2-3 Topologies of POWER5 and POWER6

System	Description
(a) POWER5	The topology of a POWER5 processor-based system consists of a first-level nodal structure containing up to four POWER5 processors. Coherence links are fully connected so that each chip is directly connected to all of the other chips in the node. Data links form clockwise and counterclockwise rings that connect all of the chips in the node. All of the processor chips within a node are designed to be packaged in the same multichip module.
(a) POWER6	The POWER6 processor first-level nodal structure is composed of up to four POWER6 processors. Relying on the traffic reduction afforded by the innovations in the coherence protocol to reduce packaging overhead, coherence and data traffic share the same physical links by using a time-division-multiplexing (TDM) approach. With this approach, the system can be configured either with 67% of the link bandwidth allocated for data and 33% for coherence or with 50% for data and 50% for coherence. Within a node, the shared links are fully connected such that each chip is directly connected to all of the other chips in the node.

System	Description
(b) POWER5	A POWER5 processor-based system can interconnect up to eight nodes with a parallel ring topology. With this approach, both coherence and data links are organized such that each chip within a node is connected to a corresponding chip in every node by a unidirectional ring. For a system with four processor s per node, four parallel rings pass through every node. The POWER5 chip also provides additional data links between nodes in order to reduce the latency and increase the bandwidth for moving data within a system. While the ring-based topology is ideal for facilitating a nonblocking-broadcast coherence-transport mechanism, it involves every node in the operation of all the other nodes. This makes it more complicated to provide isolation capabilities, which are ideal for dynamic maintenance activities and virtualization.
(b) POWER6	For POWER6 processor-based systems, the topology was changed to address dynamic maintenance and virtualization activities. Instead of using parallel rings, POWER6 process-based systems can connect up to eight nodes with a fully connected topology, in which each node is directly connected to every other node. This provides optimized isolation because any two nodes can interact without involving any other nodes. Also, system latencies do not increase as a system grows from two to eight nodes, yet aggregate system bandwidth increases faster than system size. Of the five 8-byte off-chip SMP interfaces on the POWER6 chip (which operate at half the processor frequency), the remaining two are dedicated (A, B - Intranode) to interconnecting the second-level system structure. Therefore, with a four-processor node, eight such links are available for direct node-to-node connections. Seven of the eight are used to connect a given node to the seven other nodes in an eight-node 64-core system. The five off-chip SMP interfaces on the POWER6 chip protect both coherence and data with SECDED ECCs.

With both the POWER5 and the POWER6 processor approaches, large systems are constructed by aggregating multiple nodes.

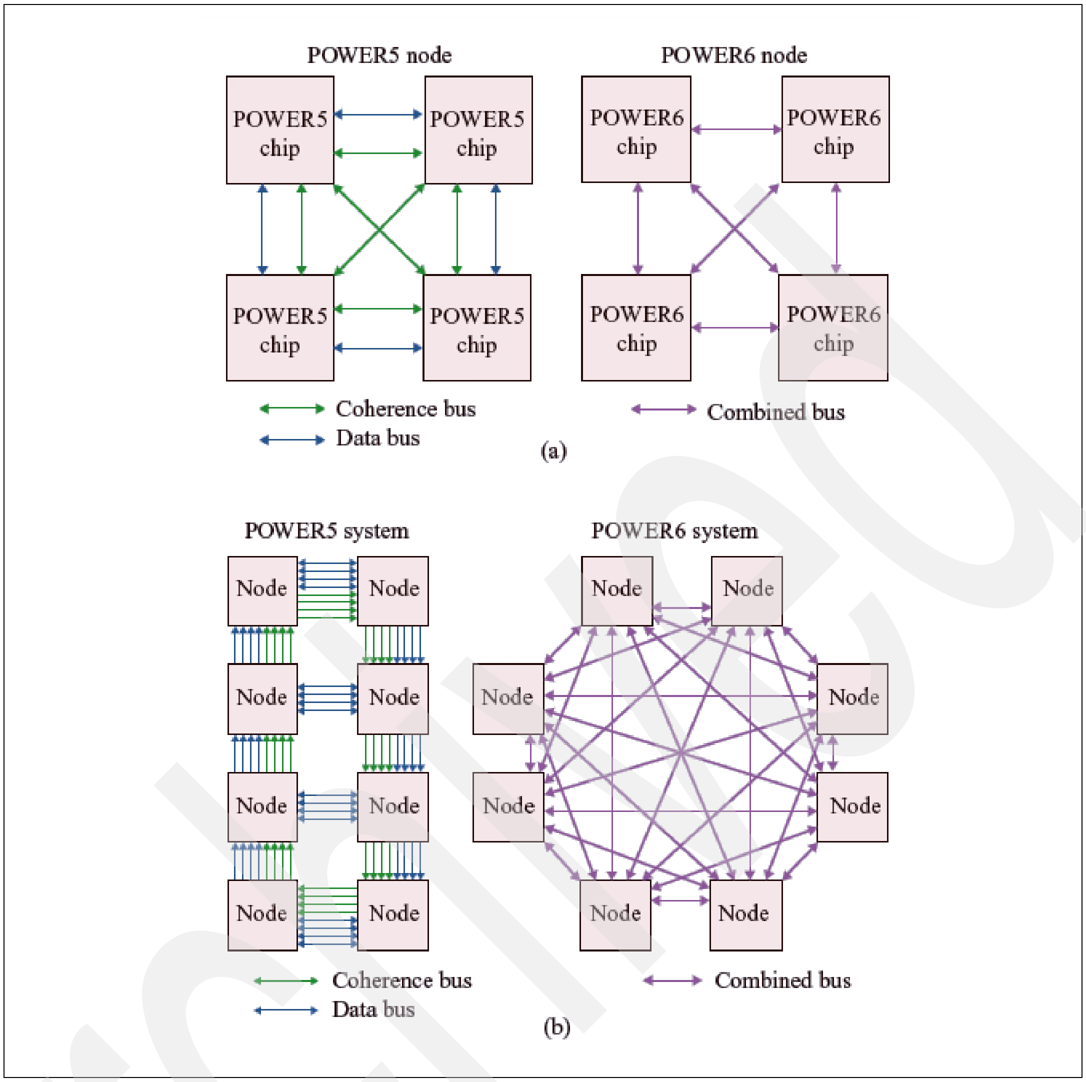


Figure 2-14 POWER5 and POWER6 processor (a) first-level nodal topology and (b) second-level system topology.

Figure 2-15 on page 55 illustrates the potential for a large, robust, 64-core system that uses 8-byte SMP interconnect links, both L3 data ports to maximize L3 bandwidth, and all eight memory channels per chip.

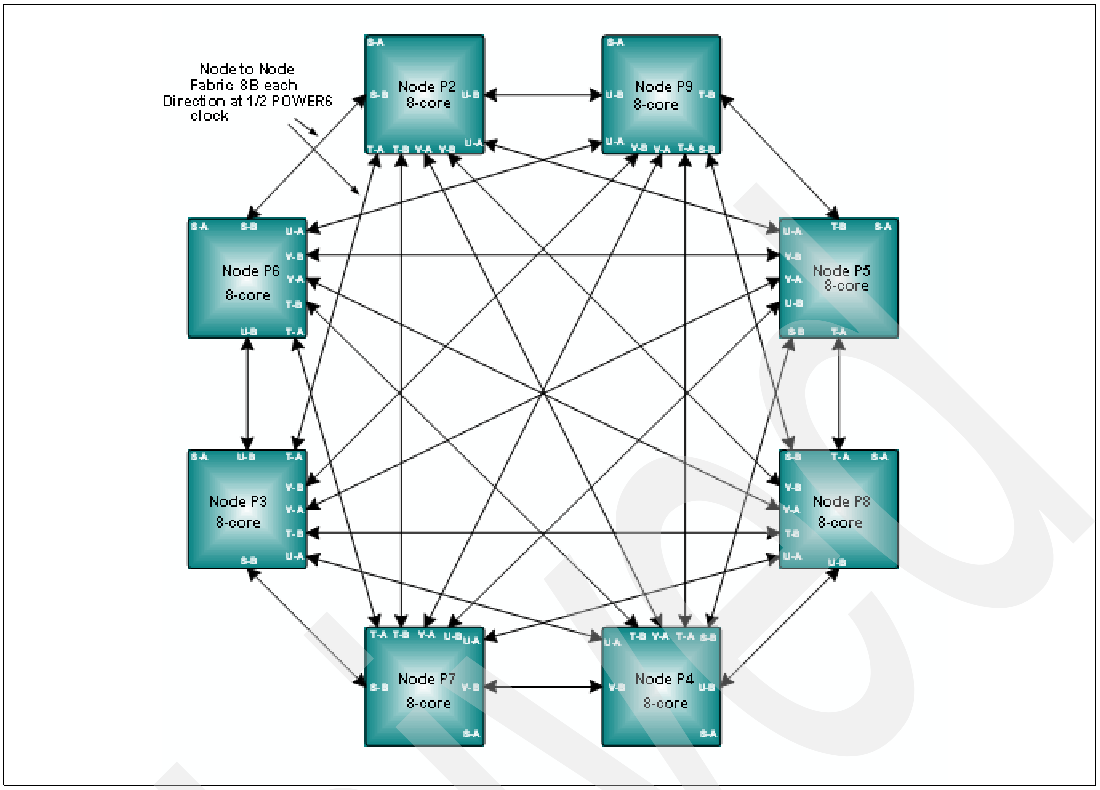


Figure 2-15 Power 595 64 core

2.2.2 I/O subsystem

The Power 595 utilizes remote I/O drawers for directly attached PCI or PCI-X adapters and disk capabilities.The 595 supports I/O DASD and media drawers through Remote I/O (RIO), High Speed Loop (HSL), and 12x Host Channel Adapters (HCA) located in the front of the processor books. These are collectively referred to as *GX adapters*.

> **Note:** RIO and HSL describe the same I/O loop technology. RIO is terminology from System p and HSL is terminology from System i.

Two types of GX adapter cards are supported in the 595 servers:

- Remote I/O-2 (RIO-2) dual port Loop Adapter (#1814)
- GX dual port 12x HCA adapter (#1816)

Drawer connections are always made in loops to help protect against a single point-of-failure resulting from an open, missing, or disconnected cable. Systems with non-looped configurations could experience degraded performance and serviceability.

RIO-2 loop connections operate bidirectional at 1 GBps (2 GBps aggregate). RIO-2 loops connect to the system CEC using RIO-2 loop attachment adapters (#1814). Each adapter has two ports and can support one RIO-2 loop. A maximum of four adapters can be installed in each 8-core processor book.

The 12x loop connections operate bidirectional at 3 GBps (6 GBps aggregate). 12x loops connect to the system CEC using 12x loop attachment adapters (#1816). Each adapter has two ports and can support one 12x loop.

A maximum of four adapters can be installed in each 8-core processor book: two wide and two narrow. Beginning with the adapter slot closest to Node Controller 0, the slots alternate narrow-wide-narrow-wide. GX slots T and U are narrow; S and V are wide.

Figure 2-16 details the GX adapter layout for a two-processor book.

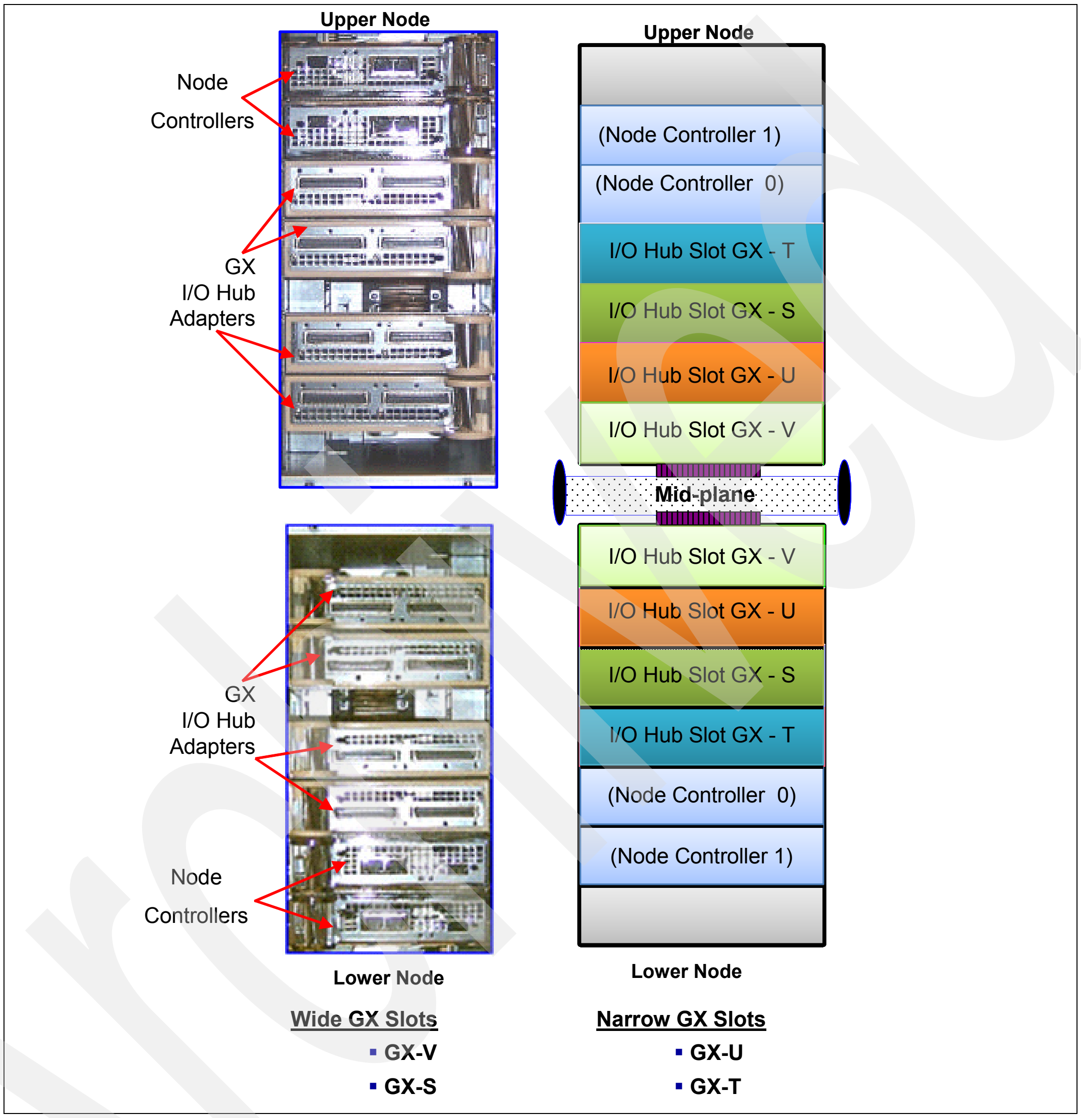


Figure 2-16 GX adapter placement

For I/O hub plugging rules, see section 2.8, “Internal I/O subsystem” on page 82.

Figure 2-17 on page 57 shows the cabling order when connecting I/O drawers to the I/O hubs. The numbers on the left show the sequence in which the I/O hubs are selected for cabling to the drawers. The designation on the left indicates the drawer location that the cables will run to. “A’ indicates system rack. P-A indicates a powered expansion rack (not shown), and P-Z indicates a nonpowered expansion rack attached to a powered expansion rack. The numbers at the right indicate the rack location for the bottom edge of the drawer. For example, A-01 is the drawer in the system rack, located at U1 (first I/O drawer) and A-09 is the drawer in the system rack, located at U9 (third I/O drawer).

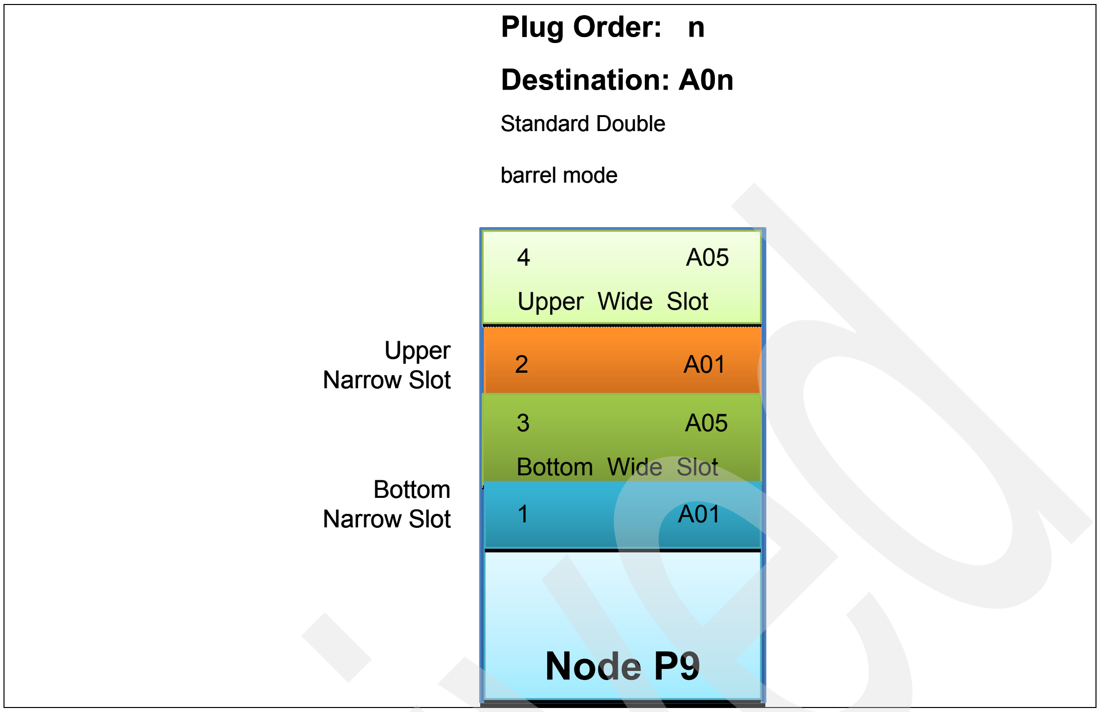


Figure 2-17 Plugging order

2.3 Bulk power assembly

The Power 595 system employs a universal front-end power system. It can accept nominal ac inputs from 200 V to 480 V at 50 or 60 Hz and converts this to a main isolated 350 V dc nominal bulk power. The Bulk Power Assembly (BPA) holds the bulk power components.

The primary system rack and powered Expansion Rack always incorporate two bulk power assemblies for redundancy. These provide 350 V dc power for devices located in those racks and associated nonpowered Expansion Racks. These bulk power assemblies are mounted in front and rear positions and occupy the top 8U of the rack. To help provide optimum system availability, the bulk power assemblies should be powered from separate power sources with separate line cords.

The Power 595 has both primary and redundant Bulk Power Assemblies (BPAs). The BPAs provide the prime power conversion and dc distribution for devices located in the POWER6 595 CEC rack. They are comprised of the following individual components, all of which support concurrent maintenance and require no special tools:

Table 2-4 BPA components

Component	Definition
Bulk power controller (BPC)	Is the BPA's main power and CEC controller.
Bulk power distributor (BPD)	Distributes 350 V dc to FRUs in the system frame, including the Air Moving Devices and Distributed Converter Assemblies. A BPA has either one or two BPDs.
Bulk power enclosure (BPE)	Is the metal enclosure containing the BPA components.

Component	Definition
Bulk power fan (BPF)	Cools the BPA components.
Bulk power hub (BPH)	Is a 24 port 10/100 Ethernet switch.
Bulk power regulator (BPR)	Is the main front-end power supply. A BPA has up to four BPRs, each capable of supplying 8 KW of 350 V dc power.

These components are shown in Figure 2-18.

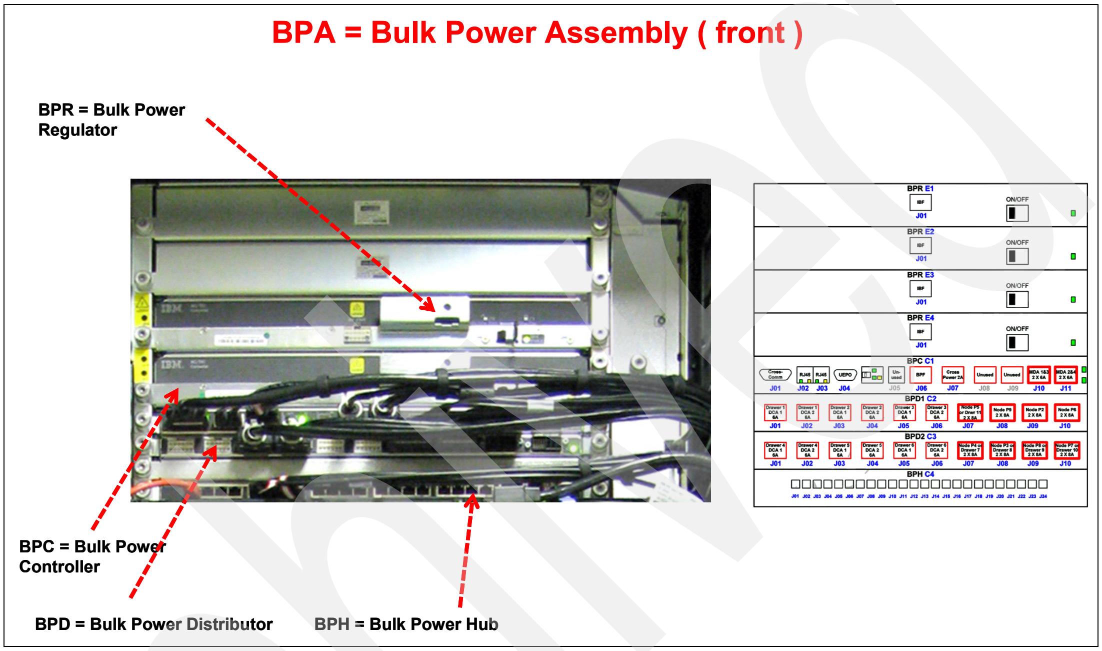

BPA = Bulk Power Assembly (front)
BPR = Bulk Power Regulator
BPR E1
BPR E2
BPR E3
BPR E4
ON/OFF
BPC C1
BPD1 C2
BPD2 C3
BPH C4
BPC = Bulk Power Controller
BPD = Bulk Power Distributor
BPH = Bulk Power Hub

Figure 2-18 Bulk power assembly

The power subsystem in the primary system rack is capable of supporting Power 595 servers with one to eight processor books installed, a media drawer, and up to three I/O drawers. The nonpowered expansion rack can only be attached to powered expansion racks. Attachment of nonpowered expansion racks to the system rack is not supported. The number of BPR and BPD assemblies can vary, depending on the number of processor books, I/O drawers, and battery backup features installed along with the final rack configuration.

2.3.1 Bulk power hub (BPH)

A 24-port 10/100 Ethernet switch serves as the 595 BPH. A BPH is contained in each of the redundant bulk power assemblies located in the front and rear at the top the CEC rack. The BPH provides the network connections for the system control structure (SCS), which in turn provide system initialization and error reporting, and facilitate service operations. The system controllers, the processor book node controllers and BPC use the BPH to communicate to the Hardware Management Console.

Bulk power hubs are shown in Figure 2-19 on page 59.

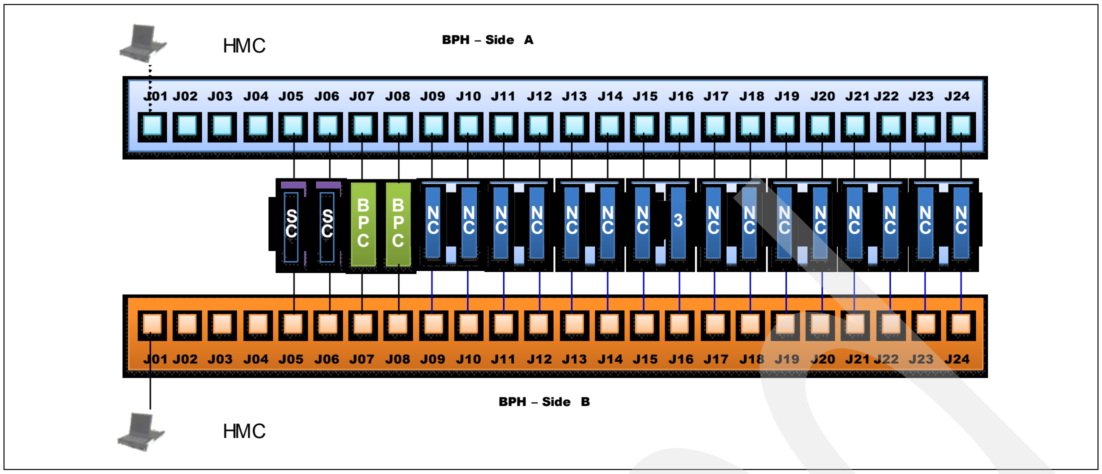

HMC
BPH – Side A
J01 J02 J03 J04 J05 J06 J07 J08 J09 J10 J11 J12 J13 J14 J15 J16 J17 J18 J19 J20 J21 J22 J23 J24
SC
SC
BPC
BPC
NC
3
J01 J02 J03 J04 J05 J06 J07 J08 J09 J10 J11 J12 J13 J14 J15 J16 J17 J18 J19 J20 J21 J22 J23 J24
BPH – Side B
HMC

Figure 2-19 Bulk power hubs (BPH)

Table 2-5 list the BPH location codes.

Table 2-5 Bulk power hub (BPH) location codes

Location code	Component	Location code	Component
U*n*-P*x*-C4	BPH (front or rear)	U*n*-P*x*-C4-J13	Processor book 6 (node P8) node controller 0
U*n*-P*x*-C4-J01	Hardware Management Console	U*n*-P*x*-C4-J14	Processor book 6 (node P8) node controller 1
U*n*-P*x*-C4-J02	Service mobile computer	U*n*-P*x*-C4-J15	Processor book 5 (node P7) node controller 0
U*n*-P*x*-C4-J03	Open	U*n*-P*x*-C4-J16	Processor book 5 (node P7) node controller 1
U*n*-P*x*-C4-J04	Corresponding BPH in powered I/O rack	U*n*-P*x*-C4-J17	Processor book 4 (node P2) node controller 0
U*n*-P*x*-C4-J05	System controller 0 (in CEC midplane)	U*n*-P*x*-C4-J18	Processor book 4 (node P2) node controller 1
U*n*-P*x*-C4-J06	System controller 1 (in CEC midplane)	U*n*-P*x*-C4-J19	Processor book 3 (node P6) node controller 0
U*n*-P*x*-C4-J07	Front BPC	U*n*-P*x*-C4-J20	Processor book 3 (node P6) node controller 1
U*n*-P*x*-C4-J08	Rear BPC	U*n*-P*x*-C4-J21	Processor book 2 (node P5) node controller 0
U*n*-P*x*-C4-J09	Processor book 8 (node P4) node controller 0	U*n*-P*x*-C4-J22	Processor book 2 (node P5) node controller 1
U*n*-P*x*-C4-J10	Processor book 8 (node P4) node controller 1	U*n*-P*x*-C4-J23	Processor book 1 (node P9) node controller 0
U*n*-P*x*-C4-J11	Processor book 7 (node P3) node controller 0	U*n*-P*x*-C4-J24	Processor book 1 (node P9) node controller 1
U*n*-P*x*-C4-J12	Processor book 7 (node P3) node controller 1	—	—

2.3.2 Bulk power controller (BPC)

One BPC, shown in Figure 2-20, is located in each BPA. The BPC provides the base power connections for the internal power cables. Eight power connectors are provided for attaching system components. In addition, the BPC contains a service processor card that provides service processor functions within the power subsystem.

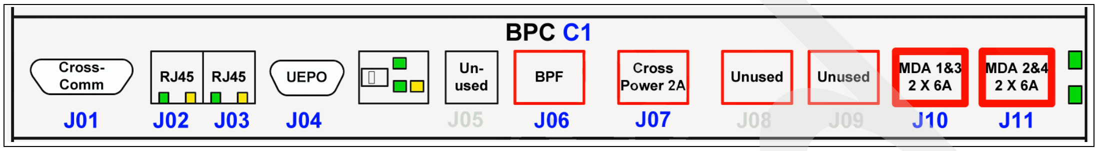


Figure 2-20 Bulk power controller (BPC)

Table 2-6 lists the BPC component location codes.

Table 2-6 Bulk power controller (BPC) component location codes

Location code	Component	Location code	Component
U*n*-P*x*-C1	BPC (front or rear)	U*n*-P*x*-C1-J06	BPF
U*n*-P*x*-C1-J01	BPC Cross Communication	U*n*-P*x*-C1-J07	BPC Cross Power
U*n*-P*x*-C1-J02	Ethernet to BPH	U*n*-P*x*-C1-J08	Not used
U*n*-P*x*-C1-J03	Ethernet to BPH	U*n*-P*x*-C1-J09	Not used
U*n*-P*x*-C1-J04	UEPO Panel	U*n*-P*x*-C1-J10	MDA 1 and MDA 3 (one Y cable powers two MDAs)
U*n*-P*x*-C1-J05	Not used	U*n*-P*x*-C1-J11	MDA 2 and MDA 4 (one Y cable powers two MDAs)

2.3.3 Bulk power distribution (BPD)

Redundant BPD assemblies provide additional power connections to support the system cooling fans, dc power converters contained in the CEC, and the I/O drawers. Each power distribution assembly provides ten power connections. Two additional BPD assemblies are provided with each Powered Expansion Rack.

Figure 2-21 details the BPD assembly.

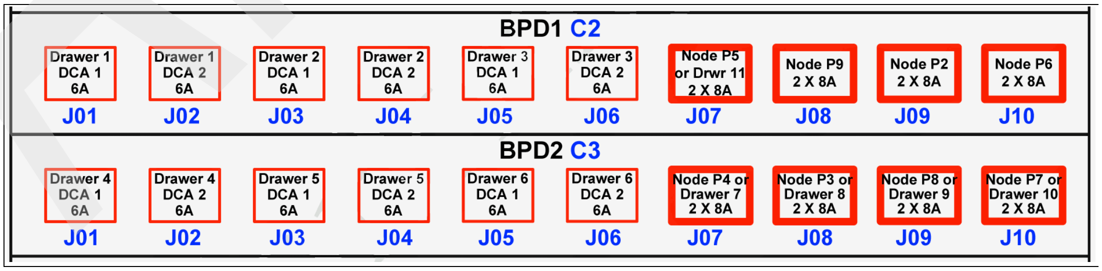


Figure 2-21 Bulk power distribution (BPD) assembly

Table 2-7 on page 61 lists the BPD component location codes.

Table 2-7 Bulk power distribution (BPD) assembly component location codes

Location code	Component	Location code	Component
U*n*-P*x*-C2	BPD 1 (front or rear)	U*n*-P*x*-C3	BPD 2
U*n*-P*x*-C2-J01	I/O Drawer 1, DCA 2	U*n*-P*x*-C3-J01	I/O Drawer 4, DCA 2
U*n*-P*x*-C2-J02	I/O Drawer 1, DCA 1	U*n*-P*x*-C3-J02	I/O Drawer 4, DCA 1
U*n*-P*x*-C2-J03	I/O Drawer 2, DCA 2	U*n*-P*x*-C3-J03	I/O Drawer 5, DCA 2
U*n*-P*x*-C2-J04	I/O Drawer 2, DCA 1	U*n*-P*x*-C3-J04	I/O Drawer 5, DCA 1
U*n*-P*x*-C2-J05	I/O Drawer 3, DCA 2	U*n*-P*x*-C3-J05	I/O Drawer 6, DCA 2
U*n*-P*x*-C2-J06	I/O Drawer 3, DCA 1	U*n*-P*x*-C3-J06	I/O Drawer 6, DCA 1
U*n*-P*x*-C2-J07	Processor book 2 (node P5)	U*n*-P*x*-C3-J07	Processor book 8 (node P4) or I/O Drawer 7
U*n*-P*x*-C2-J08	Processor book 1 (node P9)	U*n*-P*x*-C3-J08	Processor book 7 (node P3) or I/O Drawer 8
U*n*-P*x*-C2-J09	Processor book 4 (node P2)	U*n*-P*x*-C3-J09	Processor book 6 (node P8) or I/O Drawer 9
U*n*-P*x*-C2-J10	Processor book 3 (node P6)	U*n*-P*x*-C3-J10	Processor book 5 (node P7) or I/O Drawer 10

2.3.4 Bulk power regulators (BPR)

The redundant BPRs interface to the bulk power assemblies to help ensure proper power is supplied to the system components. Figure 2-22 on page 62 shows four BPR assemblies. The BPRs are always installed in pairs in the front and rear bulk power assemblies to provide redundancy. One to four BPRs are installed in each BPA. A BPR is capable of supplying 8 KW of 350 VDC power. The number of bulk power regulators required is configuration dependent, based on the number of processor MCMs and I/O drawers installed. Figure 2-22 on page 62 details the BPR assembly.

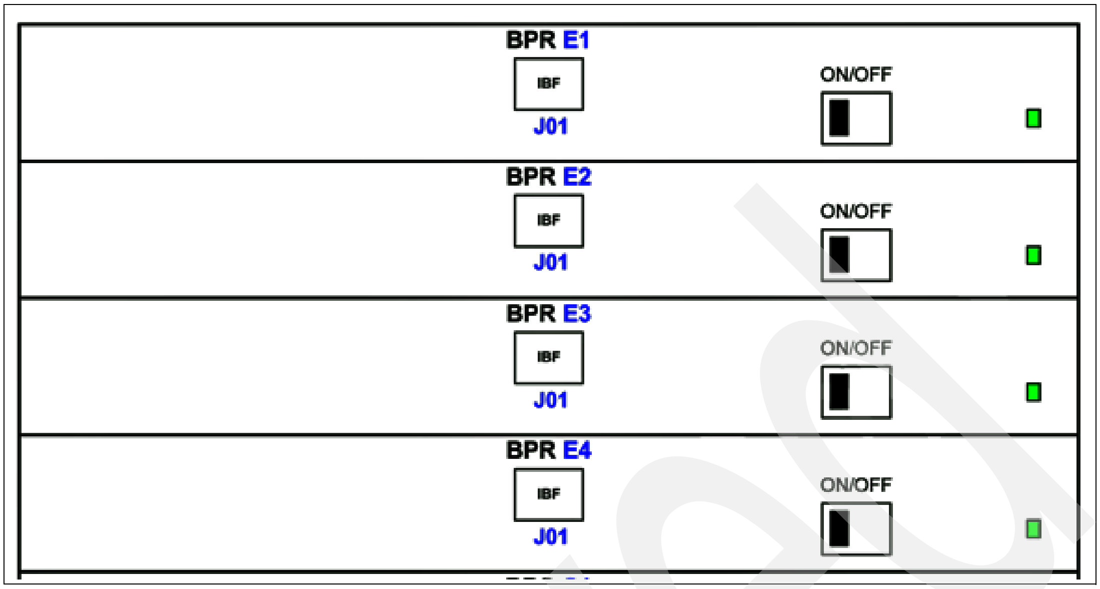


Figure 2-22 Bulk power regulator (BPR) assemblies

Table 2-8 lists the BPR component location codes.

Table 2-8 Bulk power regulator (BPR) component location codes

Location code	Component	Location code	Component
U*n*-P*x*-E1	BPR 4 (front or rear)	U*n*-P*x*-E3	BPR 2 (front or rear)
U*n*-P*x*-E1-J01	Not used	U*n*-P*x*-E3-J01	Integrated Battery feature connector
U*n*-P*x*-E2	BPR 3 (front or rear)	U*n*-P*x*-E4	BPR 1 (front or rear)
U*n*-P*x*-E2-J01	Not used	U*n*-P*x*-E4-J01	Integrated Battery feature connector

2.3.5 Bulk power fan (BPF)

Each bulk power assembly has a BPF for cooling the components of the bulk power enclosure. The bulk power fan is powered via the universal power input cable (UPIC) connected to connector J06 on the BPC. The BPF is shown in Figure 2-23 on page 63.

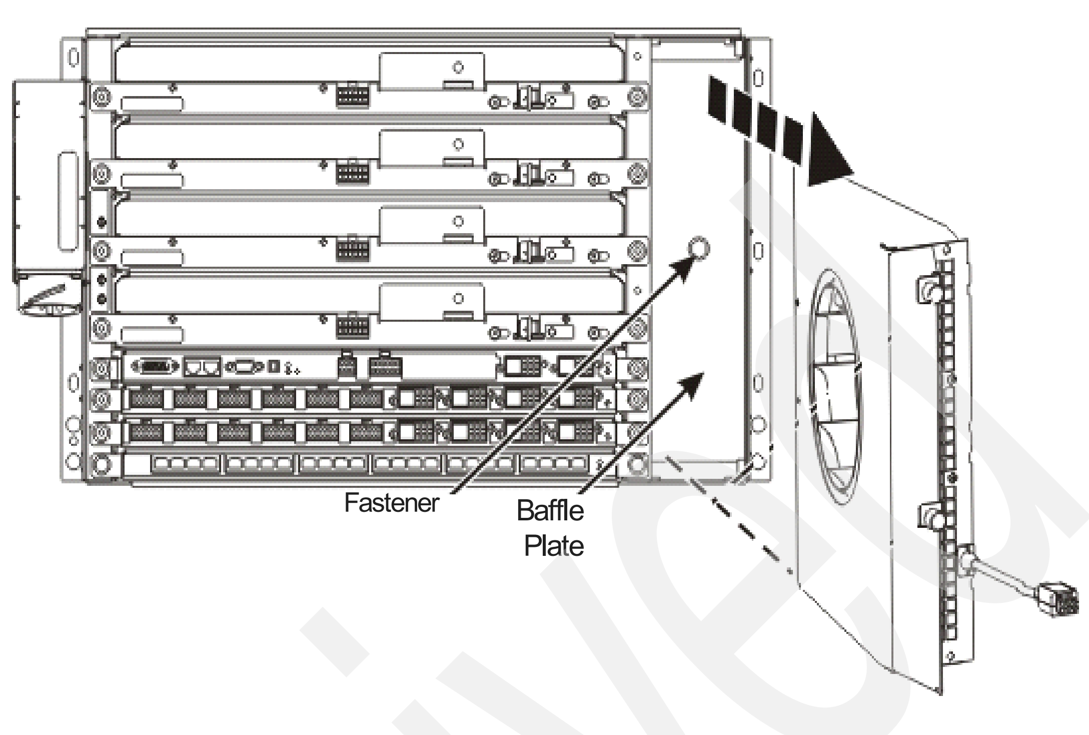


Figure 2-23 Bulk power fan (BPF)

2.3.6 Integrated battery feature (IBF)

An optional integrated battery feature (IBF) is available for the Power 595 server. The battery backup units are designed to protect against power line disturbances and provide sufficient, redundant power to allow an orderly system shutdown in the event of a power failure. The battery backup units attach to the system BPRs.

Each IBF is 2U high and IBF units will be located in each configured rack: CEC, Powered Expansion Rack, and nonpowered bolt-on rack. When ordered, the IBFs will displace the media drawer or an I/O drawer. In the CEC rack, two positions, U9 and U11 (located below the processor books) will each be occupied by redundant battery backup units. When positions U9 and U11 are occupied by battery backup units they replace one I/O drawer position. When ordered, each unit provides both primary and redundant backup power and occupy 2U of rack space. Each unit occupies both front and rear positions in the rack. The front rack positions provide primary battery backup of the power subsystem; the rear rack positions provide redundant battery backup. The media drawer is not available when the battery backup feature is ordered. In the Powered Expansion Rack (#6494), two battery backup units are located in locations 9 and 11, displacing one I/O drawer. As in the CEC rack, these battery backup units provide both primary and redundant battery backup of the power subsystem.

2.3.7 POWER6 EnergyScale

With increasing processor speed and density, denser system packaging, and other technology advances, system power and heat have become important design considerations. IBM has developed the EnergyScale architecture, a system-level power management implementation for POWER6 processor-based machines. The EnergyScale architecture uses the basic power control facilities of the POWER6 chip, together with additional board-level hardware, firmware, and systems software, to provide a complete power and thermal management solution. IBM has a comprehensive strategy for data center energy management:

- Reduce power at the system level where *work per watt* is the important metric, not *watts per core*. POWER6-based systems provide more watt per core within the same power envelope.
- Manage power at the data center level through IBM Director Active Energy Manager.
- Automate energy management policies such as:
  - Energy monitoring and management through Active Energy Manager and EnergyScale
  - Thermal and power measurement
  - Power capping
  - Dynamic power management and savings
  - Performance-aware power management

Often, significant runtime variability occurs in the power consumption and temperature because of natural fluctuations in system utilization and type of workloads being run.

Power management designs often use a worst-case conservation approach because servers have fixed power and cooling budgets (for example, a 100 W processor socket in a rack-mounted system). With this approach, the frequency or throughput of the chips must be fixed to a point well below their capability, sacrificing sizable amounts of performance, even when a non-peak workload is running or the thermal environment is favorable. Chips, in turn, are forced to operate at significantly below their runtime capabilities because of a cascade of effects. The net results include:

- Power supplies are significantly over-provisioned.
- Data centers are provisioned for power that cannot be used.
- Higher costs with minimal benefit occur in most environments.

Building adaptability into the server is the key to avoiding conservative design points in order to accommodate variability and to take further advantage of flexibility in power and performance requirements. A design in which operational parameters are dictated by runtime component, workload, environmental conditions, and by your current power versus performance requirement is less conservative and more readily adjusted to your requirements at any given time.

The IBM POWER6 processor is designed exactly with this goal in mind (high degree of adaptability), enabling feedback-driven control of power and associated performance for robust adaptability to a wide range of conditions and requirements. Explicit focus was placed on developing each of the key elements for such an infrastructure: sensors, actuators, and communications for control.

As a result, POWER6 microprocessor-based systems provide an array of capabilities for:

- Monitoring power consumption and environmental and workload characteristics
- Controlling a variety of mechanisms in order to realize the desired power and performance trade-offs (such as the highest power reduction for a given performance loss)
- Enabling higher performance and greater energy efficiency by providing more options to the system designer to dynamically tune it to the exact requirements of the server

EnergyScale is an infrastructure that enables:

- Real-time measurements feedback to address variability and unpredictability of parameters such as power, temperature, activity, and performance
- Mechanisms for regulating system activity, component operating levels, and environmental controls such as processor, memory system, fan control and disk power, and a dedicated control structure with mechanisms for interactions with OS, hypervisor, and applications.
- Interaction and integration that provides:
  - Policy-guided power management to support user-directed policies and operation modes
  - Critical event and comprehensive usage information
  - Support integration into larger scope system management frameworks

Design principles

The EnergyScale architecture is based on design principles that are used not only in POWER6 processor-based servers but also in the IBM BladeCenter® and IBM System x™ product lines. These principles are the result of fundamental research on system-level power management performed by the IBM Research Division, primarily in the Austin Research Laboratory. The major design principles are as follows:

- Implementation is primarily an out-of-band power management scheme. EnergyScale utilizes one or more dedicated adapters (thermal power management devices (TPMD)) or one or more service processors to execute the management logic. EnergyScale communicates primarily with both in-system (service processor) and off-system (for example, Active Energy Manager) entities.
- Implementation is measurement-based, that is it continuously takes measurements of voltage and current to calculate the amount of power drawn. It uses temperature sensors to measure heat, and uses performance counters to determine the characteristics of workloads. EnergyScale directly measures voltage, current, and temperature to determine the characteristics of workloads (for example sensors and critical path monitors).
- Implementation uses real-time measurement and control. When running out-of-band, the EnergyScale implementation relies on real-time measurement and control to ensure that the system meets the specified power and temperature goals. Timings are honored down to the single-millisecond range.
- System-level Power Management™ is used. EnergyScale considers a holistic view of power consumption. Most other solutions focus largely or exclusively on the system processor.
- Multiple methods are available to control power and thermal characteristics of a system, it allows sensing and allows for acting on system parameters to achieve control. The EnergyScale implementation uses multiple actuators to alter the power consumption and heat dissipation of the processors and the memory in the system.
- The architecture contains features that ensure safe, continued operation of the system during adverse power or thermal conditions, and in certain cases in which the EnergyScale implementation itself fails.
- The user has indirect control over Power Management behavior by using configurable policies, similar to existing offerings from other product offerings (Active Energy Manager).

Figure 2-24 on page 66 show the POWER6 power management architecture.

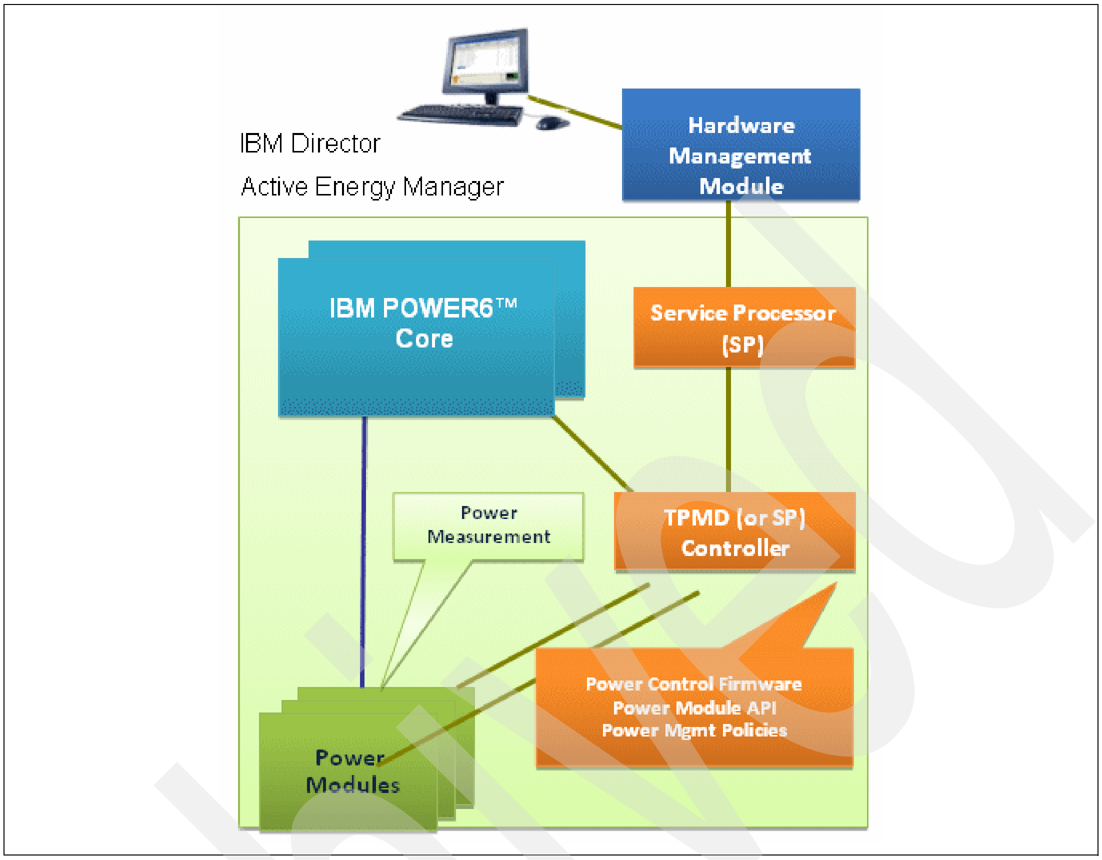


Figure 2-24 Power management architecture

EnergyScale functions

The IBM EnergyScale functions, and hardware and software requirements are described in the following list:

Power trending EnergyScale provides continuous power usage data collection (monitoring). This enables the administrators with the information to predict power consumption across their infrastructure and to react to business and processing needs. For example, an administrator could adjust server consumption to reduce electrical costs. To collect power data for the 520, having additional hardware is unnecessary because EnergyScale collects the information internally.

Power saver mode This mode reduces the voltage and frequency by a fixed percentage. This percentage is predetermined to be within a safe operating limit and is not user-configurable. Under current implementation, this is a 14% frequency drop. When CPU utilization is low, power saver mode has no impact on performance. Power saver mode can reduce the processor usage up to a 30%. Power saver mode is not supported during boot or reboot although it is a persistent condition that will be sustained after the boot when the system starts executing instructions. Power saver is only supported with 4.0 GHz processors and faster.

Power capping This enforces a user-specified limit on power usage. Power capping is not a power saving mechanism. It enforces power caps by actually throttling the one or more processors in the system, degrading performance significantly. The idea of a power cap is to set something that should never be reached but frees up margined power in the data center. The *margined power* is the amount of extra power that is allocated to a server during its installation in a data center. It is based on the server environmental specifications that usually are never reached. Server specifications are always based on maximum configurations and worst case scenarios.

Processor core nap The IBM POWER6 processor uses a low-power mode called *nap* that stops processor execution when there is no work to do on that processor core (both threads are idle). Nap mode allows the hardware to clock-off most of the circuits inside the processor core. Reducing active power consumption by turning off the clocks allows the temperature to fall, which further reduces leakage (static) power of the circuits causing a cumulative effect. Unlicensed cores are kept in core nap until they are licensed and return to core nap whenever they are unlicensed again.

EnergyScale for I/O IBM POWER6 processor-based systems automatically power off pluggable, PCI adapter slots that are empty or not being used, saving approximately 14 watts per slot. System firmware automatically scans all pluggable PCI slots at regular intervals looking for slots that meet the criteria of not being in use, and then powers them off. This support is available for all POWER6 processor-based servers, and the expansion units that they support. Note that it applies to hot pluggable PCI slots only.

Oversubscription protection

In systems with dual or redundant power supplies, additional performance can be obtained by using the combined supply capabilities of all supplies. However, if one of the supplies fails, the power management immediately switches to normal or reduced levels of operation to avoid oversubscribing the functioning power subsystem. This can also allow less-expensive servers to be built for a higher (common-case) performance requirement while maintaining the reliability, availability, and serviceability (RAS) redundancy feature expected of IBM servers.

System implementations

Although the basic design of the EnergyScale architecture is similar for all of the POWER6 processor-based systems, some differences on system implementations exist.

The Power 595 is the largest POWER6 processor-based server. These servers contain multiple boards, and the designs of their predecessor in the POWER5 product line already contain power measurement features. Such machines pose a significant challenge because of their scale and the additional complexity imposed by their hardware designs. The design approach for extending the EnergyScale architecture to them involves three changes:

- The EnergyScale architecture uses the existing power measurement function provided by the BPCs used in the power supplies.
- Rather than adding a TPMD card to each board, the design uses existing microcontrollers that are already embedded in the power distribution subsystem (inside DCA assembly). This allows real-time control on each board.
- System-wide changes, such as to the frequency and the reporting of system-wide measurements, use non-real-time implementations running on a service processor. Although this limits the responsiveness of the power management system, this allows it to scale to the scope needed to control a very large machine.

The Power 595 uses the existing MDC microcontroller found on each DCA to perform the TPMD's functions and executes the firmware that run there. It uses communication between the two MDCs on each processor book and the embedded node controller service processor (NC) to measure the book-level power. The power is sensed by using the VRMs. To collect the power consumption for the entire 595 server, the embedded node controller service processor must pass each measurement to the embedded system controller service processor for summation. Data collection is through the Ethernet connection (BPH). Voltage and frequency adjustment for power save mode is always implemented by the service processor because the system controller service processor have access to the redundant clocks of the Power 595.

Table 2-9 indicates which functions the Power 595 supports.

Table 2-9 Functions available

Server model	Power trending	Power saver mode	Power capping	Processor core nap	I/O	Oversub-scription
Power 575/595 (>= 4.0 GHz)	✓	✓	—	✓	✓	✓

Table 2-10 lists power saver mode frequency drops.

Table 2-10 Power saver mode frequency table

Power saver mode frequency table	Frequency drop	Estimated processor power saved
5.0 GHz 595	20%	25-35%
4.2 GHz 595 without GX Dual Port RIO-2 Attach	14%	20-30%
4.2 GHz 595 with GX Dual Port RIO-2 Attach	5%	5-10%

Note: Required minimum firmware and software levels:

- EM330; Active Energy Manager 3.1, IBM Director 5.20.2
- HMC V7 R320.0 - However, it is recommended that HMC code level is equal to or higher than firmware for additional feature support.

EnergyScale offers the following value proposals:

- Finer data center control and management

  Enables detailed power and temperature measurement data for trending and analysis, configurable power and thermal limits, the reduction or elimination of over-provisioning found in many data centers, and reduction or avoidance of costly capital construction in new and existing data centers.

- Enhanced availability

  Enables continuous system operation, enhanced reliability, and better performance under faulty power and thermal conditions. It allows you to better react to key component failure such as power supply and cooling faults, thus lowering operating costs, and enables you to configure power and thermal caps (where enabled).

- Improved performance at lower costs

  Enables you to dynamically maintain the power and temperature within prescribed limits, reduce your cost by simplifying the facilities needed for power and cooling,

- Consistent power management for all Power System offerings from BladeCenter to Power 595.

2.4 System cooling

CEC cooling is provided by up to four air-moving devices (high-pressure, high-flow blowers) that mount to a plenum on the rear of the CEC cage (refer to Figure 2-3 on page 40). Air is drawn through all plugged nodes in parallel. In a hot room or under certain fault conditions, blower speed can increase to maintain sufficient cooling. Figure 2-25 shows air flow through the CEC.

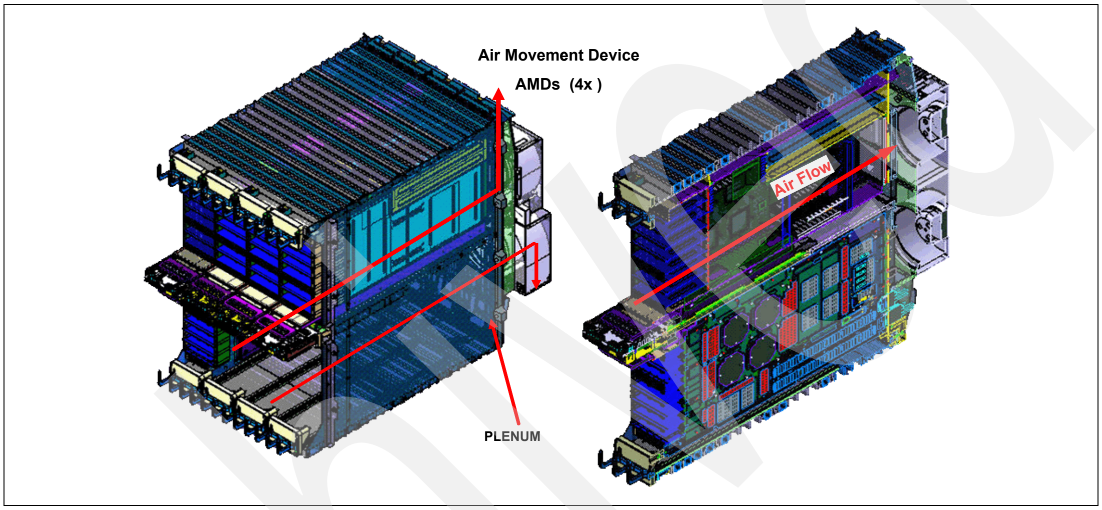


Figure 2-25 CEC internal air flow

Four motor drive assemblies (MDAs) mount on the four air moving devices (AMD™), as follows. A light strip LED identifies AMD and MDA.

- MDA 1 & 3 are powered by a Y-cable from the BPC – Connector J10.
- MDA 2 & 4 are powered by a Y-cable from the BPC – Connector J11.

Table 2-11 details the blower population

Table 2-11 Books

Processor book quantity	AMD
1 or 2 processor books	A1 and A3
3 or more processor books	A1, A2, A3, A4

2.5 Light strips

The Power 595 server uses a front and back light strip for service. The front and rear light strips each have redundant control modules that can receive input from either System Controller (SC).

To identify FRUs within a node (MCMs, DIMMs, hub cards, or node controllers), both the FRU LED (within the node, or on the light strip) and the node LED (on the light strip) must be on. The front light strip is shown in Figure 2-26.

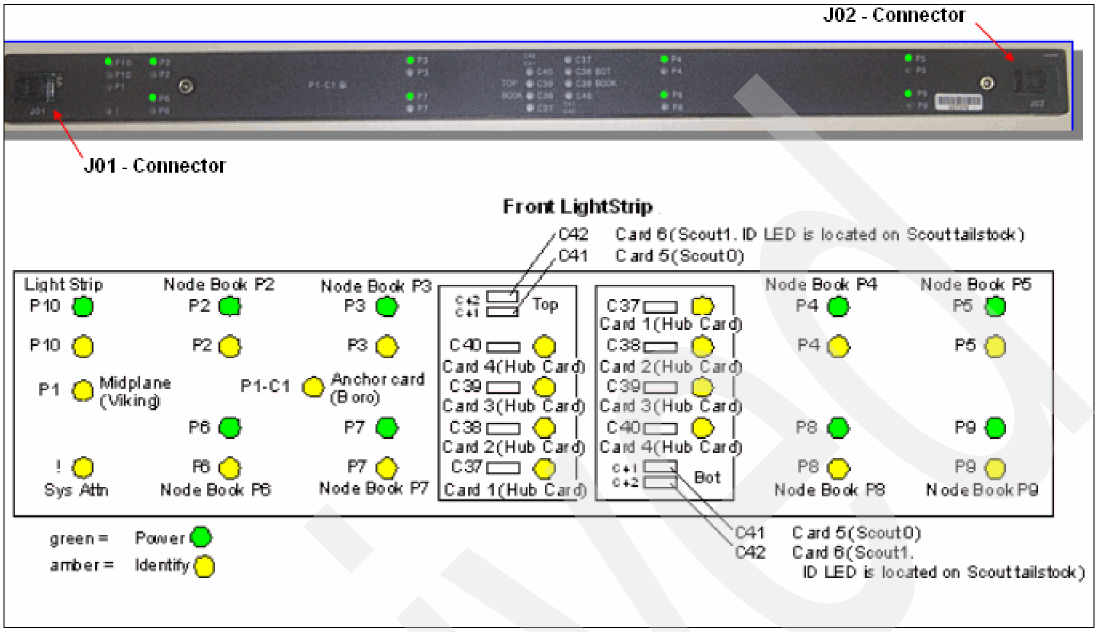


Figure 2-26 Front light strips

To identify card FRUs, both the node book LED and the card FRU LED must be on. The rear light strip is shown in Figure 2-27.

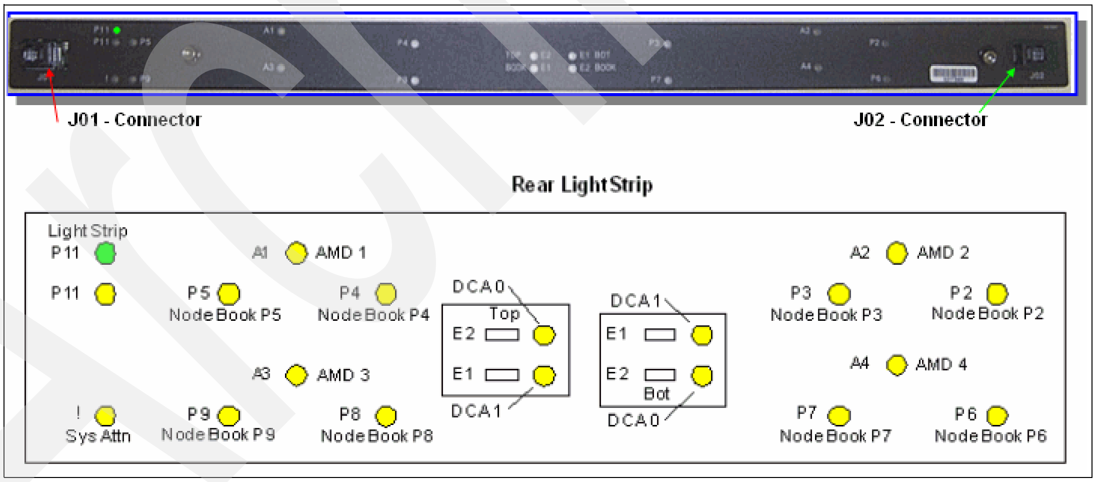


Figure 2-27 Rear light strip

To identify DCAs, both the node book LED and the DCA LED must be on.

2.6 Processor books

The 595 server can be configured with one to eight POWER6, 4.2 GHz or 5.0 GHz, 8-core processor books. All processor books installed in a 595 server must operate at the same speed. Figure 2-28 on page 71 shows the Power 595 processor book architecture.

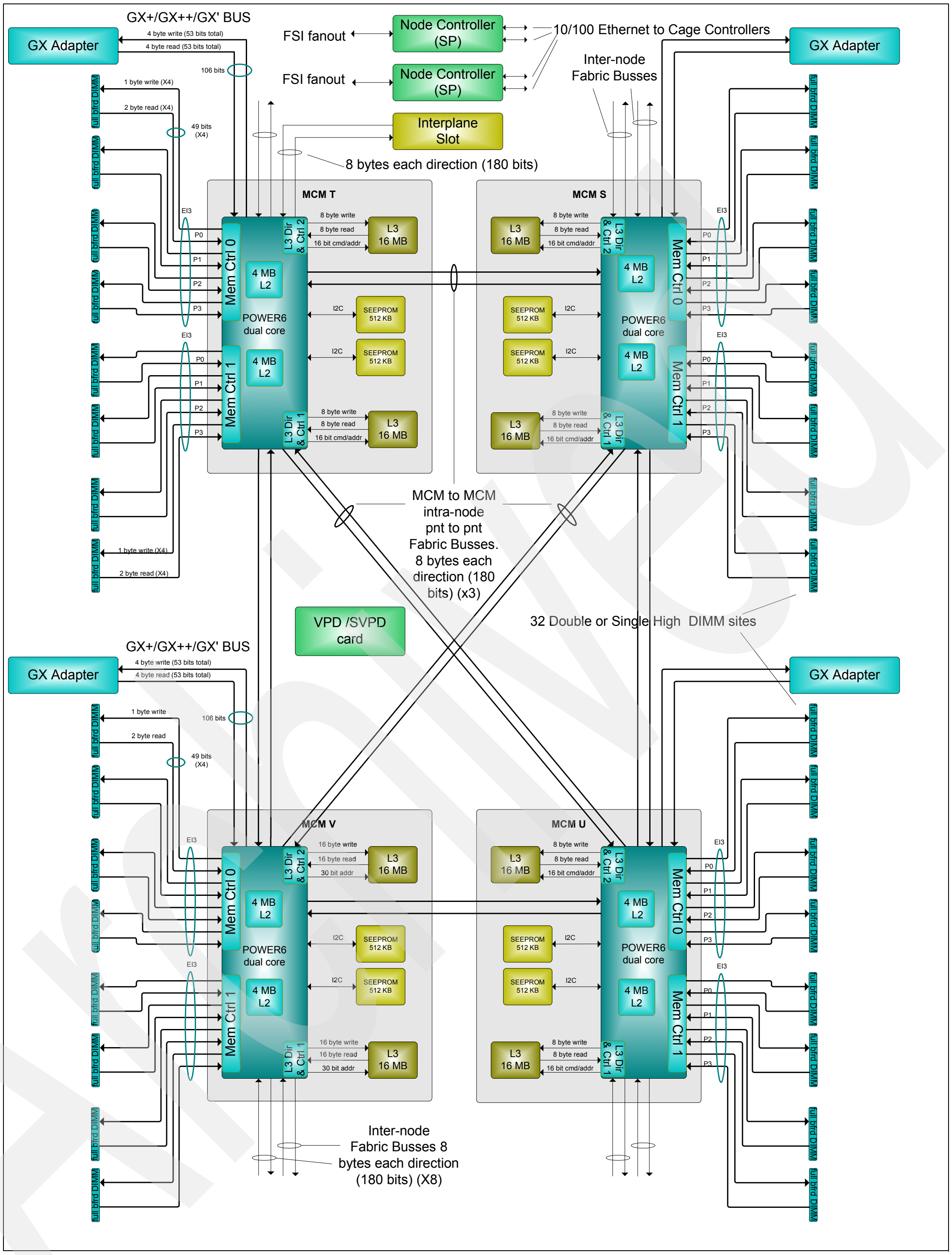


Figure 2-28 IBM Power 595 processor book architecture

The available processor books are listed in Table 2-12.

Table 2-12 Available processor books

Feature code	Description
#4694	0/8-core POWER6 4.2 GHz CoD 0-core Active Processor Book

Feature code	Description
#4695	0/8-core POWER6 5.0 GHz CoD 0-core Active Processor Book

Note: The minimum configuration requirement is one 4.2 GHz processor book with three processor activations, or two 5.0 GHz processor books with six processor activations.

Several methods for activating CoD POWER6 processors are available. Table 2-13 lists the CoD processor activation features and corresponding CoD modes. Additional information about the CoD modes are provided in section 3.3, “Capacity on Demand” on page 111.

Table 2-13 CoD processor activation features

Feature code	Description	CoD Mode	Support		
			AIX	IBM i	Linux
#4754	Processor activation for processor book #4694	Permanent	✓	✓	✓
#4755	Processor activation for processor book #4695	Permanent	✓	✓	✓
#7971	On/Off Processor Enablement	On/Off, Utility	✓	✓	✓
#7234	On/Off Processor CoD Billing, 1 Proc-Day, for #4694	On/Off	✓	–	✓
#7244	On/Off Processor CoD Billing, 1 Proc-Day, for #4695	On/Off	✓	–	✓
#5945	On/Off Processor CoD Billing, 1 Proc-Day, for #4694, IBM i	On/Off	–	✓	–
#5946	On/Off Processor CoD Billing, 1 Proc-Day, for #4695, IBM i	On/Off	–	✓	–
#5941	100 Processor Minutes for #4694	Utility	✓	–	✓
#5942	100 Processor Minutes for #4695	Utility	✓	–	✓
#5943	100 Processor Minutes for #4694, IBM i	Utility	–	✓	–
#5944	100 Processor Minutes for #4694, IBM i	Utility	–	✓	–

Each 8-core book contains four dual-threaded 64-bit SMP POWER6 processors chips packaged in four MCMs as shown in Figure 2-29 on page 73. The processor book also provides 32 DIMM slots for DDR2 memory DIMMs and four GX bus slots for remote I/O hubs cards (RIO-2 and 12x) that are used to connect system I/O drawers.

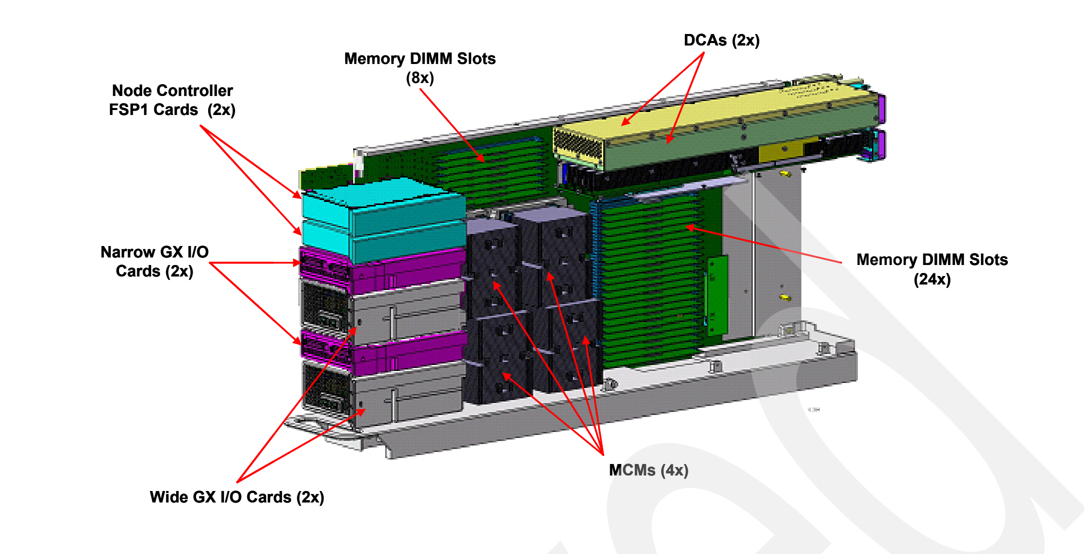


Figure 2-29 Power 595 processor book (shown in upper placement orientation)

> **Note:** All eight processor books are identical. They are simply inverted when plugged into the bottom side of the mid-plane.

Each MCM, shown in Figure 2-30, contains one dual-core POWER6 processor chip and two L3 cache chips.

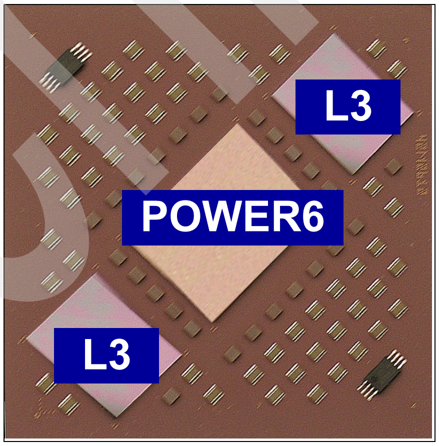


Figure 2-30 Multi-Chip Module (MCM)

The POWER6 processor chip provides 4 MB of on-board, private L2 cache per core. A total of 32 MB L3 cache is shared by the two cores.

2.6.1 POWER6 processor

The POWER6 processor capitalizes on all of the enhancements brought by the POWER5 processor. The POWER6 processor implemented in the Power 595 server includes additional features that are not implemented in the POWER6 processors within other Power Systems and System p servers. These features include:

- Dual, integrated L3 cache controllers
- Dual, integrated memory controllers

Two additional (of the many) enhancements to the POWER6 processor include the ability to perform processor instruction retry and alternate processor recovery. This significantly reduces exposure to both hard (logic) and soft (transient) errors in the processor core.

- Processor instruction retry

  Soft failures in the processor core are transient errors. When an error is encountered in the core, the POWER6 processor automatically retries the instruction. If the source of the error was truly transient, the instruction will succeed and the system will continue as before. On predecessor IBM systems, this error would have caused a checkstop.

- Alternate processor retry

  Hard failures are more challenging to recover from, being true logical errors that are replicated each time the instruction is repeated. Retrying the instruction does not help in this situation because the instruction will continue to fail. Systems with POWER6 processors introduce the ability to extract the failing instruction from the faulty core and retry it elsewhere in the system, after which the failing core is dynamically deconfigured and called out for replacement. The entire process is transparent to the partition owning the failing instruction. Systems with POWER6 processors are designed to avoid what would have been a full system outage.

Other enhancements include:

- POWER6 single processor checkstopping

  Typically, a processor checkstop would result in a system checkstop. A new feature in the 595 server is the ability to contain most processor checkstops to the partition that was using the processor at the time. This significantly reduces the probability of any one processor affecting total system availability.

- POWER6 cache availability

  In the event that an uncorrectable error occurs in L2 or L3 cache, the system is able to dynamically remove the offending line of cache without requiring a reboot. In addition, POWER6 utilizes an L1/L2 cache design and a write-through cache policy on all levels, helping to ensure that data is written to main memory as soon as possible.

  While L2 and L3 cache are physically associated with each processor module or chip, all cache is coherent. A coherent cache is one in which hardware largely hides, from the software, the fact that cache exists. This coherency management requires control traffic both within and between multiple chips. It also often means that data is copied (or moved) from the contents of cache of one core to the cache of another core. For example, if a core of chip one incurs a cache miss on some data access and the data happens to still reside in the cache of a core on chip two, the system finds the needed data and transfers it across the inter-chip fabric to the core on chip one. This is done without going through memory to transfer the data.

Figure 2-31 on page 75 shows a high-level view of the POWER6 processor. L1 Data and L1 Instruction caches are within the POWER6 core.

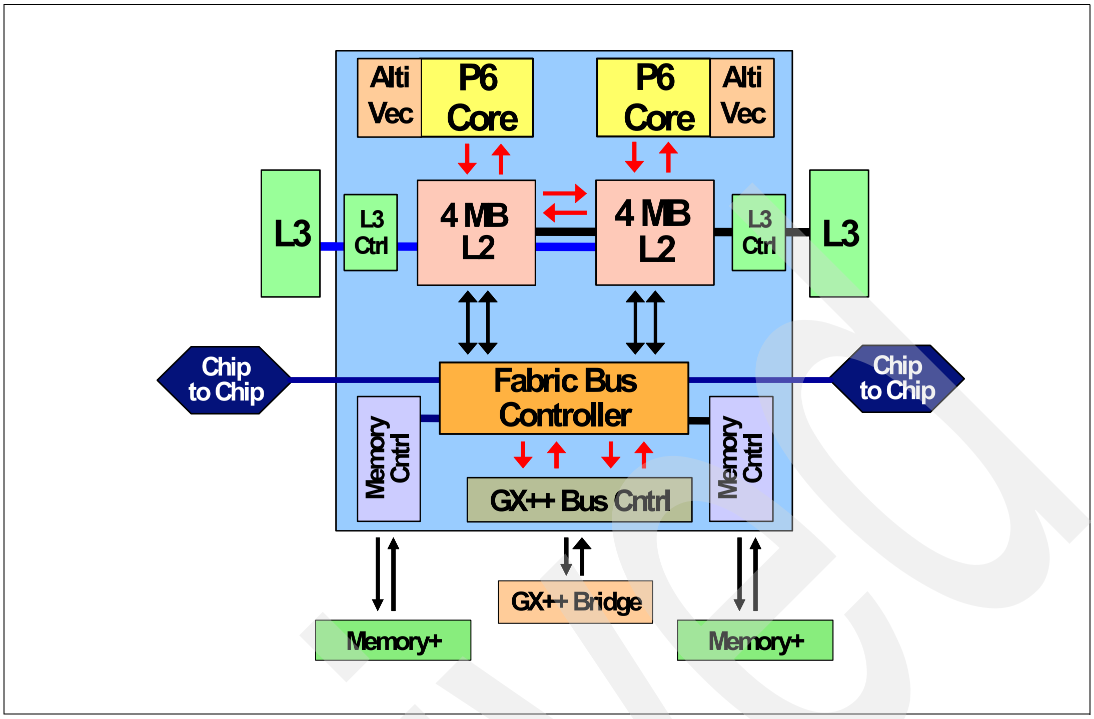


Figure 2-31 POWER6 processor

The CMOS 11S0 lithography technology in the POWER6 processor uses a 65 nm fabrication process, which enables:

- Performance gains through faster clock rates from up to 5.0 GHz
- Physical size of 341 mm

The POWER6 processor consumes less power and requires less cooling. Thus, you can use the POWER6 processor in servers where previously you could only use lower frequency chips due to cooling restrictions.

The 64-bit implementation of the POWER6 design provides the following additional enhancements:

- Compatibility of 64-bit architecture
  - Binary compatibility for all POWER and PowerPC® application code level
  - Support of partition migration
  - Support big and little endian
  - Support of four page sizes: 4 KB, 64 KB, 16 MB, and 16 GB
- High frequency optimization
  - Designed to operate at maximum speed of 5 GHz
- Superscalar core organization
  - Simultaneous multithreading: two threads
- In-order dispatch of five operations (through a single thread) or seven operations (using Simultaneous Multithreading) to nine execution units:
  - Two load or store operations
  - Two fixed-point register-register operations

- Two floating-point operations
- One branch operation

The POWER6 processor implements the 64-bit IBM Power Architecture® technology. Each POWER6 chip incorporates two ultrahigh dual-threaded Simultaneous Multithreading processor cores, a private 4 MB level 2 cache (L2) for each processor, integrated memory controller and data interconnect switch and support logic for dynamic power management, dynamic configuration and recovery, and system monitoring.

2.6.2 Decimal floating point

This section describes the behavior of the POWER6 hardware decimal floating-point processor, the supported data types, formats, and classes, and the usage of registers.

The decimal floating-point (DFP) processor shares the 32 floating-point registers (FPRs) and the floating-point status and control register (FPSCR) with the binary floating-point (BFP) processor. However, the interpretation of data formats in the FPRs, and the meaning of some control and status bits in the FPSCR are different between the DFP and BFP processors.

The DFP processor supports three DFP data formats:

- DFP32 (single precision): 4 bytes, 7 digits precision, -95/+96 exponent
- DFP64 (double precision): 8 bytes, 16 digits precision, -383/+384 exponent
- DFP128 (quad precision): 16 bytes, 34 digits precision, -6143/+6144 exponent

Most operations are performed on the DFP64 or DFP128 format directly. Support for DFP32 is limited to conversion to and from DFP64. For some operations, the DFP processor also supports operands in other data types, including signed or unsigned binary fixed-point data, and signed or unsigned decimal data.

DFP instructions that perform arithmetic, compare, test, quantum-adjustment, conversion, and format operations on operands held in FPRs or FPR pairs are:

Arithmetic instructions Perform addition, subtraction, multiplication, and division operations.

Compare instructions Perform a comparison operation on the numerical value of two DFP operands.

Test instructions Test the data class, the data group, the exponent, or the number of significant digits of a DFP operand.

Quantum-adjustment instructions
Convert a DFP number to a result in the form that has the designated exponent, which can be explicitly or implicitly specified.

Conversion instructions Perform conversion between different data formats or data types.

Format instructions Facilitate composing or decomposing a DFP operand.

Enabling applications running on POWER6 systems to take advantage of the hardware decimal floating point support depends on the programming language release level used by the application and the operating system in which the application is running.

Examples are discussed in the following list:

- Java applications: Applications running IBM Technology for Java 6.0 32-bit and 64-bit JVM automatically take advantage of the hardware assist during the initial just in time (JIT) processing. Applications running under IBM i require release level 6.1. Java 5.0 does not use DCP.

  C and C++ applications: For the C and C++ compilers running under AIX and Linux for Power, as of v9.0, DFP support through the POWER6 hardware instructions is available. Software emulation is supported on all other POWER architectures.

  Running under IBM i 6.1, support for DFP has been added to the IBM i 6.1 ILE C compiler. If a C program that uses DFP data is compiled on POWER 6 hardware, hardware DFP instructions is generated; otherwise, software emulation is used.

  IBM i support for DFP in the ILE C++ compiler is planned for a future release.

  For your information, C and C++ on z/OS®, as of V1R9, use hardware DFP support where the run time code detects hardware analogous to POWER 6.

- IBM i ILE RPG and COBOL: These languages do not use decimal floating point. The normal zoned decimal or packed decimal instructions receive *normal* performance gains merely by running under IBM i 6.1 on POWER6.

  IBM i 6.1 supports decimal floating point data, for example, in DB2 for i5/OS tables. If the RPG or COBOL compiler encounters a decimal float variable in an externally-described file or data structure, it will ignore the variable and issue an identifying information message.

- Some applications, such those available from SAP®, that run on POWER6-based systems, can provide specific ways to take advantage of decimal floating point.

  For example, the SAP NetWeaver® 7.10 ABAP™ kernel introduces a new SAP ABAP data type called *DECFLOAT* to enable more accurate and consistent results from decimal floating point computations. The decimal floating point (DFP) support by SAP NetWeaver leverages the built-in DFP feature of POWER6 processors. This allows for simplified ABAP-coding while increasing numeric accuracy and with a potential for significant performance improvements.

2.6.3 AltiVec and Single Instruction, Multiple Data

IBM semiconductor's advanced Single Instruction, Multiple Data (SIMD) technology based on the AltiVec instruction set is designed to enable exceptional general-purpose processing power for high-performance POWER processors. This leading-edge technology is engineered to support high-bandwidth data processing and algorithmic-intensive computations, all in a single-chip solution

With its computing power, AltiVec technology also enables high-performance POWER processors to address markets and applications in which performance must be balanced with power consumption, system cost and peripheral integration.

The AltiVec technology is a well known environment for software developers who want to add efficiency and speed to their applications. A 128-bit vector execution unit was added to the architecture. This engine operates concurrently with the existing integer and floating-point units and enables highly parallel operations, up to 16 operations in a single clock cycle. By leveraging AltiVec technology, developers can optimize applications to deliver acceleration in performance-driven, high-bandwidth computing.

The AltiVec technology is not comparable to the IBM POWER6 processor implementation, using the simultaneous multithreading functionality.

2.7 Memory subsystem

The Power 595 server uses fully buffered, Double Data Rate (DDR2) DRAM memory DIMMs. The DIMM modules are X8 organized (8 data bits per module). Support is provided for migrated X4 DIMM modules. Memory DIMMs are available in the following capacities: 1 GB, 2 GB, 4 GB, 8 GB, and 16 GB. Each orderable memory feature (memory unit) provides four DIMMs.

The memory subsystem provides the following levels of reliability, availability, and serviceability (RAS):

- ECC, single-bit correction, double-bit detection
- Chip kill correction
- Dynamic bit steering
- Memory scrubbing
- Page deallocation (AIX only)
- Dynamic I/O bit line repair for bit line between the memory controller and synchronous memory interface chip (SMI) and between SMI chips. The SMI chips connect the memory controllers to memory DIMMs.
- ECC on DRAM addressing provided by SMI chip
- Service processor interface

Each of the four dual-core POWER6 processors within a processor book has two memory controllers as shown in Figure 2-32 on page 79. Each memory controller is connected to a memory unit. The memory controllers use an elastic interface to the memory DIMMs that runs at four times the memory speed.

Note: One memory unit for each POWER6 processor must be populated at initial order (four units per installed processor book).

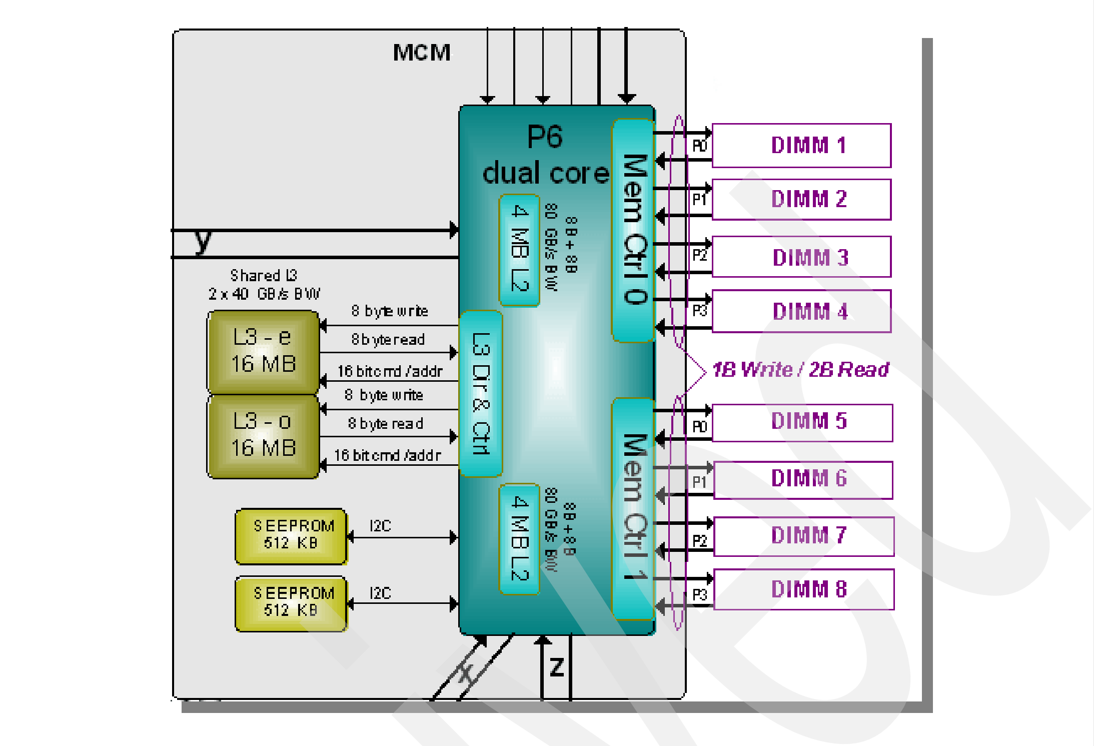


Figure 2-32 Memory system logical view

Each processor book supports a total of eight memory units (32 DIMMS or eight memory features). A fully configured Power 595 server with eight processor books supports up to 64 memory units (256 DIMMs). Using memory features that are based on 64 GB DIMMs, the resulting maximum memory configuration is four TB.

Note: One memory unit is equal to an orderable memory feature. One memory unit contains four memory DIMMs.

2.7.1 Memory bandwidth

The Power 595 memory subsystem consists of L1, L2, and L3 caches along with the main memory. The bandwidths for these memory components is shown in Table 2-14

Table 2-14 Memory bandwidth

Description	Bus size	Bandwidth
L1 (data)	2 x 8 bytes	80 GBps
L2	2 x 32 bytes	160 GBps
L3	4 x 8 bytes	80 GBps (per 2-core MCM) 2.56 TBps (per 64-core system)
Main memory	4 x 1 byte (write) 4 x 2 bytes (read)	42.7 GBps (per 2-core MCM) 1.33 TBps (per 64-core system)

2.7.2 Available memory features

The available memory features (4 DIMM units) for the 595 server are shown in Table 2-15.

Table 2-15 Available memory features

Feature code	Description	Speed (MHz)	Minimum activation	Maximum system memory
#5693	0/4 GB DDR2 Memory (4X1 GB)	667	100%	256 GB
#5694	0/8 GB DDR2 Memory (4X2 GB)	667	50%	512 GB
#5695	0/16 GB DDR2 Memory (4X4 GB)	533	50%	1024 GB
#5696	0/32 GB DDR2 Memory (4X8 GB)	400	50%	2048 GB
#5697[a]	0/64 GB DDR2 Memory(4X16 GB)[b]	400	100%	4096 GB
#8201	0/256 GB 533 MHz DDR2 Memory Package (32 x #5694)	667	100%	512 GB
#8202	0/256 GB 533 MHz DDR2 Memory Package (16 x #5695)	533	100%	1024 GB
#8203	0/512 GB 533 MHz DDR2 Memory Package (3 x #5695)	533	100%	1024 GB
#8204	0/512 GB 400 MHz DDR2 Memory Package (16 x #5696)	400	100%	2048 GB
#8205[c]	0/2 TB 400 MHz DDR2 Memory Package (32x #5697)[1]	400	100%	4096 GB

[a] Memory feature #5697, which uses 16 GB memory DIMMs, has a planned availability date of November 21, 2008.
[b] The 16 GB DIMMS are only available with the 5.0 GHz processor option.
[c] Memory feature #8205, which uses 16 GB memory DIMMs, has a planned availability date of November 21, 2008.

All memory features for the 595 server are shipped with zero activations. A minimum percentage of memory must be activated for each memory feature ordered.

For permanent memory activations, choose the desired quantity of memory activation features from Table 2-16 that corresponds to the amount of memory that you would like to permanently activate.

Table 2-16 Permanent memory activation features

Feature code	Description
#5680	Activation of 1 GB DDR2 POWER6 memory
#5681	Activation of 256 GB DDR2 POWER6 memory

Memory can also be temporarily activated using the feature codes provided in Table 2-17. For additional discussion on CoD options, see Chapter 3, CoD Options.

Table 2-17 CoD memory activation features

Feature code	Description
#5691	On/Off, 1 GB-1Day, memory billing POWER6 memory
#7973	On/Off Memory Enablement

2.7.3 Memory configuration and placement

Each processor book features four MCMs. The layout of the MCMs and their corresponding memory units (a unit is a memory feature, or 4 DIMMs) is shown in Figure 2-33.

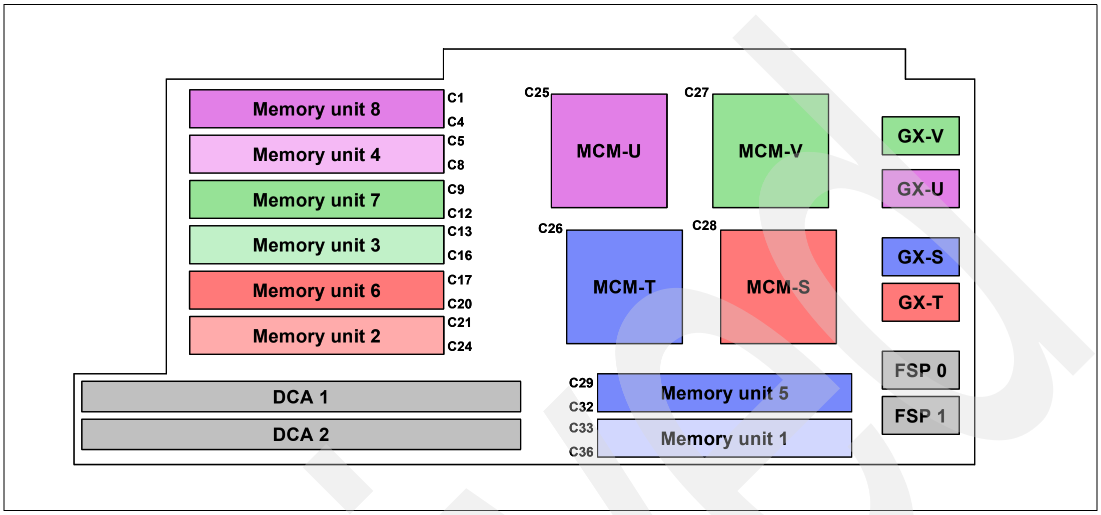


Figure 2-33 Processor book with MCM and memory locations

Table 2-18 shows the sequence in which the memory DIMMs are populated within the processor book. Memory units one through four must be populated on every processor book. Memory units five through eight are populated in pairs (5 and 6, 7 and 8) and do not have to be uniformly populated across the installed processor books. For example, on a system with three processor books, it is acceptable to have memory units 5 and 6 populated on just one of the processor books.

Table 2-18 Memory DIMM installation sequence

Installation sequence	Memory unit	MCM
1	C33-C36	MCM-S (C28)
2	C21-C24	MCM-T (C26)
3	C13-C16	MCM-V (C27)
4	C5-C8	MCM-U (C25)
5	C29-C32	MCM-S (C28)
6	C17-C20	MCM-T(C26)
7	C9-C12	MCM-V (C27)
8	C1-C4	MCM-U (C25)

Within a 595 server, individual processor books can contain memory different from that contained in another processor book. However, within a processor book, all memory must be comprised using identical memory features.

For balanced memory performance within a 595 server, it is recommended that mixed memory should not be different by more than 2x in size. That is, a mix of 8 GB and 16 GB features is acceptable, but a mix of 4 GB and 16 GB is not recommended within a server.

When multiple DIMM sizes are ordered, smaller DIMM sizes are placed in the fewest processor books possible, while insuring that the quantity of remaining larger DIMMs are adequate to populate at least one feature code per MCM module. The largest DIMM size is spread out among all remaining processor books. This tends to balance the memory throughout the system.

For memory upgrades, DIMMs are added first to those books with fewer DIMMs until all books have the same number of DIMMs. Any remaining memory is then distributed round robin amongst all books having that size DIMM.

The following memory configuration and placement rules apply to the 595 server:

- At initial order, each installed processor book must have a minimum of:
  - Four memory units installed (50% populated). The memory units must use the same DIMM size within the processor book. Different DIMM sizes can be used within the 595 server. For 16 GB DIMMs, memory units must be installed in groups of eight.
  - 16 GB of memory activated
- Memory upgrades can be added in groups of two units (16 GB DIMMs must be added in groups of eight units), as follows:
  - For memory upgrades, you are not required to add memory to all processor books.
  - You must maintain the same DIMM sizes within a processor book when adding memory.

  Processors books are 50% (initial), 75%, or 100% populated. Put another way, each processor book will have either, four, six, or eight memory units installed.

2.8 Internal I/O subsystem

Each processor book on the 595 server provides four GX busses for the attachment of GX bus adapters. A fully configured 595 server with eight processor books supports up to 32 GX bus adapters. The GX bus adapter locations are shown in Figure 2-34.

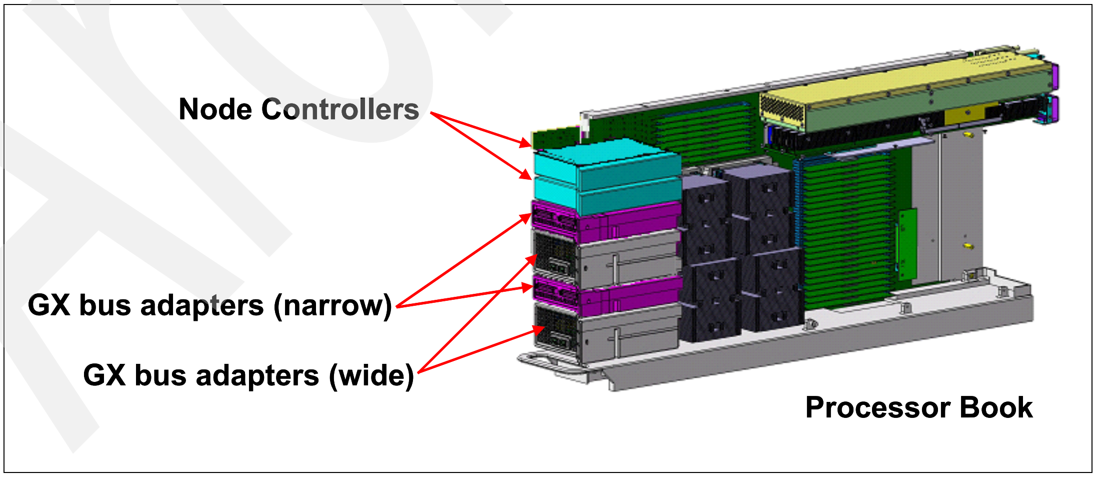


Figure 2-34 GX bus adapters

The processor book provides two narrow and two wide GX bus adapter slots. Narrow adapters fit into both narrow and wide GX bus slots.

2.8.1 Connection technology

RIO-2 and 12x connectivity is provided using GX bus adapter based, remote I/O hubs. These remote I/O hubs are listed in Table 2-19.

Table 2-19 Remote I/O hubs

Feature	Description	Form factor	Attach to drawer(s)	Support		
				AIX	IBM i	Linux
#1814	Remote I/O-2 (RIO-2) Loop Adapter, Two Port	narrow	5791	✓	—	✓
#1816	GX Dual-Port 12x HCA	narrow	5797, 5798	✓	✓	✓

Each I/O hub provides two ports that are used to connect internal 24-inch I/O drawers to the CEC.

The RIO-2 I/O hubs are currently available and the 12x I/O hubs have a planned-availability date of November 21, 2008.

I/O hub adapter plugging rules

The I/O hubs are evenly distributed across the installed processor books. The installation order follows the processor book plugging sequence listed in table Table 2-1 on page 42 with the following order of priority:

1. Bottom narrow slots are across all processor nodes.
2. Upper narrow slots are across all processor nodes.
3. Bottom wide slots are across all processor nodes.
4. Upper wide slots are across all processor nodes.

This information (bottom and upper notation) is applicable regardless of the orientation of the processor books (upper or lower). For example, bottom means bottom whether you are plugging into a processor book installed in an upper or lower location.

> **Important:** When your Power 595 server is manufactured, the I/O hubs are evenly distributed across the installed processor books. I/O connections will then be distributed across these installed I/O hubs. If you add more I/O hubs during an upgrade, install them so that the end result is an even balance across all new and existing processor books. Therefore, the cabling relationship between the I/O hubs and drawers can vary with each Power 595 server. We suggest that you document these connections to assist with system layout and maintenance. I/O hubs cards can be hot-added. Concurrent re-balancing of I/O hub cards is not supported.

An example of the I/O hub installation sequence for a fully configured system with eight processor books and 32 I/O hubs is shown in Figure 2-35 on page 84.

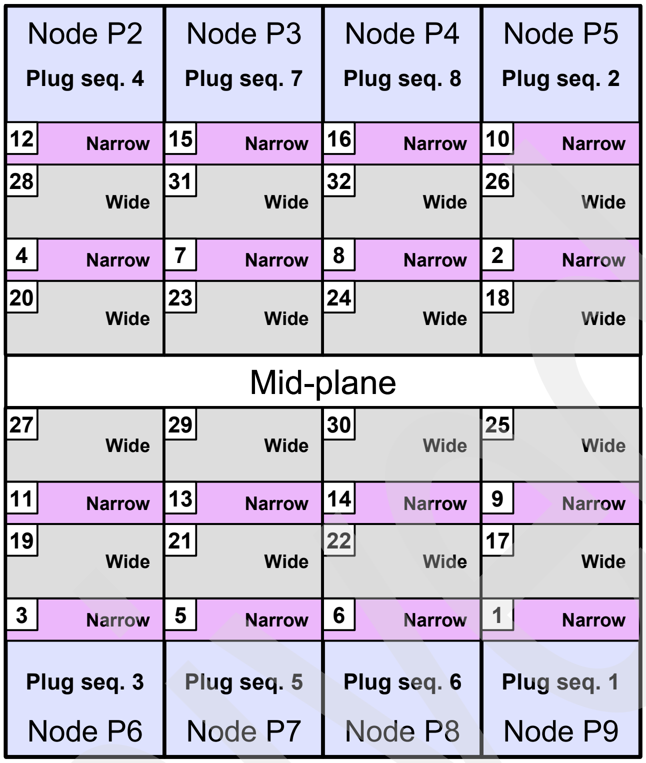


Figure 2-35 I/O hub installation sequence

2.8.2 Internal I/O drawers

The internal I/O drawers (24 inches) provide storage and I/O connectivity for the 595 server. The available internal I/O drawers are listed in Table 2-20

Table 2-20 Internal I/O drawers

Feature	Description	Connection Adapter	Support		
			AIX	IBM i	Linux
#5791	I/O drawer, 20 slots, 16 disk bays	1814	✓	—	✓
#5797	12x I/O drawer, 20 slots, 16 disk bays, with repeater	1816	✓	✓	✓
#5798	12x I/O drawer, 20 slots, 16 disk bays, no repeater	1816	✓	✓	✓

I/O drawers #5791 and #5797 (with repeater) are supported in the system (CEC) rack, powered expansion racks, and nonpowered expansion racks. I/O drawer #5798 (without repeater) is only supported in the system rack.

Note: I/O drawers #5797 and #5798 have a planned availability date of November 21, 2008.

Figure 2-36 shows the components of an internal I/O drawer. The I/O riser cards provide RIO-2 or 12x ports that are connected via cables to the I/O hubs located in the processor books within the CEC.

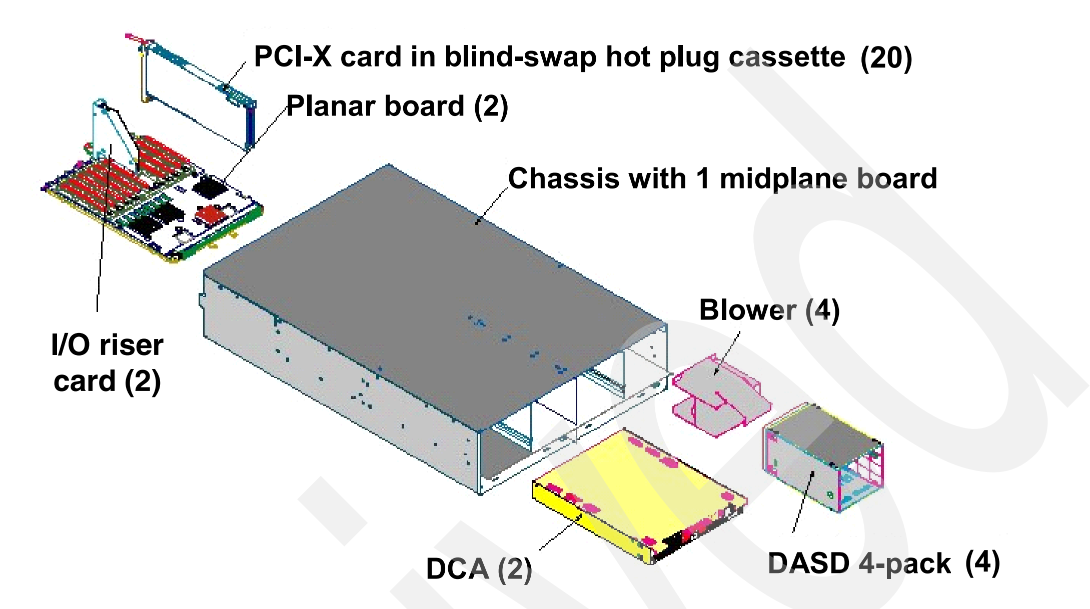

PCI-X card in blind-swap hot plug cassette (20)
Planar board (2)
Chassis with 1 midplane board
Blower (4)
I/O riser card (2)
DCA (2)
DASD 4-pack (4)

Figure 2-36 I/O drawer internal view

Each I/O drawer is divided into two halves. Each half contains 10 blind-swap adapter slots (3.3 V) and two Ultra3 SCSI 4-pack backplanes for a total of 20 adapter slots and 16 hot-swap disk bays per drawer. The internal SCSI backplanes provide support for the internal drives and do not have an external SCSI connector. Each half of the I/O drawer is powered separately.

Additional I/O drawer configuration requirements:

- A blind-swap hot-plug cassette is provided in each PCI-X slot of the I/O drawer. Cassettes not containing an adapter are shipped with a plastic filler card installed to help ensure proper environmental characteristics for the drawer. Additional blind-swap hot-plug cassettes can be ordered: #4599, PCI blind-swap cassette kit.
- All 10 adapter slots on each I/O drawer planar are capable of supporting either 64-bit or 32-bit 3.3 V based adapters.
- For maximum throughout, use two I/O hubs per adapter drawer (one I/O hub per 10 slot planar). This is also known as double-barrel cabling configuration (dual loop). Single-loop configuration is supported for configurations with a large number of internal I/O drawers.

Table 2-21 compares features of the RIO-2 and 12x based internal I/O drawers.

Table 2-21 Internal I/O drawer feature comparison

Feature or Function	#5791 drawer	#5797, #5798 drawers
Connection technology	RIO-2	12x
Bandwidth per connection port (4 ports per drawe)r	1.7 GBps sustained 2 GBps peak	5 GBps sustained 6 GBps peak
PCI-X (133 MHz) slots	10 per planar (20 total)	3 per planar (6 total)

Feature or Function	#5791 drawer	#5797, #5798 drawers
PCI-X 2.0 (266 MHz) slots	none	7 per planar (14 total)
Ultra3 SCSI busses	2 per planar (4 total)	2 per planar (4 total)
SCSI disk bays	8 per planar (16 total)	8 per planar (16 total)
Maximum drawers per system	12	30 (#5797) 3 (#5798) 30 (#5797 and #5798)

RIO-2 based internal I/O drawer (#5791)

The 5791 internal I/O drawer uses RIO-2 connectivity to the CEC. All 20 slots are PCI-X based. An internal diagram of the #5791 internal I/O drawer is shown in Figure 2-37.

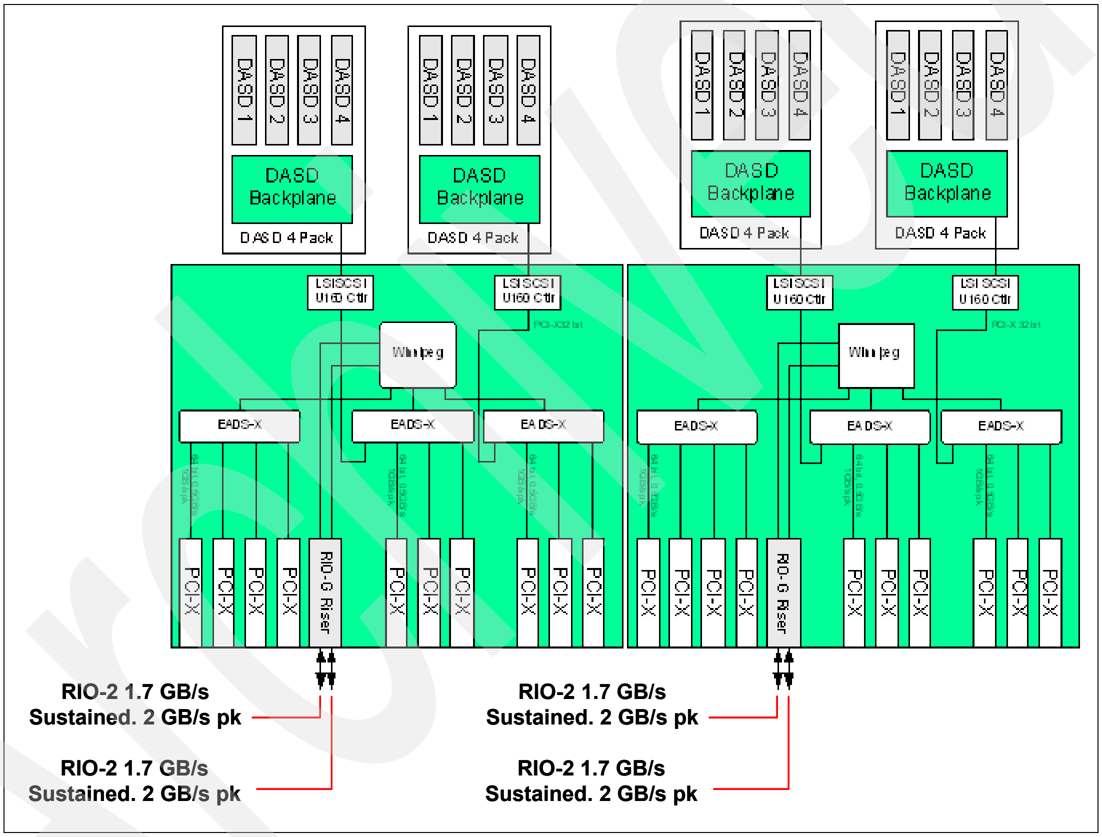


Figure 2-37 #5791 internal I/O Expansion Drawer (RIO-2)

12x based internal I/O drawers (#5797 and #5798)

The #5797 and #5798 internal I/O drawers use 12x connectivity to the CEC. Each I/O drawer provides a total of 14 PCI-X 2.0 (266 MHz) slots and 6 PCI-X (133 MHz) slots. An internal diagram of the #5797 and #5798 internal I/O drawers is shown in Figure 2-38 on page 87.

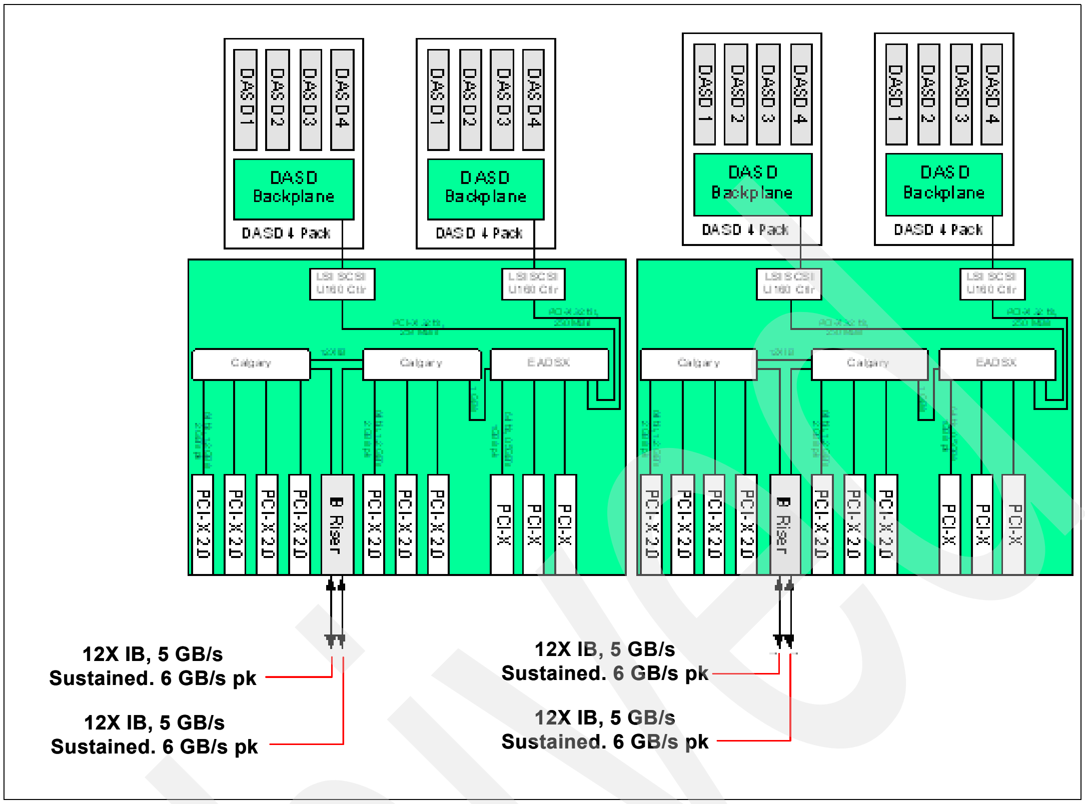

DASD 1
DASD 2
DASD 3
DASD 4
DASD Backplane
DASD 4 Pack
LSI SCSI U160 Ctlr
Calgary
EADSX
12X IB
PCI-X 2.0
IB Riser
PCI-X
12X IB, 5 GB/s Sustained. 6 GB/s pk
12X IB, 5 GB/s Sustained. 6 GB/s pk
12X IB, 5 GB/s Sustained. 6 GB/s pk
12X IB, 5 GB/s Sustained. 6 GB/s pk

Figure 2-38 #5797 and #5798 internal I/O drawers (12x)

The Power 595 server supports up to 30 expansion drawers (maximum of 12 for RIO-2).

Figure 2-39 on page 88 shows the drawer installation sequence when the integrated battery feature (IBF) is not installed. If the IBF is installed, the battery backup units will be located where I/O drawer #2 would have been located. Subsequent drawer numbering with IBF is shown in parenthesis.

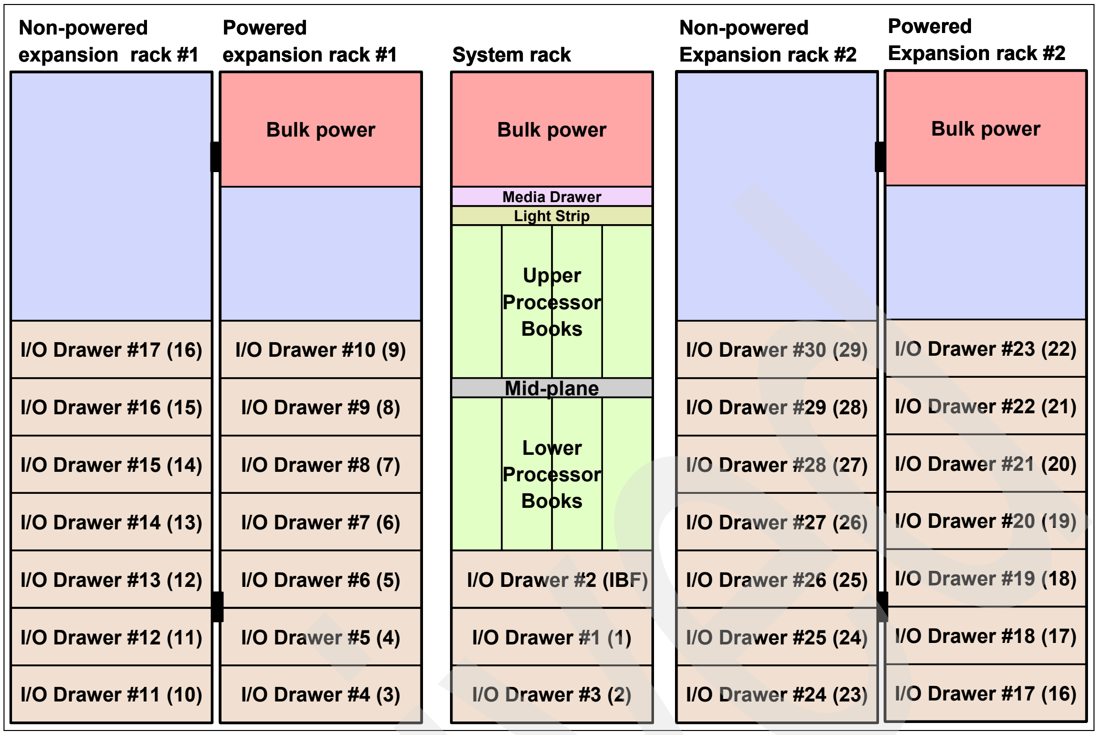


Figure 2-39 Power 595 I/O Expansion Drawer locations

2.8.3 Internal I/O drawer attachment

The internal I/O drawers are connected to the 595 server CEC using RIO-2 or 12x technology. Drawer connections are made in loops to help protect against errors resulting from an open, missing, or disconnected cable. If a fault is detected, the system can reduce the speed on a cable, or disable part of the loop to maintain system availability.

Each RIO-2 or 12x I/O attachment adapter (I/O hub) has two ports and can support one loop. A maximum of one internal I/O drawer can be attached to each loop. Up to four I/O hub attachment adapters can be installed in each 8-core processor book. Up to 12 RIO-2 or 30 12x I/O drawers are supported per 595 server.

I/O drawers can be connected to the CEC in either single-loop or dual-loop mode:

- Single-loop (Figure 2-40 on page 89) mode connects an entire I/O drawer to the CEC using one RIO-2 or 12x loop. In this configuration, the two I/O planars in the I/O drawer are connected together using a short cable. Single-loop connection requires one RIO-2 Loop Attachment Adapter (#1814) or GX Dual-Port 12x (#1816) per I/O drawer.
- Dual-loop (Figure 2-41 on page 90) mode connects each of the two I/O planars (within the I/O drawer) to the CEC on separate loops. Dual-loop connection requires two I/O hub attachment adapters (#1814 or #1816) per connected I/O drawer. With a dual-loop configurations, the overall I/O bandwidth per drawer is higher.

Note: Use dual-loop mode whenever possible to provide the maximum bandwidth between the I/O drawer and the CEC.

Table 2-22 on page 89 lists the number of single-looped and double-looped I/O drawers that can be connected to a 595 server based on the number of processor books installed.

Table 2-22 Number of RIO drawers that can be connected

Number of installed processor books	RIO-2 (#5791)		12x (#5797 & #5798)	
	Single-looped	Dual-looped	Single-looped	Dual-looped
1	4	2	4	2
2	8	4	8	4
3	12	6	12	6
4	12	8	16	8
5	12	10	20	10
6	12	12	24	12
7	12	12	28	14
8	12	12	30	16

2.8.4 Single loop (full-drawer) cabling

Single loop I/O drawer connections are shown in Figure 2-40.

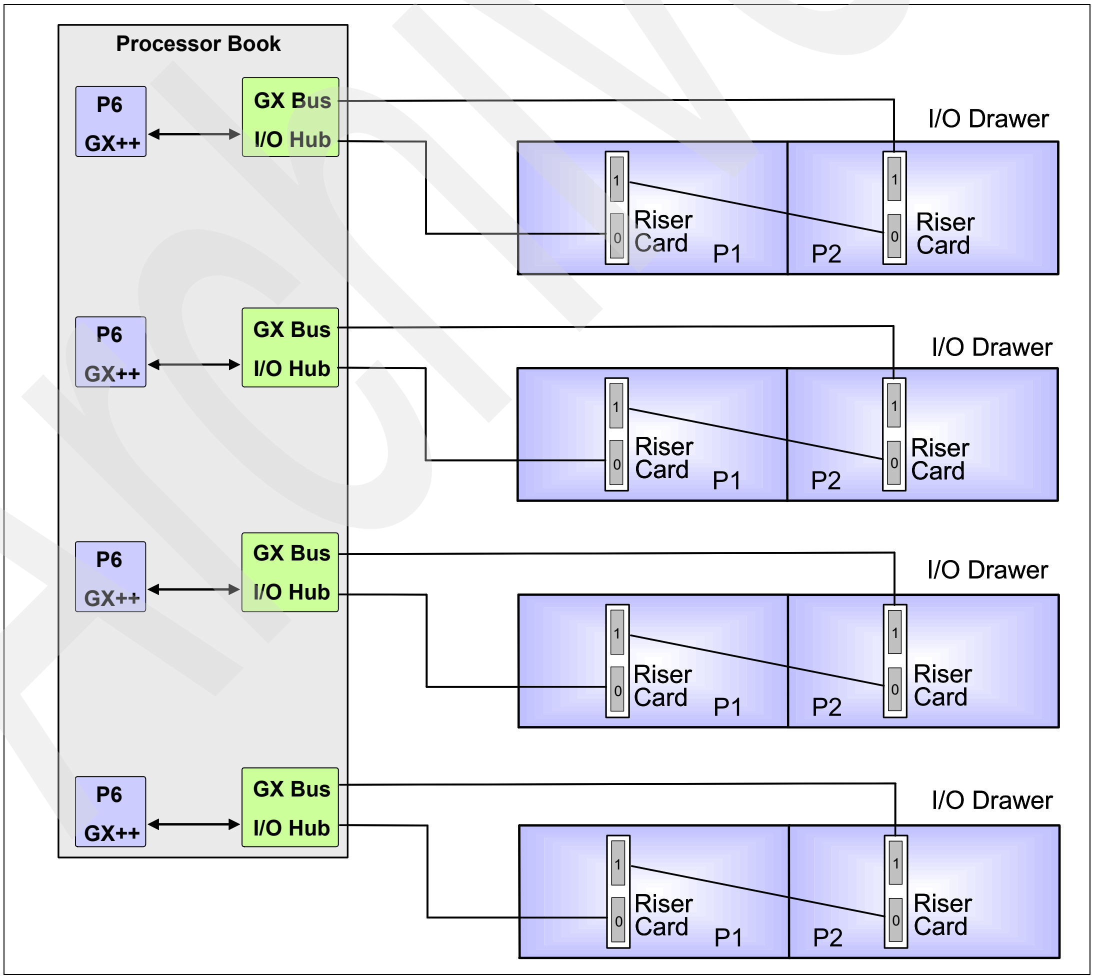


Figure 2-40 Single loop I/O drawer cabling

The short cable connecting the two halves of the drawer ensures that each planar (P1 and P2) within the drawer can access the I/O hub adapter card, even if one of the main connection cables is disconnected or damaged.

2.8.5 Dual looped (half-drawer) cabling

Each of the two internal I/O drawer planars can be cabled and addressed by an I/O hub individually using the preferred, dual loop (half drawer) cabling as shown in Figure 2-41.

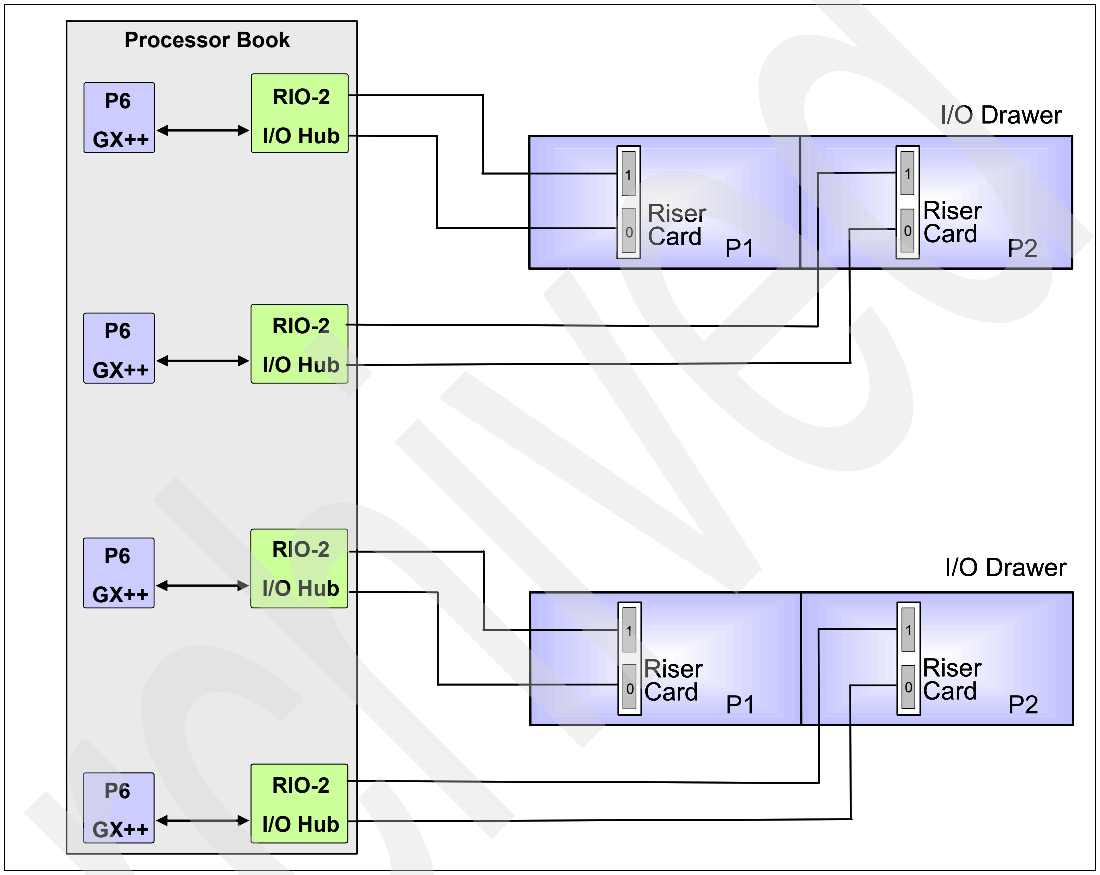


Figure 2-41 Dual loop I/O drawer (#5791)

2.8.6 I/O drawer to I/O hub cabling sequence

The I/O expansion drawers are connected using loops to the I/O hubs in the same order that the I/O hubs were installed as described in section 2.8.1, "Connection technology" on page 83 section. Most installations use dual loop configurations for the I/O expansion drawers and therefore each planar (half) of the drawer is connected to an individual I/O hub adapter as follows:

- The I/O expansion drawers are cabled in numerical order.
- Viewed from the rear side of the rack, the left planar is cabled first, followed by the right planar of the same I/O expansion drawer.
- For each planer:
  - The lower connector on the I/O expansion drawer riser card connects to the left connector on the I/O hub adapter in the processor book.

- – The upper connector on the I/O expansion drawer riser card will connect to the right connector on the I/O hub adapter in the processor book.

Loop connection sequence

The I/O expansion drawer loops is cabled to the I/O hubs using an even distribution across all of the installed processor books. The order follows the processor book plugging sequence listed in table Table 2-1 on page 42, with the following order of priority:

1. I/O hubs are installed in bottom narrow slots across all processor nodes.
2. I/O hubs are installed in upper narrow slots across all processor nodes.
3. I/O hubs are installed in bottom wide slots across all processor nodes.
4. I/O hubs are installed in upper wide slots across all processor nodes.

This information (bottom and upper notation) is applicable regardless of the orientation of the processor books (upper or lower). For example, bottom means bottom whether you are plugging into an upper or lower processor book.

Figure 2-42 shows an example of the loop cabling between an I/O expansion drawer and an I/O hub for a Power 595 server with one processor book, four I/O hubs, and two I/O expansion drawers.

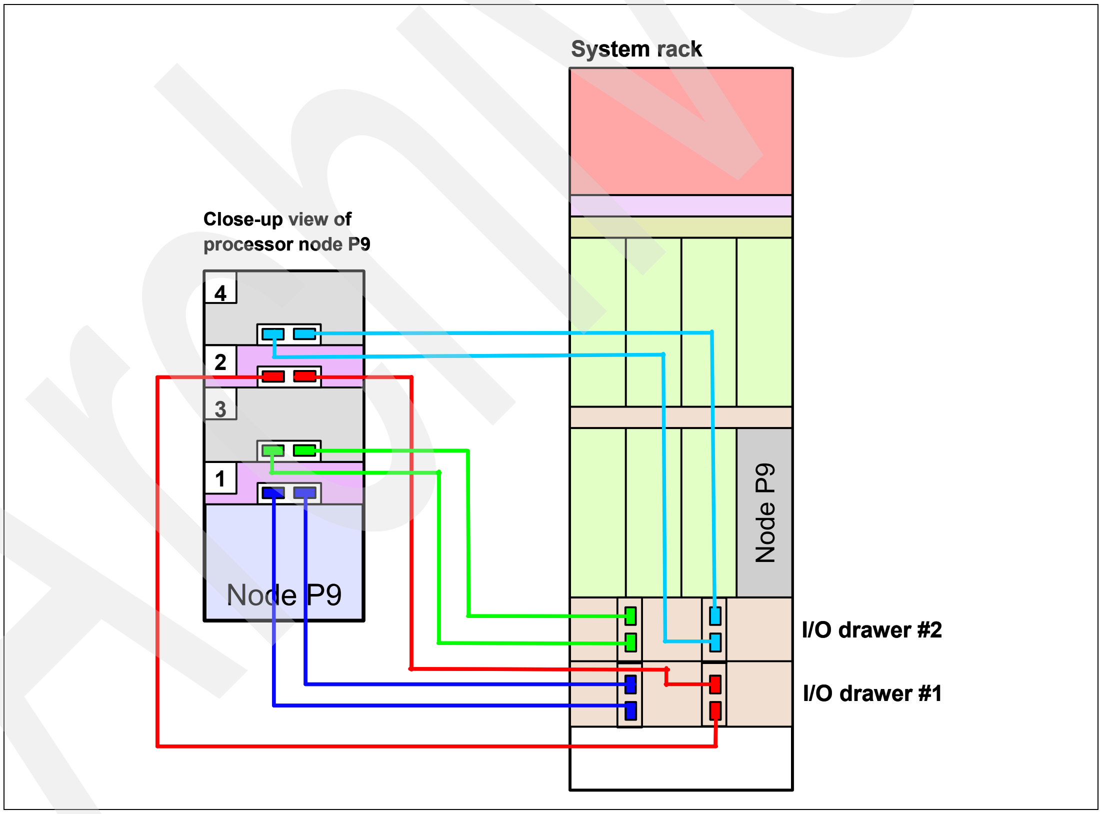


Figure 2-42 Expansion drawer to I/O hub loop cabling

Note: For ongoing management of the 595 system, it is important to keep up-to-date cabling documentation for the internal I/O drawers. Depending on when processor books and I/O hubs are added to your server, the cabling layout might be different from the standard cabling diagrams provided in the installation guide.

2.9 PCI adapter support

IBM offers PCI and PCI-extended (PCI-X) adapters for the 595 server. All adapters support Enhanced Error Handling (EEH). A PCI, PCI-X, and PCI-X 2.0 adapters can be installed in any available PCI-X pr PCI-X 2.0 slot.

Most of the PCI-X and PCI-X 2.0 adapters for the 595 server are capable of being hot-plugged. Any PCI adapter supporting a boot device or system console should not be hot-plugged. The POWER GXT135P Graphics Accelerator with Digital Support (#2849) is not hot-plug-capable.

System maximum limits for adapters and devices might not provide optimal system performance. These limits are given to help assure connectivity and function.

For complete PCI card placement guidance in a POWER6 configuration, including the system unit and I/O enclosures attached to loops, refer to the *Power Systems PCI Adapter Placement Guide for Machine Type 820x and 91xx,* SA76-0090, which is available at:

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp

Select:

IBM Systems Hardware Information Center → **Power Systems information** → **9119-FHA (IBM Power 595)** → **PDF files for the 9119-FHA (IBM Power 595)**

Before adding or rearranging adapters, use the IBM System Planning Tool to validate the new adapter configuration. See the IBM System Planning Tool Web site at:

http://www-03.ibm.com/servers/eserver/support/tools/systemplanningtool/

If you are installing a new feature, ensure that you have the software required to support the new feature and determine whether there are any existing PTF prerequisites to install. To do this, use the IBM Prerequisite Web site at:

http://www-912.ibm.com/e_dir/eServerPrereq.nsf

In the feature summary tables of this section, note the following information:

- Some adapters exclusively supported with the IBM i operating system might require an IOP adapter to function properly.
- Over time, additional features might be withdrawn from marketing. This information is typically included in announcement letters. To ensure you have the latest information about ordering a feature, go to the Support for IBM Systems Web site. at:

  http://www.ibm.com/systems/support

  On the navigation panel, select the technology (Power in our case). Then, select the product (Hardware), and machine type and then a Planning tab. For example:

  a. http://www.ibm.com/systems/support
  b. Select **Power**.
  c. Select **9119-FHA** from the **Hardware** list.
  d. Click **Go**.
  e. Select **Planning** tab.

  Alternatively, for example for System i models, go to:

  http://www.ibm.com/systems/support/i/planning/upgrade/index.html

Note: When using the IBM i operating system, some adapters might require feature #2844 PCI IOP adapter card. These adapters are note by ✓^{iop} in the support column under the IBM i heading.

2.9.1 LAN adapters

Table 2-23 lists the LAN adapters that are available for the 595 server.

Table 2-23 Available LAN adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#5700	Gigabit Ethernet-SX PCI-X Adapter	Short	640	✓	✓	✓
#5701	10/100/1000 Base-TX Ethernet PCI-X Adapter	Short	640	✓	✓	✓
#5706	2-Port 10/100/1000 Base-TX Ethernet PCI-X Adapter	Short	640	✓	✓	✓
#5707	2-Port Gigabit Ethernet-SX PCI-X Adapter	Short	640	✓	✓	✓
#5721	10 Gigabit Ethernet-SR PCI-X (Fiber)	Short	448	✓	✓	✓
#5722	10 Gigabit Ethernet-LR PCI-X (Fiber)	Short	448	✓	✓	✓
#5740	4-port 10/100/1000 Gigabit Ethernet PCI-X	Short	640	✓	—	✓

Note: AIX Network Install Manager (NIM) boot capability is supported with adapter feature codes #5700, #5701, and #5707.

2.9.2 SCSI adapters

Table 2-24 lists the SCSI adapters that are available for the 595 server.

Table 2-24 Available SCSI adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#5583	#5777 Controller w/AUX Write Cache	Long	288	—	✓	—
#5736	PCI-X Dual Channel Ultra320 SCSI Adapter	Long	128	✓	✓	✓
#5776	PCI-X Disk Controller-90 MB No IOP	Long	192	—	✓	—
#5777	PCI-X Disk Controller-1.5 GB No IOP	Long	288	—	✓	—
#5778	PCI-X EXP24 Ctl-1.5 GB No IOP	Long	192	—	✓	—
#5780	PCI-X EXP24 Ctl-1.5 GB No IOP	Long	256	—	✓	—
#5782	PCI-X EXP24 Ctl-1.5 GB No IOP	Long	384	—	✓	—
#5806	PCI-X DDR Dual Channel Ultra320 SCSI Adapter	Long	168	—	✓^{iop}	—

2.9.3 iSCSI

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

For more information about this standard, see:

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

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

The IBM iSCSI adapters in the 595 server offer the advantage of increased bandwidth through the hardware support of the iSCSI protocol. The 1 Gigabit iSCSI TOE PCI-X adapters support hardware encapsulation of SCSI commands and data into TCP and transport it over the Ethernet using IP packets. The adapter operates as an iSCSI TCP/IP Offload Engine. This offload function eliminates host protocol processing and reduces CPU interrupts. The adapter uses a small form factor LC type fiber optic connector or copper RJ45 connector.

Table 2-25 lists the iSCSI adapters that are available for the 595 server.

Table 2-25 Available iSCSI adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#5713	Gigabit iSCSI TOE PCI-X on copper media adapter	Short	48	✓	✓	✓
#5714	Gigabit iSCSI TOE PCI-X on optical media adapter	Short	48	✓	✓	✓

IBM iSCSI software Host Support Kit

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

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

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

2.9.4 SAS adapters

Serial Attached SCSI (SAS) is a new interface that provides enhancements over parallel SCSI with its point-to-point high frequency connections. SAS physical links are a set of four wires used as two differential signal pairs. One differential signal transmits in one direction while the other differential signal transmits in the opposite direction. Data can be transmitted in both directions simultaneously. Table 2-26 lists the SAS adapters that are available for the 595 server.

Table 2-26 Available SAS adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#5902[a]	PCI-X DDR Dual - x4 3 Gb SAS RAID Adapter	Long	192	✓	—	✓
#5912	PCI-X DDR Dual - x4 SAS Adapter	Short	192	✓	—	✓

[a] The SAS RAID adapter must be installed in pairs. The SAS RAID adapter cannot be used to drive. tape, and DVD media devices.

2.9.5 Fibre Channel adapters

The 595 server supports direct or SAN connection to devices using Fibre Channel adapters. 4 Gbps Fibre Channel adapters are available in either single-port or dual-port configuration.

All of these adapters have LC connectors. If you are attaching a device or switch with an SC type fiber connector, an LC-SC 50 Micron Fiber Converter Cable (FC 2456) or an LC-SC 62.5 Micron Fiber Converter Cable (FC 2459) is required.

Supported data rates between the server and the attached device or switch are as follows: Distances of up to 500 meters running at 1 Gbps, distances up to 300 meters running at 2 Gbps data rate, and distances up to 150 meters running at 4 Gbps. When these adapters are used with IBM supported Fibre Channel storage switches supporting long-wave optics, distances of up to 10 kilometers are capable running at either 1 Gbps, 2 Gbps, or 4 Gbps data rates.

Table 2-27 summarizes the Fibre Channel adapters that are available for the 595 server.

Table 2-27 Available Fibre Channel adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#5749	4 Gbps Fibre Channel (2-port)	Short	512	—	✓	—
#5758	4 Gigabit single-port Fibre Channel PCI-X 2.0 Adapter (LC)	Short	512	✓	✓	✓
#5759	4 Gigabit dual-port Fibre Channel PCI-X 2.0 Adapter (LC)	Short	512	✓	—	✓
#5760	PCI-X Fibre Channel Disk Controller	Short	512	—	✓^{iop}	—
#5761	PCI-X Fibre Channel Tape Controller	Short	512	—	✓^{iop}	—

2.9.6 Asynchronous, WAN, and modem adapters

The asynchronous PCI-X adapters provide connection of asynchronous EIA-232 or RS-422 devices. The PCI 2-Line WAN IOA has two ports that provide support for V.21, V.24/EIA232, V.35 and V.36 communication protocols. The PCI 4-Modem WAN IOA provides four RJ-11 modem ports and the PCI 2-Line WAN w/Modem provides one RJ-11 modem port and one

WAN port that provides support for V.21, V.24/EIA232, V.35 and V.36 communication protocols. Table 2-28 lists the asynchronous, WAN, and modem adapters that are available for the 595 server.

Table 2-28 Available Asynchronous, WAN, and modem adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#2943	8-Port Asynchronous Adapter EIA-232/RS-422	Short	18	✓	—	—
#5723	2-Port Asynchronous EIA-232 PCI Adapter	Short	16	✓	—	✓
#6805	PCI 2-Line WAN IOA No IOP	Short	199	—	✓	—
#6808	PCI 4-Modem WAN IOA No IOP	Long	99	—	✓	—
#6833	PCI 2-Line WAN w/Modem IOA No IOP	Short	239	—	✓	—

2.9.7 PCI-X Cryptographic Coprocessor

The PCI-X Cryptographic Coprocessor (FIPS 4) (FC 4764) for selected servers provides both cryptographic coprocessor and secure-key cryptographic accelerator functions in a single PCI-X card. The coprocessor functions are targeted to banking and finance applications. Financial PIN processing and credit card functions are provided. EMV is a standard for integrated chip-based credit cards. The secure-key accelerator functions are targeted at improving the performance of Secure Sockets Layer (SSL) transactions. FC 4764 provides the security and performance required to support on demand business and the emerging digital signature application.

The PCI-X Cryptographic Coprocessor (FIPS 4) (FC 4764) for selected servers provides secure storage of cryptographic keys in a tamper resistant hardware security module (HSM), that is designed to meet FIPS 140 security requirements. FIPS 140 is a U.S. Government National Institute of Standards and Technology (NIST)-administered standard and certification program for cryptographic modules. The firmware for the FC 4764 is available on a separately ordered and distributed CD. This firmware is an LPO product: 5733-CY1 Cryptographic Device Manager. The FC 4764 also requires LPP 5722-AC3 Cryptographic Access Provider to enable data encryption.

Table 2-29 lists the cryptographic adapter that is available for the 595 server.

Table 2-29 Available cryptographic adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#4764	PCI-X Cryptographic Coprocessor (FIPS 4)	Short	32	✓	✓	—

Note: This feature has country-specific usage. Refer to your IBM marketing representative for availability or restrictions.

2.9.8 IOP adapter

The PCI IOP adapter is used to drive PCI IOA adapters located in expansion drawers, units, and towers. PCI IOP and PCI IOA adapters are used exclusively by the IBM i operating system. Each PCI IOP adapter can support up to four PCI IOA adapters. Table 2-30 lists the IOP adapter that is available for the 595 server.

Table 2-30 Available IOP adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#2844	PCI IOP	Short	182	—	✓	—

2.9.9 RIO-2 PCI adapter

The RIO-2 Remote I/O loop adapter is used to attach #5790 PCI drawer to the 595 server. This feature provides two RIO-2 ports and supports a total of one loop per adapter. Up to four #5790 PCI drawers can be configured in a single loop. Additional information about this adapter is provided in Table 2-31.

Table 2-31 Available RIO-2 PCI adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#6438	RIO-2 Remote I/O Loop Adapter	Short	128	—	✓	—

2.9.10 USB and graphics adapters

The 2-Port USB PCI adapter is available for the connection of a keyboard and a mouse. The POWER GXT135P is a 2-D graphics adapter that provides support for analog and digital monitors. Table 2-32 lists the available USB and graphics adapters.

Table 2-32 USB and Graphics adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
#2738	2-Port USB PCI Adapter	Short	16	✓	—	✓
#2849	POWER GXT135P Graphics Accelerator with Digital Support	Short	8	✓	—	✓

2.10 Internal storage

A variety of SCSI and SAS disk drives are available for installation in the expansion drawers, units, and towers. Table 2-33 on page 98 lists the disk drives that are available for the 595 server.

Table 2-33 Disk drive options

Feature code	Description	Supported I/O drawer(s)	Support		
			AIX	IBM i	Linux
#3279[a]	146.8 GB 15 K RPM Ultra320 SCSI Disk Drive Assembly	#5786, #5791, #5797. #5798	✓	—	✓
#3646	73 GB 15 K RPM SAS Disk Drive	#5886	✓	—	✓
#3647	146 GB 15 K RPM SAS Disk Drive	#5886	✓	—	✓
#3648	300 GB 15 K RPM SAS Disk Drive	#5886	✓	—	✓
#3676	69.7 GB 15 K RPM SAS Disk Drive	#5886	—	✓	—
#3677	139.5 GB 15 K RPM SAS Disk Drive	#5886	—	✓	—
#3678	283.7 GB 15 K RPM SAS Disk Drive	#5886	—	✓	—
#4328	141.12 GB 15 K RPM Disk Unit (SCSI)	#5786	—	✓	—

[a] #5786 is supported only with IBM i. #5791 is supported only with AIX and Linux for Power. For IBM i, use 4328.

2.11 Media drawers

Tape and DVD support is provided though the use of a media drawer. As listed in Table 2-34, two media drawers are available for the 595 server. Only one #5720 is supported per 595 system.

Table 2-34 Media drawers

Feature	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
#7214-1U2	Media Drawer, 19-inch	1	✓	✓	✓
#5720[a]	DVD/Tape SAS External Storage Unit	1	✓	—	✓

[a] #5720 DVD/Tape SAS External Storage Unit is not available when the battery backup feature is ordered.

A DVD media device must be available to perform OS installation, maintenance, problem determination, and service actions such as maintaining system firmware and I/O microcode. Certain configurations can utilize an AIX NIM (Network Install Manager) server. The installation and use of an AIX NIM server is a client responsibility. All Linux for Power only systems must have a DVD media device available. The available tape and media devices are listed in Table 2-35

Table 2-35 Available tape and DVD media devices

Feature	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
#5619	80/160 GB DAT160 SAS Tape Drive	1	✓	—	✓

Feature	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
#5746	Half High 800 GB / 1.6 TB LTO4 SAS Tape Drive	1	✓	—	✓
#5756	Slimline DVD-ROM Drive	2	✓	✓	✓
#5757	IBM 4.7 GB Slimline DVD-RAM Drive	2	✓	✓	✓

2.11.1 Media drawer, 19-inch (7214-1U2)

The Media Drawer, 19-inch is a rack mounted media drawer with two media bays as shown in figure Figure 2-43

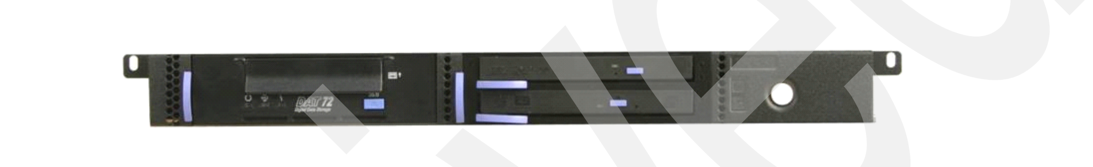


Figure 2-43 19" Media Drawer, 19-inch (7214-1U2)

The first bay in the 7214-1U2 media drawer supports a tape drive. The second bay can support either a tape device, or two slim-line DVD devices. Media devices and is connected by a single SAS controller (#5912) that drives all of the devices in the media drawer, or other drawers. The SAS controller must be placed into a PCI Expansion Drawer (#5790). The 7214-1U2 media drawer must be mounted in a 19-inch rack w/1U of available space. Up to two SAS tape drives can be configured in this media drawer.

2.11.2 DVD/Tape SAS External Storage Unit (#5720)

The DVD/Tape SAS External Storage is a 24-inch rack mounted media drawer with two media bays as shown in figure Figure 2-44.

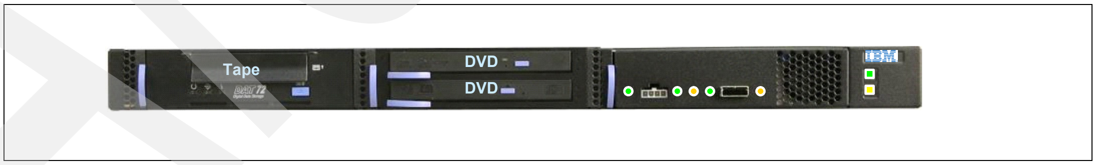


Figure 2-44 DVD/Tape SAS External Storage Unit (#5720)

Each bay supports a tape drive or DVD (one or two) media device. A maximum of one tape drive and two DVD devices is allowed per drawer. The media drawer is connected using a single SAS controller (#5912) that drives all of the devices in the media drawer. The SAS controller must be placed into an internal I/O drawer (#5791, #5797, or #5798).

The #5720 media drawer is mounted in the systems (CEC) rack at position U12 or U34. Rack location U34 is the default. If it is mounted in the U12 location (below the processor books), one I/O drawer position in the system rack will be eliminated. To move the drawer to U12 during the system build, order #8449.

2.12 External I/O enclosures

The 595 server supports three external I/O enclosures that are mounted in 19-inch racks. These enclosures are listed in Table 2-36.

Table 2-36 External I/O enclosures

Feature	Description	Maximum per system	Support		
			AIX	IBM i	Linux
#5786	TotalStorage® EXP24 Disk Drawer	110	—	✓	—
#5790	PCI Expansion Drawer	96	—	✓	—
#5886	EXP 12S Expansion Drawer	12	✓	✓	✓

2.12.1 TotalStorage EXP24 Disk Dwr (#5786)

The TotalStorage EXP24 Disk Dwr provides 24 disk bays. The EXP24 requires 4U of mounting space in a 19” rack and features redundant power, redundant cooling. The EXP24 uses Ultra™ 320 SCSI drive interface connections. The 24 disk bays are organized into four independent groups of six drive bays.

Disk groups are enabled using Ultra 320 SCSI repeater cards (#5741/#5742) that are connected to an Ultra 320 SCSI adapters. Up to four repeater cards are supported per #5786 unit.

Repeater card #5741 can be used to connect one group of six disk drives to a SCSI initiator. Repeater card #5742 can be used to connect one group of six drives to one or two different SCSI initiators, or it can be used to connect two groups of six drives to one SCSI initiator.

2.12.2 PCI Expansion Drawer (#5790)

The PCI Expansion Drawer (#5790) provides six full-length, 64-bit, 3.3-V 133 MHz hot-plug PCI slots. PCI cards are mounted in blind swap cassettes that allow adapter cards to be added/removed without opening the drawer. The PCI Expansion Drawer is connected using the RIO-2 interface adapter (#6438). Redundant and concurrently maintainable power and cooling is included.

The #5790 drawer mounts in a 19” rack using a Dual I/O unit Enclosure (#7307). Each Dual I/O Enclosure supports one or two #5790 PCI Expansion drawers.

> **Note:** The PCI Expansion Drawer (#5790) has a planned availability date of September 9, 2008.

2.12.3 EXP 12S Expansion Drawer (#5886)

The EXP 12S Expansion Drawer (#5886) provides 12 hot-swap SAS storage bays. Redundant power supplies and two Service Managers are included in the drawer. The EXP 12S requires 2U of mounting space in a 19-inch rack. The enclosure attaches to the 595 server using a SAS controller card (#5900 or #5912), or SAS RAID controllers #5902.

2.13 Hardware Management Console (HMC)

The Hardware Management Console (HMC) is a dedicated workstation that provides a graphical user interface for configuring, operating, and performing basic system tasks for POWER6 processor-based (as well as POWER5 and POWER5+ processor-based) systems that function in either non-partitioned, partitioned, or clustered environments. In addition the HMC is used to configure and manage partitions. One HMC is capable of controlling multiple POWER5, POWER5+, and POWER6 processor-based systems.

- The HMC 7042 Model C06 is a deskside model with one integrated 10/100/1000 Mbps Ethernet port, and two additional PCI slots.
- The HMC 7042 Model CR4 is a 1U, 19-inch rack-mountable drawer that has two native 10/100/1000 Mbps Ethernet ports and two additional PCI slots.

> **Note:** When you order the IBM 2-Port 10/100/1000 Base-TX Ethernet PCI-X Adapter (FC 5767, or the older adapter FC 5706 if existing), consider ordering the HMC to provide additional physical Ethernet connections.

At the time of writing, one HMC supports up to 48 POWER5, POWER5+ and POWER6 processor-based systems and up to 254 LPARs using the HMC machine code Version 7.3. For updates of the machine code and HMC functions and hardware prerequisites, refer to the following Web site:

https://www14.software.ibm.com/webapp/set2/sas/f/hmc/home.html

The 595 server requires Ethernet connectivity between the HMC and the bulk power hub (BPH) as shown in Figure 2-45.

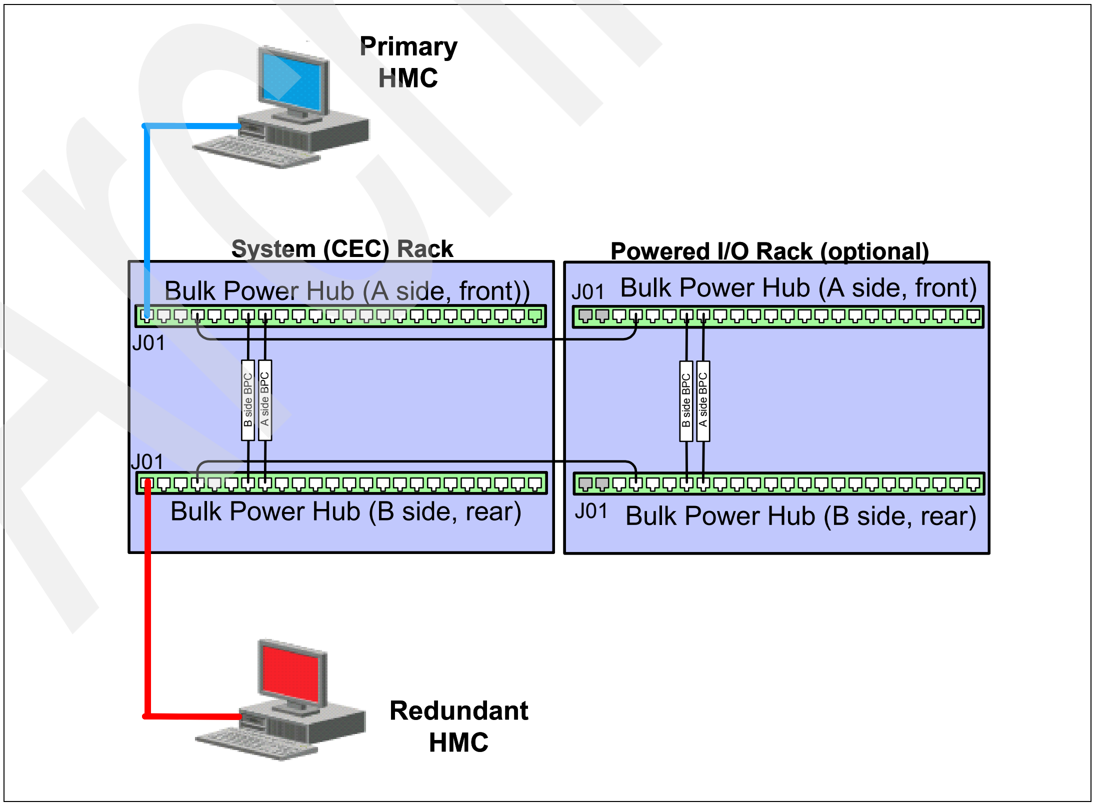

Primary HMC
System (CEC) Rack
Bulk Power Hub (A side, front))
J01
B side BPC
A side BPC
J01
Bulk Power Hub (B side, rear)
Powered I/O Rack (optional)
J01
Bulk Power Hub (A side, front)
B side BPC
A side BPC
J01
Bulk Power Hub (B side, rear)
Redundant HMC

Figure 2-45 HMC connections to the 595 server

The 595 server requires one primary HMC that can communicate to all BPHs in the system. The primary HMC is connected to port J01 on the BPH located on the front side of the system (CEC) rack.

On a systems that support mission-critical workloads, for improved system availability, a redundant HMC is highly recommended. The redundant HMC is connected to port J01 on the BPH located on the back side of the system rack. It is common to make use of an Ethernet hub or switch to facilitate the connections between the The HMC and the 595 server.

The default mechanism for allocation of the IP addresses for the 595 server HMC ports is dynamic. The HMC must be configured as a DHCP server, which provides an IP address to the system controller (service processor) when the server is powered on.

For dynamic LPAR operations and system performance reporting, it is recommended that all AIX and Linux for Power partitions be configured to communicate over an administrative or public network that is shared with the HMC. Figure 2-46 shows a simple network configuration that enables communications between the HMC and server LPARs for the enablement of Dynamic LPAR operations. For more details about HMC and the possible network connections, refer to the *Hardware Management Console V7 Handbook*, SG24-7491.

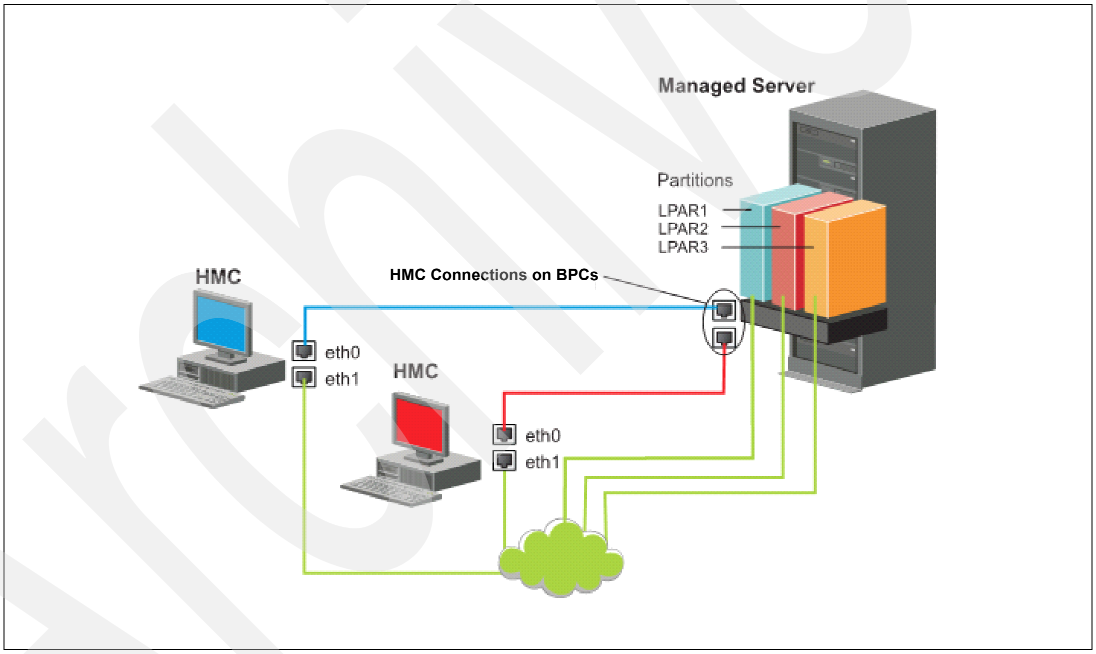


Figure 2-46 HMC and LPAR public network

For any logical partition (dedicated or shared processor pool-based) in a server, it is possible to configure a Shared Ethernet Adapter using the Virtual I/O Server to satisfy the public network requirement. This can help to reduce the quantity of physical adapters required to communicate with the HMC.

2.13.1 Determining the HMC serial number

For some HMC or service processor troubleshooting situations, an IBM service representative must log into the HMC. The service password changes daily and is not generally available for client use. If an IBM project engineer (PE) determines a local service engineer can sign onto the HMC, the service representative might request the HMC serial number.

To find the HMC serial number, open a restricted shell window and run the following command:

```
#lshmc -v
```

2.14 Advanced System Management Interface

The Advanced System Management Interface (ASMI) is the interface to the service processor that is required to perform general and administrator-level service tasks, such as reading service processor error logs, reading vital product data, setting up the service processor, and controlling the system power. The ASMI can also be referred to as the service processor menus. The ASMI can be accessed in one of the following ways:

- Web browser
- ASCII console
- HMC

Often, you might use the service processor's default settings, in which case, accessing the ASMI is not necessary.

2.14.1 Accessing the ASMI using a Web browser

The Web interface to the ASMI is available during all phases of system operation including the initial program load (IPL) and run time. To access the ASMI menus using a Web browser, first connect a PC or mobile computer to the server. Then, using an Ethernet cable, connect your computer to one of the service ports labeled J02, on either the front or back side Bulk Power Hub.

The Web interface to the ASMI is accessible through Microsoft® Internet Explorer® 6.0, Netscape 7.1, or Opera 7.23 running on a PC or mobile computer connected one of the two HMC ports. The Web interface is available during all phases of system operation, including the initial program load and run time. However, some of the menu options in the Web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The default ASMI IP addresses are listed in Table 2-37.

Table 2-37 ASMI default IP addresses

System Controller Port	Subnet mask	IP address
J01 on Bulk Power Hub (front)	255.255.255.0	169.254.2.147
J02 on Bulk Power Hub (rear)	255.255.255.0	169.254.3.147

The default ASMI user IDs and passwords are listed in Table 2-38.

Table 2-38 Default user IDs and passwords for the ASMI web interface

User ID	Password
general	general
admin	admin

2.14.2 Accessing the ASMI using an ASCII console

The ASMI on an ASCII console supports a subset of the functions provided by the Web interface and is available only when the system is in the platform standby state. The ASMI on an ASCII console is not available during some phases of system operation, such as the initial program load and run time. More information about how to access the ASMI menus using an ASCII console can be found in the *Service Guide for the 9119-FHA,* SA76-0162, located at:

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp

Search for 9119-FHA, select **PDF files for the 9119-FHA (IBM Power 595)**.

2.14.3 Accessing the ASMI using an HMC

To access the ASMI using the Hardware Management Console (HMC):

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

System Management Services

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

To start SMS, perform the following steps:

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

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

   For a full server partition, watch the firmware console.
3. Look for the power-on self-test (POST) indicators: memory, keyboard, network, SCSI, and speaker that appear across the bottom of the window. Press the numeric 1 key after the word `keyboard` appears and before the word `speaker` appears.

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

For more information about usage of SMS, refer to the IBM Systems Hardware Information Center:

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

2.14.4 Server firmware

Server firmware is the part of the Licensed Internal Code that enables hardware, such as the service processor. Check for available server firmware fixes regularly. Depending on your service environment, you may download your server firmware fixes using different interfaces and methods. The 595 servers must use the HMC to install server firmware fixes. Firmware is loaded on to the server and to the bulk power controller over the HMC to the frame's Ethernet network.

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

Server firmware image

The server firmware binary image is a single image that includes code for the service processor, the POWER Hypervisor™ firmware, and platform partition firmware. This server firmware binary image is stored in the service processor's flash memory and executed in the service processor main memory.

Because each Central Electronics Complex (CEC) has dual service processors (system controllers), both service processors are updated when firmware updates are applied and activated using the Licensed Internal Code Updates section of the HMC.

Firmware is available for download at:

http://www14.software.ibm.com/webapp/set2/firmware/gjsn

Power subsystem firmware

Power subsystem firmware is the part of the Licensed Internal Code that enables the power subsystem hardware in the 595 server. You must use an HMC to update or upgrade power subsystem firmware fixes.

The bulk power controller (BPC) has its own service processor. The power firmware not only has the code load for the BPC service processor itself, but it also has the code for the distributed converter assemblies (DCAs), bulk power regulators (BPRs), fans, and other more granular field replaceable units (FRUs) that have firmware to help manage the frame and its power and cooling controls. The BPC service processor code load also has the firmware for the cluster switches that can be installed in the frame.

In the same way that the CEC has dual service processors, the power subsystem has dual BPCs. Both are updated when firmware changes are made using the Licensed Internal Code Updates section of the HMC.

The BPC initialization sequence after the reboot is unique. The BPC service processor must check the code levels of all the power components it manages, including DCAs, BPRs, fans, cluster switches, and it must load those if they are different than what is in the active flash side of the BPC. Code is cascaded to the downstream power components over universal power interface controller (UPIC) cables.

Platform initial program load

The main function of the 595 system controller is to initiate platform initial program load (IPL), also referred to as *platform boot*. The service processor has a self-initialization procedure and then initiates a sequence of initializing and configuring many components on the CEC backplane.

The service processor has various functional states, which can be queried and reported to the POWER Hypervisor component. Service processor states include, but are not limited to, standby, reset, power up, power down, and run time. As part of the IPL process, the primary service processor checks the state of the backup. The primary service processor is responsible for reporting the condition of the backup service processor to the POWER Hypervisor component. The primary service processor waits for the backup service processor to indicate that it is ready to continue with the IPL (for a finite time duration). If the backup service processor fails to initialize in a timely fashion, the primary service processor reports the backup service processor as a non-functional device to the POWER Hypervisor component and marks it as a *guarded* resource before continuing with the IPL. The backup service processor can later be integrated into the system.

Open Firmware

The 595 server has one instance of Open Firmware, both when in the partitioned environment and when running as a full system partition. Open Firmware has access to all devices and data in the system. Open Firmware is started when the system goes through a power-on reset. Open Firmware, which runs in addition to the Hypervisor firmware in a partitioned environment, runs in two modes: global and partition. Each mode of Open Firmware shares the same firmware binary that is stored in the flash memory.

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

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

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

http://www.redbooks.ibm.com/redpieces/abstracts/SG247039.html?Open

Temporary and permanent sides of the service processor

The service processor and the BPC maintain two copies of the firmware:

- One copy is considered the *permanent* or *backup* copy and is stored on the permanent side, sometimes referred to as the *p* side.
- The other copy is considered the *installed* or *temporary* copy and is stored on the temporary side, sometimes referred to as the *t* side. Start and run the server from the temporary side.
- The copy actually booted from is called the *activated level*, sometimes referred to as *b*.

The concept of *sides* is an abstraction. The firmware is located in flash memory, and pointers in NVRAM determine which is *p* and *t*.

Firmware levels

The levels of firmware are:

- *Installed Level* indicates the level of firmware that has been installed and is installed in memory after the managed system is powered off, and powered on using the default temporary side.
- *Activated Level* indicates the level of firmware that is active and running in memory.
- *Accepted Level* indicates the backup level (or permanent side) of firmware. You can return to the backup level of firmware if you decide to remove the installed level.

Server firmware fix

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

Note: The following points are of special interest:

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

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

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

Choosing which firmware to use when powering on the system is done using the Power-On Parameters tab in the server properties box.

3

Virtualization

The Power 595 server combined with optional PowerVM technology provides powerful virtualization capabilities that can help you consolidate and simplify your environment. This chapter provides an overview of these virtualization technologies.

3.1 Virtualization feature support

The 595 server and PowerVM provide a variety of virtualization capabilities. Most of them are available to the AIX, IBM i, and Linux for Power operating systems. Several virtualization features are operating system-specific. Table 3-1 lists AIX, IBM i and Linux for Power operating system support for each of the virtualization capabilities that will be discussed in this chapter. Note that these capabilities are either a standard component of the 595 server or they are provided through the optional PowerVM offering.

Table 3-1 Virtualization feature support

Feature	Source	AIX V5.3	AIX V6.1	IBM i 5.4 w/LIC 5.4.5	IBM i 6.1	RHEL V4.5 for Power	RHEL V5.1 for Power	SLES V10 SP1 for Power
Hypervisor	Standard	✓	✓	✓	✓	✓	✓	✓
Capacity Upgrade on Demand	Standard	✓	✓	✓	✓	✓	✓	✓
Dynamic LPARs[a]	Standard	✓	✓	✓	✓	✓	✓	✓
Share Processor Pool LPARs	PowerVM	✓	✓	✓	✓	✓	✓	✓
Shared Dedicated Processor Capacity	PowerVM	✓	✓	—	✓	—	✓	✓
Multiple Shared Processor Pools	PowerVM	✓	✓	—	✓	—	✓	✓

Feature	Source	AIX V5.3	AIX V6.1	IBM i 5.4 w/LIC 5.4.5	IBM i 6.1	RHEL V4.5 for Power	RHEL V5.1 for Power	SLES V10 SP1 for Power
Virtual Ethernet	Standard	✓	✓	✓ as a server	✓ as a server, client	✓	✓	✓
Virtual I/O Server	PowerVM	✓	✓	—	✓ as a client	✓	✓	✓
Virtual SCSI	PowerVM	✓	✓	✓ as server	✓ as a server, client	✓	✓	✓
Shared Ethernet Adapter	PowerVM	✓	✓	✓ as a server	✓ as a server, client	✓	✓	✓
Live Partition Mobility[b]	PowerVM	✓	✓	—	✓	—	✓	✓

[a] Dynamic memory removal is not supported by Linux at the time of writing.

[b] Requires PowerVM Enterprise Edition.

3.2 PowerVM and PowerVM editions

PowerVM is a combination of hardware feature enablement and value-added software that provides additional virtualization functions beyond the standard 595 server virtualization capabilities. PowerVM enables:

- Shared processor pool LPARs (based on Micro-Partitioning technology)
- Shared dedicated capacity
- Multiple shared processor pools
- Virtual I/O Server (virtual SCSI and shared Ethernet adapter)
- Lx86
- Live Partition Mobility

PowerVM is an enhanced version of the former virtualization product called Advanced Power Virtualization (APV).

PowerVM is available in both Standard and Enterprise Editions for the 595 server. Table 3-2, lists the virtualization features that are enabled (provided) with each of these PowerVM editions.

Table 3-2 PowerVM features by edition

Virtualization Feature	PowerVM Standard Edition (#8506)	PowerVM Enterprise Edition (#8507)
Shared Processor Pool LPARs	✓	✓
Shared Dedicated Capacity	✓	✓
Multiple Shared-Processor Pools	✓	✓
Virtual I/O Server (Virtual SCSI, Shared Ethernet Adapter)	✓	✓

Virtualization Feature	PowerVM Standard Edition (#8506)	PowerVM Enterprise Edition (#8507)
PowerVM Lx86	✓	✓
PowerVM Live Partition Mobility	—	✓

PowerVM Standard Edition can be upgraded to PowerVM Enterprise Edition by providing a key code, which is entered at the HMC. This upgrade operation is nondisruptive.

Additional information about the different PowerVM Editions can be found at the IBM PowerVM Editions Web site:

http://www-03.ibm.com/systems/power/software/virtualization/editions/index.html

Detailed information about the use of PowerVM technology can be found in the following IBM Redbooks:

- *PowerVM Virtualization on IBM System p: Introduction and Configuration Fourth Edition*, SG24-7940
- *PowerVM Virtualization on IBM System p: Managing and Monitoring*, SG24-7590

The IBM Redbooks Web site is located at:

http://www.ibm.com/redbooks

3.3 Capacity on Demand

The 595 server can be configured with inactive processors and memory, which can be activated at a later date when they are needed. This feature is called Capacity on Demand (CoD). The options for activating CoD resources are listed in Table 3-3.

Table 3-3 CoD options for processors and memory

Option	Processors	Memory
Permanent activation	✓	✓
On/Off	✓	✓
Utility CoD	✓	—
Trial CoD	✓	✓
Capacity Backup (CBU)	✓	—

> **Note:** CoD processors that have not been activated are available to the 595 server for dynamic processor sparing. If the server detects the impending failure of an active processor, it attempts to activate one of the unused CoD processors and add it to the running system configuration. This process does not affect the licensed number of processors for your system.

When processors are temporarily enabled using the On/Off, Utility CoD, Trial CoD, or CBU option, no additional license charges are applicable for the AIX and IBM i operating systems or IBM i 5250 Enterprise Enablement. Different rules might apply for Linux for Power. Consult with your Linux vendor regarding the temporary activation of processors for Linux partitions.

For more information about Capacity on Demand, visit

http://www.ibm.com/systems/power/hardware/cod

3.3.1 Permanent activation

You can permanently activate CoD processors and memory by ordering the appropriate activation feature codes. Minimum activation requirements exist for CoD processors and memory that are ordered as part of 595 server system.

Processors

All processor books available on the 595 are shipped as 8-core CoD offerings with zero processors activated. The minimum number of permanently activated processors required is based on the number of processor books installed as shown in Table 3-4

Table 3-4 Minimum processor activation requirements

Qty of installed processor books	Qty of installed processors	Minimum number of activated processors
1	8	3
2	16	6
3	24	9
4	32	12
5	40	15
6	48	18
7	56	21
8	64	24

Additional processors can be permanently activated in increments of one by ordering the appropriate activation feature code. If more than one processor is to be activated at the same time, order the activation feature code in multiples. After receiving an order for a CoD processor activation feature, IBM will provide you with a 32-character encrypted activation key. Enter this key into the HMC associated with your 595 server. Newly activated processors are automatically put into use as part of the shared processor pool. Newly activated processors can also be dynamically added to dedicated processor LPARs by using a DLPAR operation.

Memory

All 595 server memory is shipped as a CoD offering with zero activations. Minimum activation requirements exist for each memory feature as shown in Table 3-5.

Table 3-5 Minimum memory activation requirements

Memory feature	Description	Minimum activation
5693	0/4 GB DDR2 Memory (4X1 GB)	100%
5694	0/8 GB DDR2 Memory (4X2 GB)	50%
5695	0/16 GB DDR2 Memory (4X4 GB)	50%

Memory feature	Description	Minimum activation
5696	0/32 GB DDR2 Memory (4X 8 GB)	50%
5697	0/64 GB DDR2 Memory(4X16 GB)	100%
8201	0/256 GB 533 MHz DDR2 Memory Package (32x #5694)	100%
8202	0/256 GB 533 MHz DDR2 Memory Package (16x #5695)	100%
8203	0/512 GB 533 MHz DDR2 Memory Package (32x #5695)	100%
8204	0/512 GB 400 MHz DDR2 Memory Package (16x #5696)	100%
8205	0/2 TB 400 MHz DDR2 Memory Package (32x #5697)	100%

3.3.2 On/Off CoD

On/Off Capacity on Demand offers flexibility in meeting peak demands that are temporary in duration. Inactive CoD processors and memory can be temporarily activated with a simple request made by an operator from the HMC. The On/Off CoD usage is measured and billed in one day increments. The one-day timing is based on a 24-hour time period, starting at the time of activation. If the 595 server is shutdown during the 24-hour time period, the *clock* is stopped while the server is powered off, and restarted when the server is powered on.

Implementing On/Off CoD for processors and memory consists of three steps:

1. Enablement
2. Activation
3. Billing

On/Off CoD enablement

Before you can use temporary capacity on a server, you must *enable* it for On/Off CoD. To do this, order the no-charge On/Off processor or memory enablement feature. This initiates the generation of usage and billing contracts that must be signed and returned to IBM. Upon receipt of the signed contracts, IBM generates an enablement code, mails it to you, and posts it on the Web. Enter the enablement code on the managing HMC to enable On/Off functionality.

On/Off CoD activation

When you require On/Off CoD temporary capacity (for a processor or for memory), use the HMC menu for On/Off CoD. Specify how many of the inactive processors or gigabytes of memory you would like temporarily activated for a specified number of days. At the end of the temporary period (days you requested), you must ensure the temporarily activated capacity is available to be reclaimed by the server (not assigned to partitions), or you are billed for any unreturned processor or memory days.

Time limits on enablement codes are:

- The processor enablement code allows you to request a maximum of 360 processor days of temporary capacity. If you have consumed 360 processor days, place an order for another processor enablement code to reset the number of days that you can request back to 360.
- The memory enablement code lets you request up to 999 memory days of temporary capacity. A memory day represents one GB of memory activated for one day. If you have

consumed 999 memory days, you will need to place an order for another memory enablement code to reset the number of days that you can request to 999.

On/Off CoD billing

The On/Off CoD billing contract requires that you report usage data at least once a month to IBM whether or not there is activity. Billing data can be automatically delivered to IBM by the HMC if it is configured with Electronic Service Agent™. You can also manually report billing data through a fax or e-mail.

The billing data is used to determine the correct amount to bill at the end of each billing period (calendar quarter). Your IBM sales representative or IBM Business Partner is notified of temporary capacity usage and will submit an order that corresponds to the amount of On/Off usage. Failure to report temporary processor or memory usage during a billing quarter results in the default billing equivalent of 90 processor days of temporary capacity.

3.3.3 Utility CoD

Utility CoD automatically provides additional processor performance on a temporary basis to the shared processor pool. Utility CoD enables you to assign a quantity of inactive processors to the server's shared processor pool. When the Hypervisor recognizes that the combined processor utilization of the shared pool has exceeded 100% of the original shared pool baseline (activated CPUs), then a Utility CoD Processor Minute is charged and this level of performance is available for the next minute of use. If additional workload requires a higher level of performance, the system will automatically allow additional Utility CoD processors to be used. The system automatically and continuously monitors and charges for the performance needed above the baseline. Registration and usage reporting for Utility CoD is made using a public Web site and payment is based on reported usage.

3.3.4 Trial CoD

Trial CoD is available for 595 servers configured with CoD resources. CoD processors and memory can be temporarily activated using a one-time, no-cost activation for a maximum period of 30 consecutive days. This enhancement allows for benchmarking with the activation of CoD resources or can be used to provide immediate access to standby resources when the purchase of a permanent activation is pending.

If you purchase a permanent processor activation for a 595 server after having made use of the Trial CoD, that server will be eligible for another trial of 30 consecutive days. These subsequent trials are limited to the use of two processors and 4 GB of memory.

Trial CoD is a complimentary service offered by IBM. Although IBM intends to continue it for the foreseeable future, IBM reserves the right to withdraw Trial CoD at any time, with or without notice.

> **Important:** Trial CoD should not be used as part of a disaster recovery plan.

3.3.5 Capacity Backup

595 servers configured with the Capacity Backup (CBU) designation are useful as secondary systems for backup, high availability, and disaster recovery. CBU systems are populated with a high percentage of on demand (CoD) processors and include a specified number of On/Off processor days that can be activated as needed for testing or failover. Capacity BackUp 595

systems allow you to have an off-site, disaster recovery machine at a lower entry price in comparison to a standard 595 server.

The CBU offering includes:

- 4 processors that are permanently activated and can be used for any workload
- 28 or 60 standby processors available for testing or for use in the event of disaster
- 1800 (4/32-core) or 3600 (4/64-core) On/Off CoD processor days

CBU processor resources can be turned on at any time for testing or in the event of a disaster by using the On/Off CoD activation procedure. If you have used the initial allotment of processor days (1800 or 3600), additional processor capacity can be purchased using On/Off CoD activation prices. Standby processors cannot be permanently activated.

Minimum and maximum configuration requirements for memory and I/O are the same as for the standard 595 server offerings. For further information about CBU, visit:

http://www.ibm.com/systems/power/hardware/cbu

3.4 POWER Hypervisor

Combined with features designed into the POWER6 processors, the POWER Hypervisor delivers functions that enable other system technologies, including logical partitioning, virtualized processors, IEEE VLAN compatible virtual switch, virtual SCSI adapters, shared and virtual consoles. The POWER Hypervisor is a basic component of the system's firmware and offers the following functions:

- Provides an abstraction between the physical hardware resources and the logical partitions that use them.
- Enforces partition integrity by providing a security layer between logical partitions.
- Controls the dispatch of virtual processors to physical processors (see 3.5.2, "Shared processor pool partitions" on page 117).
- Saves and restores all processor state information during a logical processor context switch.
- Controls hardware I/O interrupt management facilities for logical partitions.
- Provides virtual LAN channels between logical partitions that help to reduce the need for physical Ethernet adapters for inter-partition communication.
- Monitors the service processor and will perform a reset/reload if it detects the loss of the service processor, notifying the operating system if the problem is not corrected.

The POWER Hypervisor is always active, regardless of the system configuration (LPAR or full system partition). It requires memory to support the resource assignment to the logical partitions on the server. The amount of memory required by the POWER Hypervisor firmware varies according to several factors. The following factors influence the POWER Hypervisor memory requirements:

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

The minimum amount of physical memory to create a partition is the size of the system's Logical Memory Block (LMB). The default LMB size varies according to the amount of memory configured for the system as shown in Table 3-6 on page 116.

Table 3-6 Configurable memory-to-default logical memory block size

Configurable memory	Default logical memory block
Less than 4 GB	16 MB
Greater than 4 GB up to 8 GB	32 MB
Greater than 8 GB up to 16 GB	64 MB
Greater than 16 GB up to 32 GB	128 MB
Greater than 32 GB	256 MB

In most cases, the actual requirements and recommendations are between 256 MB and 512 MB. Physical memory is assigned to partitions in increments of the logical memory block (LMB).

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

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

Virtual SCSI and virtual Ethernet are discussed later in this section.

Virtual (TTY) console

Each partition must have access to a system console. Tasks such as operating system installation, network setup, and some problem analysis activities require access to a system console. The POWER Hypervisor provides the virtual console using a virtual TTY or serial adapter and a set of Hypervisor calls to operate on them. Virtual TTY is a standard feature of the 595 server and hardware management console

3.5 Logical partitioning

Logical partitions (LPARs) provide a level of virtualization which provides a level of abstraction between the physical processors and the operating system. Implementing LPARs can help you increase utilization of system resources and add a new level of configuration possibilities. This section provides details and configuration specifications about the various logical apportioning technologies available for the 595 server.

> **Note:** The terms processor, cpu, and core are used interchangeably in this document.

Detailed information about logical partition definitions and configuration can be found in *PowerVM Virtualization on IBM System p: Introduction and Configuration, Fourth Edition*, SG24-7940, which is on the IBM Redbooks Web site is located at:

http://www.ibm.com/redbooks

3.5.1 Dynamic logical partitions

Logical partitioning technology allows you to divide a single physical system into separate logical partitions (LPARs). Each LPAR receives a user specified allocation of dedicated computing resources, such as:

- Processors (whole CPU increments)
- Memory (16 to 256MB increments, depending on memory configured)
- PCI slots and I/O components (individual slots or components)

The Hypervisor presents a collection of these resources as a *virtualized machine* to the supported operating environment of your choice.

Logical partitioning was introduced with System p POWER4™ servers and the AIX 5L Version 5.1 operating system. Logical partitioning was introduced on System i models prior to POWER4 and IBM i (then known as OS/400) V4.5. LPAR is remains a valid option for configuring LPARs on the 595 server

Later, dynamic logical partitioning (DLPAR) became available, which increased flexibility by allowing system resources (processors, memory, and I/O components) to be dynamically added or removed from the running LPARs. Dynamic reconfiguration of partitions allows system administrators to move system resources between LPARs as necessary to satisfy application performance requirements and to improve utilization of system resources. DLPAR capabilities became available on POWER4 based servers running AIX 5L Version 5.2 and IBM i 5.3. shows to LPAR with dedicated processing resources. Figure 3-1 shows LPAR #1 and LPAR #2 configured as dedicated processor LPARs.

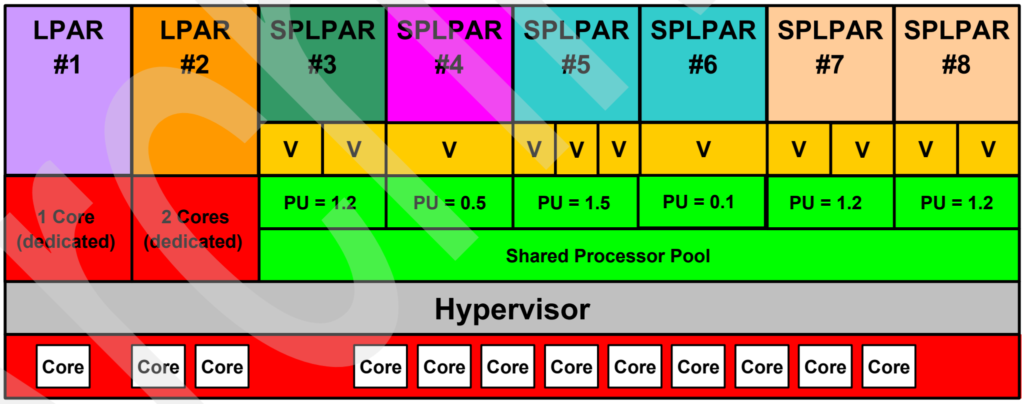


Figure 3-1 Logical partitioning processor allocation options

3.5.2 Shared processor pool partitions

IBM Micro-Partitioning technology (introduced with POWER5 systems) allows you to allocate processors with a finer granularity than with the basic logical partition. Processor allocations can now be specified in hundredths of a cpu and can start as small as one tenth of a cpu. Processing resources for these logical partitions are sourced from the shared processor pool which is comprised of all activated processors that are not currently allocated to running dedicated processor LPARs.

Logical partitions, which source their processing resources from the shared processor pool, are known as shared processor LPARs (SPLPAR). A maximum of ten SPLPARs can be configured per processor or a maximum 254 SPLPARs per 595 server. An important point is that these maximums are supported by the hardware, however the practical limits depend on the application workload demands.

The processing resources allocated to an SPLAR are called processing units (PU). The PU setting represents a guaranteed amount of processing power that is available to a given SPLPAR no matter how busy the shared processor pool is.

If the SPLAR is configured as *uncapped*, it will have the ability to access excess processing units above is specified PU setting. If needed, the PU setting can be changed using a dynamic LPAR operation. The shared processor pool, SPLARs, and PUs are shown in Figure 3-1 on page 117.

> **Note:** The sum total of the PU settings for all running SPLPARs within a 595 server must be less than or equal to the size of the shared processor pool. This ensures that all SPLARs can simultaneously access their guaranteed amount of processing power.

Shared processor pool resources are abstracted to the SPLPARs through virtual processors. Virtual processors (VPs) map the processing power (PUs) assigned to the operating system running on the SPLPAR. VPs are specified in whole numbers.

The number of VPs supported on SPLPAR depends on the number of PUs assigned to that SPLPAR. Each VP can represent from 0.1 to 1 processing units. The minimum number of VPs allowed is equal to the PUs assigned rounded up to the nearest whole number. The maximum number of VPs allowed is equal to the assigned PUs multiplied by ten. The assigned number of VPs sets an upper limit for an SPLPAR ability to consume excess processor units. For example, a SPLPAR with three VPs can access up to three processing units. The number of VPs can be changed dynamically through a dynamic LPAR operation.

> **Note:** Your 595 server is positioned to begin sharing physical processors when the total number of running VPs exceeds the size of the shared processor pool.

Every ten milliseconds, the POWER Hypervisor recalculates each SPLAR's processing needs. If a SPLAR is not busy, it receives an allocation that is smaller than the assigned processing units, making this SPLAR a *donor* of processor cycles to the shared processor pool. If the SPLPAR is busy (and *uncapped*) the Hypervisor can assign excess processor cycles to the SPLPAR up to the number of VP assigned.

3.5.3 Shared dedicated capacity

On POWER6 servers, you can now configure dedicated partitions to donate excess processor cycles to the shared processor pool when their assigned processors are idle. At all times, the dedicated partition receives priority access to its assigned processors. Enabling the shared dedicated capacity feature can help to increase system utilization, without compromising the computing power for critical workloads running in partition with dedicated processors.

3.5.4 Multiple shared processor pools

The original shared processor pool implementation provides support for one shared processor pool. Starting with POWER6, multiple shared processor pools can be defined. SPLARs can now be assigned to run in a specific pool. Each pool has a maximum processing unit (MaxPU) setting that limits the total amount of processing units available to the collection of LPARs running in that pool. A system with three pools configured is shown in Figure 3-2 on page 119.

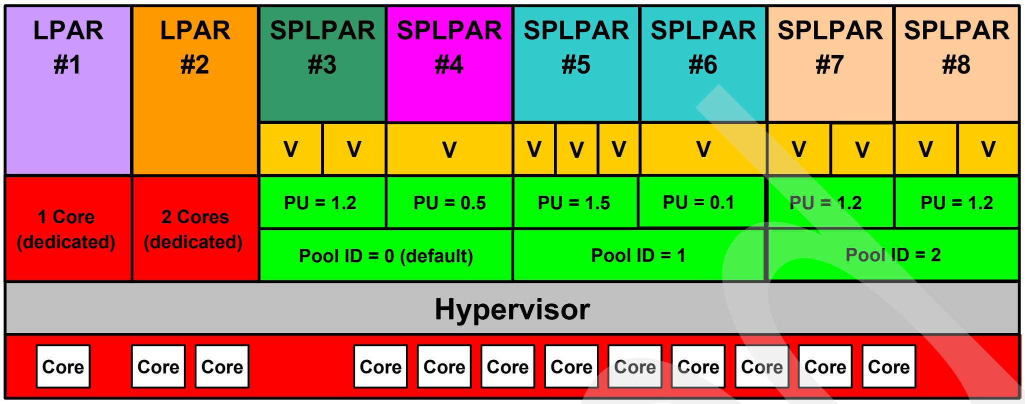


Figure 3-2 Multiple shared processor pools

Multiple shared pools can help you manage expenses for operating systems and ISV software by grouping common applications in the same pool and then limiting the processor resources that can be used by that pool. Use of multiple shared processor pools also provides an additional level of control for processor allocation when system usage is high.

3.6 Virtual Ethernet

Virtual Ethernet enables logical partitions (within a single 595 server) to communicate with each other directly through the POWER Hypervisor and without using a physical Ethernet interface. The virtual Ethernet function is provided by the POWER Hypervisor. The POWER Hypervisor implements the Ethernet transport mechanism as well as an Ethernet switch which supports VLAN capability. Virtual LAN allows secure communication between logical partitions without the need for a physical I/O adapter or cabling. The ability to securely share Ethernet bandwidth across multiple partitions can improve hardware utilization.

The POWER Hypervisor implements an IEEE 802.1Q VLAN style virtual Ethernet switch. Similar to a physical IEEE 802.1Q Ethernet switch it can support tagged and untagged ports. A virtual switch does not really need ports, so the virtual ports correspond directly to virtual Ethernet adapters that can be assigned to partitions from the HMC. There is no need to explicitly attach a virtual Ethernet adapter to a virtual Ethernet switch port. To draw on the analogy of physical Ethernet switches, a virtual Ethernet switch port is configured when you configure the virtual Ethernet adapter on the HMC.

For AIX, a virtual Ethernet adapter is not much different from a physical Ethernet adapter. It can be used:

- To configure an Ethernet interface with an IP address
- To configure VLAN adapters (one per VID)
- As a member of a Network Interface Backup adapter

EtherChannel or Link Aggregation is not applicable to virtual Ethernet LAN adapters on the Power 595 server.

The POWER Hypervisor's virtual Ethernet switch can support virtual Ethernet frame sizes of up to 65408 bytes, which is much larger than what physical switches support: 1522 bytes is standard and 9000 bytes are supported with Gigabit Ethernet Jumbo Frames. Thus, with the POWER Hypervisor's virtual Ethernet, you can increase TCP/IP's MTU size to:

- 65394 (= 65408 - 14 for the header, no CRC) if you do not use VLAN
- 65390 (= 65408 - 14, four for the VLAN, again no CRC) if you use VLAN

Increasing the MTU size could benefit performance because it can improve the efficiency of the transport. This is dependant of the communication data requirements of the running workload.

Note: Virtual Ethernet interfaces can be configured for both dedicated CPU (LPAR) and shared processor pool (SPLPAR) logical partitions.

3.7 Virtual I/O Server

The Virtual I/O Server is a special purpose partition that allows the sharing of physical resources between logical partitions to facilitate improvements in server utilization (for example consolidation). The Virtual I/O Server owns physical resources (SCSI, Fibre Channel, network adapters, and optical devices) and allows client partitions to access them through virtual adapter. This can help to reduce the number of physical adapters in the system.

The Virtual I/O server eliminates the requirement that every partition owns a dedicated network adapter, disk adapter, and disk drive. OpenSSH is supported for secure remote login to the Virtual I/O server. The Virtual I/O server also provides a firewall for limiting access by ports, network services and IP addresses. Figure 3-3 shows an overview of a Virtual I/O Server configuration.

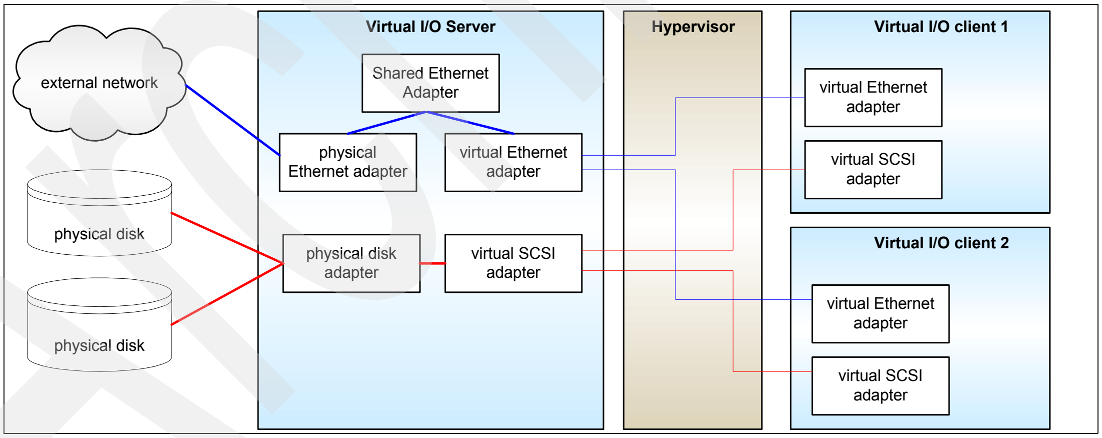


Figure 3-3 Architectural view of the Virtual I/O Server

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

Installation of the Virtual I/O Server partition is performed from a special system backup DVD that is provided to clients with PowerVM Standard and Enterprise editions. This dedicated software is only for the Virtual I/O Server and is only supported in special Virtual I/O Server partitions. Two major functions are provided with the Virtual I/O Server: Virtual SCSI and a Shared Ethernet Adapter.

3.7.1 Virtual SCSI

Virtual SCSI refers to a virtualized implementation of the SCSI protocol. Virtual SCSI is based on a client/server relationship. The Virtual I/O Server logical partition owns physical disk and resources and then acts as server or, in SCSI terms, target device to share these devices with client logical partitions.

Virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured using an HMC. The virtual SCSI server (target) adapter on the Virtual I/O Server is responsible for executing any SCSI commands it receives. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the Virtual I/O Server.

Physical disks presented to the Virtual I/O Server can be exported and assigned to a client partition in a number of different ways:

- The entire disk is presented to the client partition.
- The disk is divided into several logical volumes, which can be presented to a single client or multiple different clients.
- As of Virtual I/O Server 1.5, files can be created on these disks and file backed storage devices can be created.

The Logical volumes or files associate with a single SCS or Fibre Channel adapter can be assigned to different client partitions. Therefore, virtual SCSI enables sharing of adapters and also disk devices.

Figure 3-4 shows an example where one physical disk is divided into two logical volumes by the Virtual I/O Server. Each of the two client partitions is assigned one logical volume, which is then accessed through a virtual I/O adapter (VSCSI Client Adapter). Inside the partition, the disk is seen as a normal hdisk.

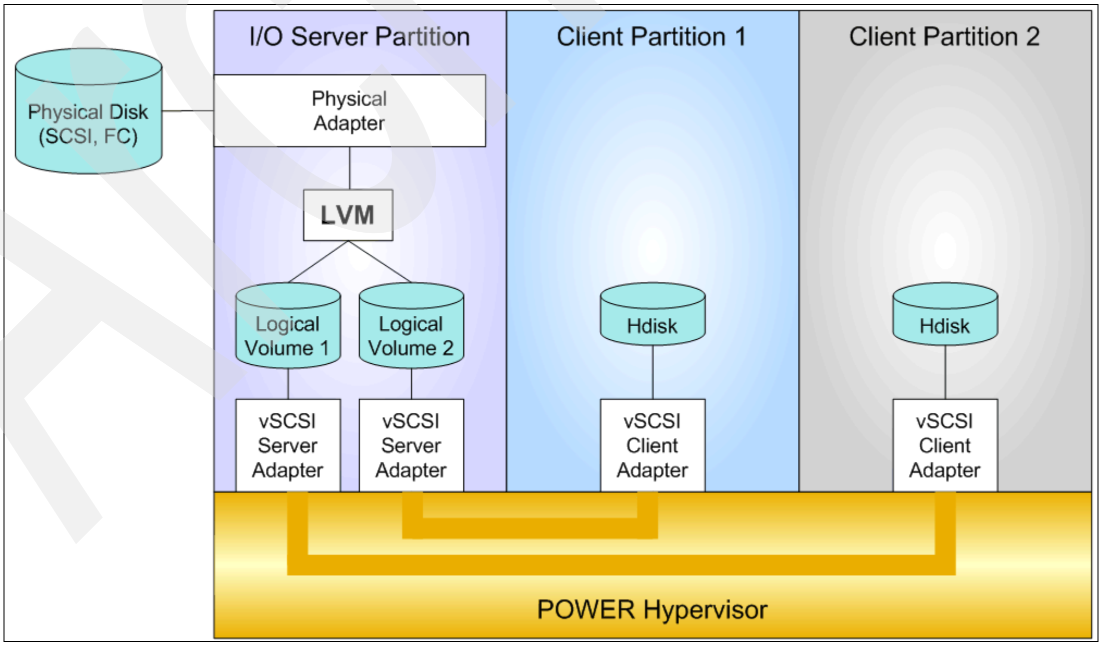


Figure 3-4 Architectural view of virtual SCSI

At the time of writing, virtual SCSI supports Fibre Channel, parallel SCSI, iSCSI, SAS, SCSI RAID devices and optical devices, including DVD-RAM and DVD-ROM. Other protocols such as SSA and tape devices are not supported.

For more information about specific storage devices supported for Virtual I/O Server, see:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html

Note: Virtual SCSI interfaces can be configured for both dedicated CPU (LPAR) and shared processor pool (SPLPAR) logical partitions.

3.7.2 Shared Ethernet Adapter

A Shared Ethernet Adapter (SEA) is used to connect a physical Ethernet network to a virtual Ethernet network. The SEA provides this access by connecting the internal Hypervisor VLANs with the VLANs on external switches. Because the SEA processes packets at layer 2, the original MAC address and VLAN tags of the packet are visible to other systems on the physical network. IEEE 802.1 VLAN tagging is supported.

The SEA also provides the ability for several client partitions to share one physical adapter. Using an SEA, you can connect internal and external VLANs using a physical adapter. The SEA service can only be hosted in the Virtual I/O Server, not in a general purpose AIX or Linux for Power partition, and acts as a layer-2 network bridge to securely transport network traffic between virtual Ethernet networks (internal) and one or more (EtherChannel) physical network adapters (external). These virtual Ethernet network adapters are defined by the POWER Hypervisor on the Virtual I/O Server

Tip: A Linux for Power partition can provide bridging function as well, by using the **`brctl`** command.

Figure 3-5 on page 123 shows a configuration example of an SEA with one physical and two virtual Ethernet adapters. An SEA can include up to 16 virtual Ethernet adapters on the Virtual I/O Server that share the same physical access.

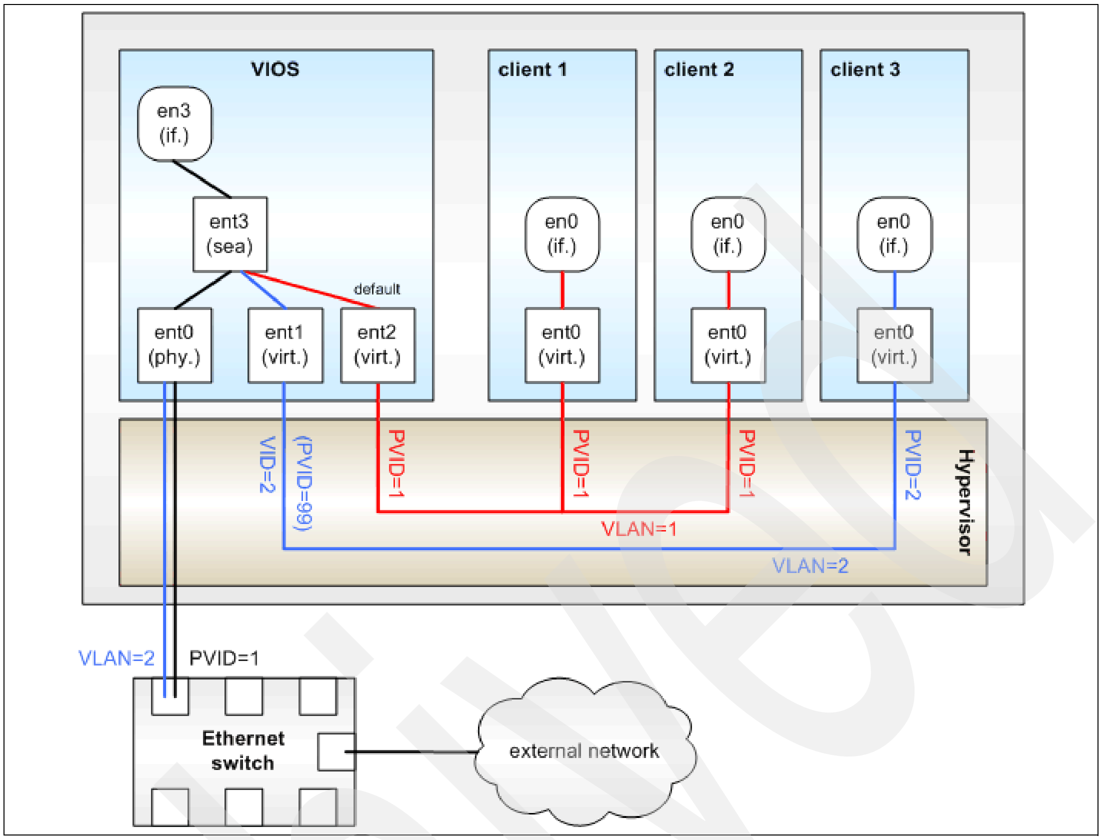


Figure 3-5 Architectural view of a Shared Ethernet Adapter

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

Unicast, broadcast, and multicast is supported, so protocols that rely on broadcast or multicast, such as Address Resolution Protocol (ARP), Dynamic Host Configuration Protocol (DHCP), Boot Protocol (BOOTP), and Neighbor Discovery Protocol (NDP) can work across an SEA.

A SEA does not require a configured IP address to be able to perform the Ethernet bridging functionality. Configuring an IP address on the Virtual I/O Server is helpful so that it can be reached by TCP/IP. This is necessary, for example, to perform dynamic LPAR operations or to enable remote login. This can be done either by configuring an IP address directly on the SEA device, or on an additional virtual Ethernet adapter in the Virtual I/O Server. This leaves the SEA without the IP address, allowing for maintenance on the SEA without losing IP connectivity in case SEA failover is configured.

Note: Shared Ethernet Adapter interfaces can be used by both dedicated CPU (LPAR) and shared processor pool (SPLPAR) client logical partitions.

For a more detailed discussion about virtual networking, see:

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

3.8 PowerVM Lx86

The IBM PowerVM Lx86 feature creates a virtual x86 application environment within a Linux operating system-based partition running on your 595 server. Most 32-bit x86 Linux applications can run in the Lx86 environment without requiring clients or ISVs to recompile the code. This brings new benefits to organizations who want the reliability and flexibility of consolidating (through virtualization) on Power Systems and desire to use applications that have not yet been ported to the POWER platform.

PowerVM Lx86 dynamically translates x86 instructions to Power Architecture instructions, operating much like the just-in-time compiler (JIT) in a Java system. The technology creates an environment on which the applications being translated run on the new target platform, in this case Linux for Power. This environment encapsulates the application and runtime libraries and runs them on the Linux for Power operating system kernel. These applications can be run side by side with Linux for Power native applications on a single system image and do not require a separate partition.

Figure 3-6 shows the diagram of the Linux x86 application environment.

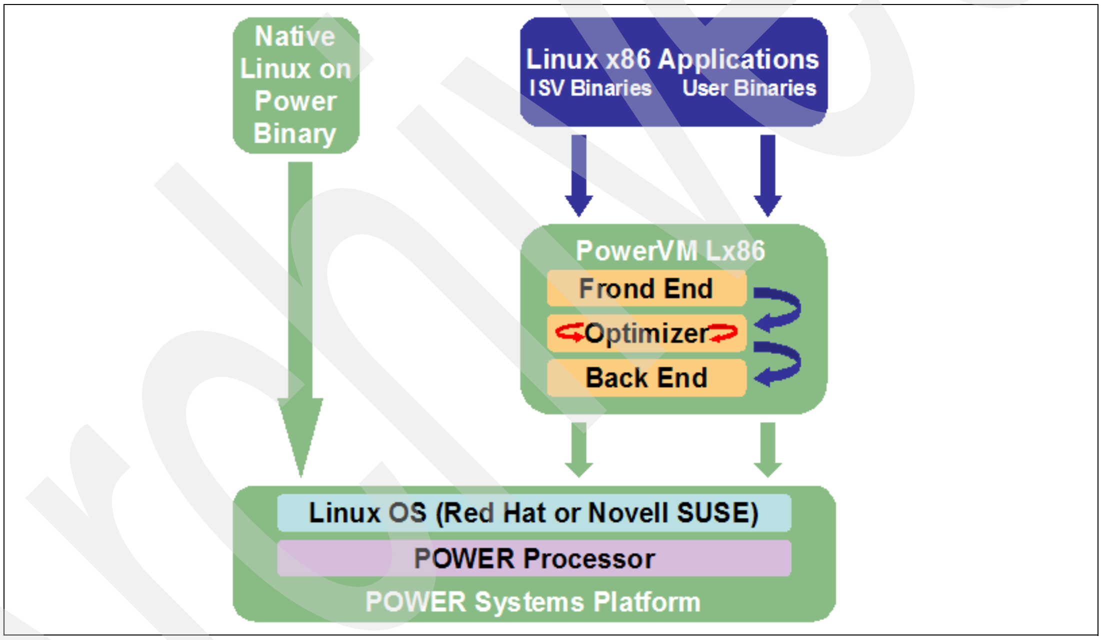


Figure 3-6 Diagram of the Linux x86 application environment

Supported Linux for Power operating systems

PowerVM Lx86 version 1.1 supports the following Linux for Power operating systems:

- Red Hat Enterprise Linux 4 (RHEL 4) for POWER version 4.4 and 4.5. Also supported are x86 Linux applications running on RHEL 4.3.
- SUSE Linux Enterprise Server 9 (SLES 9) for POWER Service Pack 3
- SUSE Linux Enterprise Server 10 (SLES 10) for POWER Service Pack 1

Notes:

- PowerVM LX86 is supported under the VIOS Software Maintenance Agreement (SWMA).
- When using PowerVM Lx86 on an IBM System p POWER6 processor-based system, only SLES 10 with SP1 and RHEL 4.5 are supported.
- Make sure the x86 version is the same as your Linux for Power version. Do not try to use any other version because it is unlikely to work. One exception is with Red Hat Enterprise Linux, both the Advanced Server and Enterprise Server option at the correct release will work.

Although PowerVM Lx86 runs most x86 Linux applications, PowerVM Lx86 cannot run applications that:

- Directly access hardware devices (for example, graphics adapters).
- Require nonstandard kernel module access or use kernel modules not provided by the Linux for Power Systems operating system distribution.
- Do not use only the Intel® IA-32 instruction set architecture as defined by the 1997 Intel Architecture Software Developer's Manual consisting of Basic Architecture (order number 243190), Instruction Set Reference Manual (Order Number 243191) and the System Programming Guide (order number 243192) dated 1997.
- Do not run correctly on Red Hat Enterprise Linux 4 starting with version 4.3 or Novell SUSE Linux Enterprise Server (SLES) 9 starting with version SP3 or Novell SLES 10.
- Use x86 Linux specific system administration or configuration tools.

For more information about PowerVM Lx86, refer to *Getting Started with PowerVM Lx86*, REDP-4298, located at:

http://www.ibm.com/redbooks

3.9 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility allows you to move a fully virtualized, running logical partition between any two POWER6 based servers without a shutdown or disruption to the operation of that logical partition. A companion feature, Inactive Partition Mobility, allows you to move a powered off logical partition from one system to another.

At a high level, the required environment to support Live Partition Mobility includes:

- Two POWER6-based servers (they can be different models).
- An external disk subsystem, capable of concurrent access, and zoned to a Virtual I/O partition on each POWER6 server.
- A running logical partition configured with virtual Ethernet and SCS adapters (no dedicated adapters).
- Sufficient resources on the target POWER6 server so that a new logical partition with similar resources can be configured.

Figure 3-7 on page 126 provides a visual representation of these requirements.

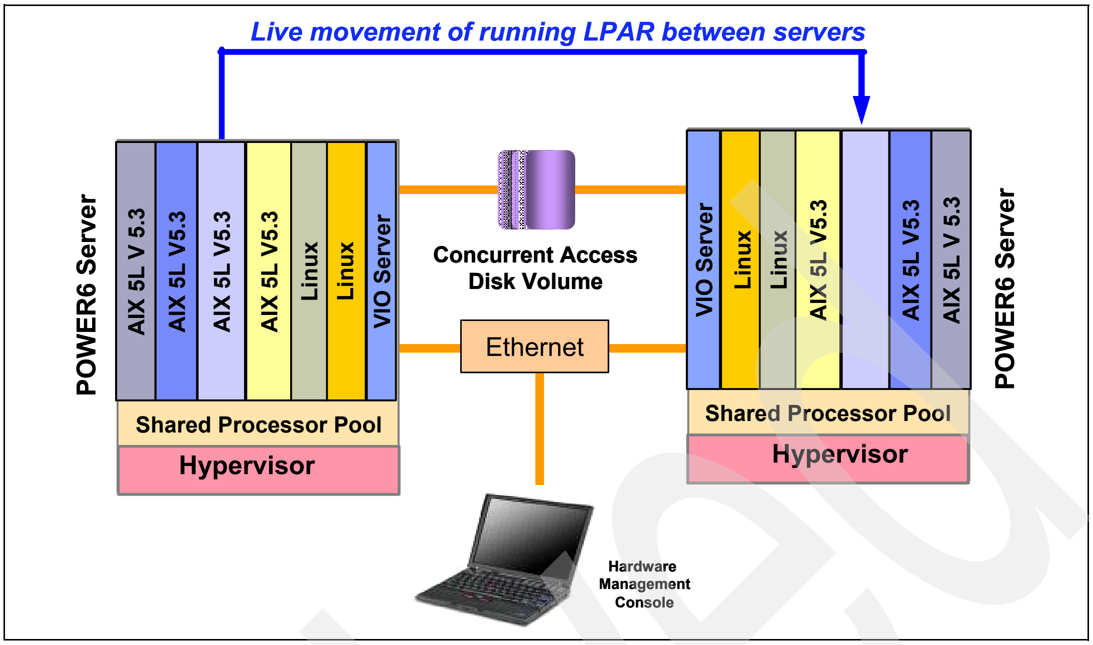


Figure 3-7 Live partition mobility example

Partition mobility provides systems management flexibility and can improve system availability when you:

- Avoid planned outages for hardware or firmware maintenance by moving logical partitions to another server and then performing the maintenance. Live partition mobility can help lead to zero downtime hardware maintenance because you can use it to evacuate the server on which you have to perform scheduled maintenance activities.
- Avoid downtime for a server upgrade by moving logical partitions to another server and then performing the upgrade. This allows your end users to continue their work without disruption.
- Perform preventive failure management: If a server indicates a potential failure, using Partition Mobility you can move its logical partitions to another server before the failure occurs.
- Perform server optimization by:
  - Consolidation: You can move workloads running on several small, under-used servers onto a single larger server.
  - Deconsolidation: You can move workloads from server to server to optimize resource use and workload performance within your computing environment. With active partition mobility, you can manage workloads with minimal downtime.

Mobile partition’s operating system requirements

Partition mobility is currently supported with fully virtualized AIX and Linux for Power partitions. The Virtual I/O Server logical partition must be at release level 1.5 or higher. Note that Virtual I/O Server partitions cannot be migrated. The operating system must be at one of the following levels:

- AIX 5L V5.3 with 5300-07 Technology Level or later
- AIX V6.1 or later
- Red Hat Enterprise Linux Version V5.1 or later

– SUSE Linux Enterprise Services 10 (SLES 10) Service Pack 1 or later

Previous versions of AIX and Linux for Power can participate in Inactive Partition Mobility, if the operating systems support virtual devices and IBM System p POWER6 processor-based systems.

Source and destination system requirements

The source partition must be one that only has virtual devices. If any physical devices are in its allocation, they must be removed before the validation or migration is initiated.

The Hypervisor must support the Partition Mobility functions, also called migration process. POWER 6 processor-based hypervisors have this capability. If migrating to a POWER6 processor-based server (other than a 595 server), the server must be at firmware level eFW3.2 or later. Source and destination systems could have different firmware levels, but they must be compatible with each other. The firmware instructions note any incompatibilities.

The Virtual I/O Server on the source system provides the access to the clients resources and must is identified as a Mover Service Partition (MSP). The Virtual Asynchronous Services Interface (VASI) device allows the mover service partition to communicate with the hypervisor. It is created and managed automatically by the HMC and is configured on both the source and destination Virtual I/O Servers designated as the mover service partitions. Other requirements include:

- The time of day setting must be the same on each server.
- Systems should not be running on battery power.
- Shared access to external storage (external hdisk with reserve_policy=no_reserve) must exist.
- All logical partitions should be on the same open network with RMC established to the HMC.

The HMC is used to configure, validate and to orchestrate a live partition mobility operation. You will use the HMC to configure the Virtual I/O Server as an MSP and to configure the VASI device. An HMC wizard validates your configuration and identifies any items that would cause the migration to fail. The HMC controls all phases of the migration process.

For more information about Live Partition Mobility and how to implement it, refer to *PowerVM Live Partition Mobility on IBM System p*, SG24-7460 which can be found at:

http://www.ibm.com/redbooks

3.10 AIX 6 workload partitions

Workload partitions is a software-based virtualization feature of AIX V6 that provides the ability to run multiple applications within the same instance of an AIX operating system. Each workload partition (WPAR) runs in separate application space which provides security and administrative isolation between the applications. You can control processor and memory allocations for each WPAR.

Multiple WPARs can be created within a single AIX 6 instance which is running on a stand-alone server, logical partition, or shared processor pool partition. Workload partitions (WPARs) can improve administrative efficiency by reducing the number of AIX operating system instances that must be maintained and can increase the overall utilization of systems by consolidating multiple workloads on a single system.

Live Application Mobility is part of workload partitions, and enables a way to relocate a WPAR from one system to another.

To assist with the management of WPARs, IBM offers an optional product called Workload Partition Manager (5765-WPM). WPAR Manager provides a centralized point of control for managing WPARs across one or more managed systems running AIX 6.1.

> **Note:** Workload partitions are only supported with AIX V6.

For a detailed list of the workload partitions concepts and function, refer to *Introduction to Workload Partition Management in IBM AIX Version 6.1*, SG24-7431, located at:

http://www.ibm.com/redbooks

3.11 System Planning Tool

The IBM System Planning Tool (SPT) helps you design a system or systems to be partitioned with logical partitions. You can also plan for and design non-partitioned systems using the SPT. The resulting output of your design is called a *system plan*, which is stored in a .sysplan file. This file can contain plans for a single system or multiple systems. The .sysplan file can be used for the following purposes:

- To create reports
- As input to the IBM configuration tool (eConfig)
- To create and deploy partitions on your systems automatically

The SPT is the next generation of the IBM LPAR Validation Tool (LVT). It contains all the functions from the LVT, and significant functional enhancements. It is integrated with the IBM Systems Workload Estimator (WLE). System plans generated by the SPT can be deployed on the system by the Hardware Management Console (HMC).

You can create an entirely new system configuration, or you can create a system configuration based upon any of the following information:

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

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

Use the SPT before you order a system to determine what you must order to support your workload. You can also use the SPT to determine how to partition an existing system.

The SPT is a PC-based browser application designed to be run in a standalone environment. The SPT can be used to plan solutions for the following IBM systems:

- IBM Power Systems
- System p5™ and System i5™
- eServer p5 and eServer i5
- OpenPower®
- iSeries® 8xx and 270 models

SPT can create two files from your configuration work (which can include any combination of multiple AIX, IBM i, and Linux for Power partitions):

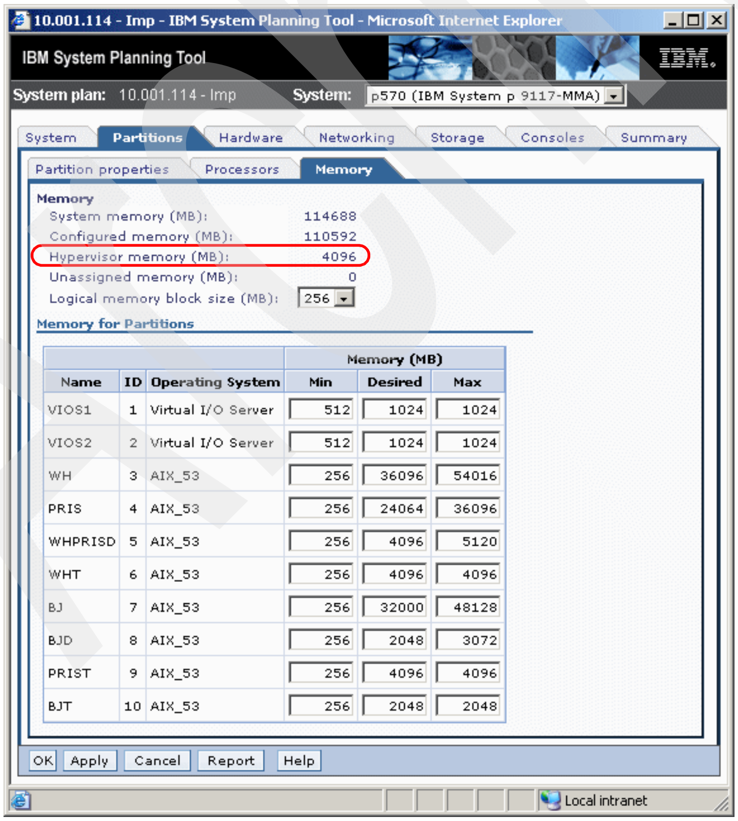
- Configuration file: The configuration file, denoted by the *.cfr* file extension, can be imported to the IBM sales configurator tool (e-Config) as input to the ordering process. Using SPT to create the configuration files helps IBM manufacturing, as much as is possible, configure your new order so that all the hardware is physically located and ready for use by the logical partitions you have configured. Using this file helps minimize the moving of cards to different cards slots when deploying partition configurations on the delivered system.
- System plan file: When the system plan file, denoted by the *.sysplan* file extension, is on the HMC, it can be used to deploy the partition configurations defined in the .sysplan on the installed system. This file can be exported to media and used by an HMC to deploy the partition configurations stored within the .sysplan file.

Consider using the IBM Systems Workload Estimator (Workload Estimator) tool to *performance size* your POWER6 partitions. This tool can be accessed at:

http://www-912.ibm.com/estimator

This helps to determine the processor capacity, and memory and disk configuration sufficient to meet your performance objectives for each partition (or only partition) on a system. Based on that output, you can use the IBM System Planning Tool to configure the hardware for each partition. As part of its output, SPT also estimates POWER Hypervisor requirements as part of the memory resources required for all partitioned and non-partitioned servers.

Figure 3-8 is an SPT window example showing the estimated Hypervisor memory requirements based on sample partition requirements.



System plan: 10.001.114 - Imp
System: p570 (IBM System p 9117-MMA)

Memory
System memory (MB): 114688
Configured memory (MB): 110592
Hypervisor memory (MB): 4096
Unassigned memory (MB): 0
Logical memory block size (MB): 256

Memory for Partitions

			Memory (MB)		
Name	ID	Operating System	Min	Desired	Max
VIOS1	1	Virtual I/O Server	512	1024	1024
VIOS2	2	Virtual I/O Server	512	1024	1024
WH	3	AIX_53	256	36096	54016
PRIS	4	AIX_53	256	24064	36096
WHPRISD	5	AIX_53	256	4096	5120
WHT	6	AIX_53	256	4096	4096
BJ	7	AIX_53	256	32000	48128
BJD	8	AIX_53	256	2048	3072
PRIST	9	AIX_53	256	4096	4096
BJT	10	AIX_53	256	2048	2048

Figure 3-8 IBM System Planning Tool window showing Hypervisor memory requirements

Note: In previous releases of the SPT, you could view an HMC system plan, but not edit it. The SPT now allows you to convert an HMC system plan to a format you can edit in the SPT.

Also note that SPT 2.0 is the last release to support .lvt and .xml files. You should load your old .lvt and .xml plans, and then save them as .sysplan files.

The SPT and its supporting documentation can be found on the IBM System Planning Tool site at:

http://www.ibm.com/systems/support/tools/systemplanningtool/

Note: Consider using #0456 on new orders to specify adapter card placement within the system unit and I/O enclosures to expedite getting your LPAR configuration up and running after your system is delivered. This is where the SPT configuration file (for the order) and system plan file (for deployment) can be most used.

4

Continuous availability and manageability

This chapter provides information about IBM Power Systems design features that help lower the total cost of ownership (TCO). The advanced IBM RAS (Reliability, Availability, and Serviceability) technology allows the possibility to improve your architecture's TCO by reducing unplanned down time.

In April 2008, IBM announced the newest Power Architecture technology-based server: the Power 595, incorporating innovative IBM POWER6 processor technology to deliver both outstanding performance and enhanced RAS capabilities. The IBM Power Servers complement the IBM POWER5 processor-based server family, coupling technology innovation with new capabilities designed to ease administrative burdens and increase system use. In addition, IBM PowerVM delivers virtualization technologies for IBM Power Systems product families, enabling individual servers to run dozens or even hundreds of mission-critical applications.

In IBMs view, servers must be designed to avoid both planned and unplanned outages, and to maintain a focus on application uptime. From an RAS standpoint, servers in the IBM Power Systems family include features to increase availability and to support new levels of virtualization, building upon the leading-edge RAS features delivered in the IBM eServer p5, pSeries and iSeries families of servers.

IBM POWER6 processor-based systems have a number of new features that enable systems to dynamically adjust when issues arise that threaten availability. Most notably, POWER6 processor-based systems introduce the POWER6 Processor Instruction Retry suite of tools, which includes Processor Instruction Retry, Alternate Processor Recovery, Partition Availability Prioritization, and Single Processor Checkstop. Taken together, in many failure scenarios these features allow a POWER6 processor-based system to recover with no impact on a partition using a failing core.

This chapter discusses several features that are based on the benefits available when using AIX and IBM i as the operating system. Support for these features when using Linux for Power should be verified with your Linux supplier.

4.1 Reliability

Highly reliable systems are built with highly reliable components. On IBM POWER6 processor-based systems, this basic principle is expanded on with a clear design for reliability architecture and methodology. A concentrated, systematic, architecture-based approach is designed to improve overall system reliability with each successive generation of system offerings.

4.1.1 Designed for reliability

Systems that have fewer components and interconnects have fewer opportunities to fail. Simple design choices such as integrating two processor cores on a single POWER chip can dramatically reduce the opportunity for system failures. In this case, a 4-core server will include half as many processor chips (and chip socket interfaces) in comparison to a single CPU core per processor design. Not only does this reduce the total number of system components, it reduces the total amount of heat generated in the design, resulting in an additional reduction in required power and cooling components.

Parts selection also plays a critical role in overall system reliability. IBM uses three grades of components, with Grade 3 defined as industry standard (off-the-shelf). As shown in Figure 4-1, using stringent design criteria and an extensive testing program, the IBM manufacturing team can produce Grade 1 components that are expected to be 10 times more reliable than industry standard. Engineers select Grade 1 parts for the most critical system components. Newly introduced organic packaging technologies, rated Grade 5, achieve the same reliability as Grade 1 parts.

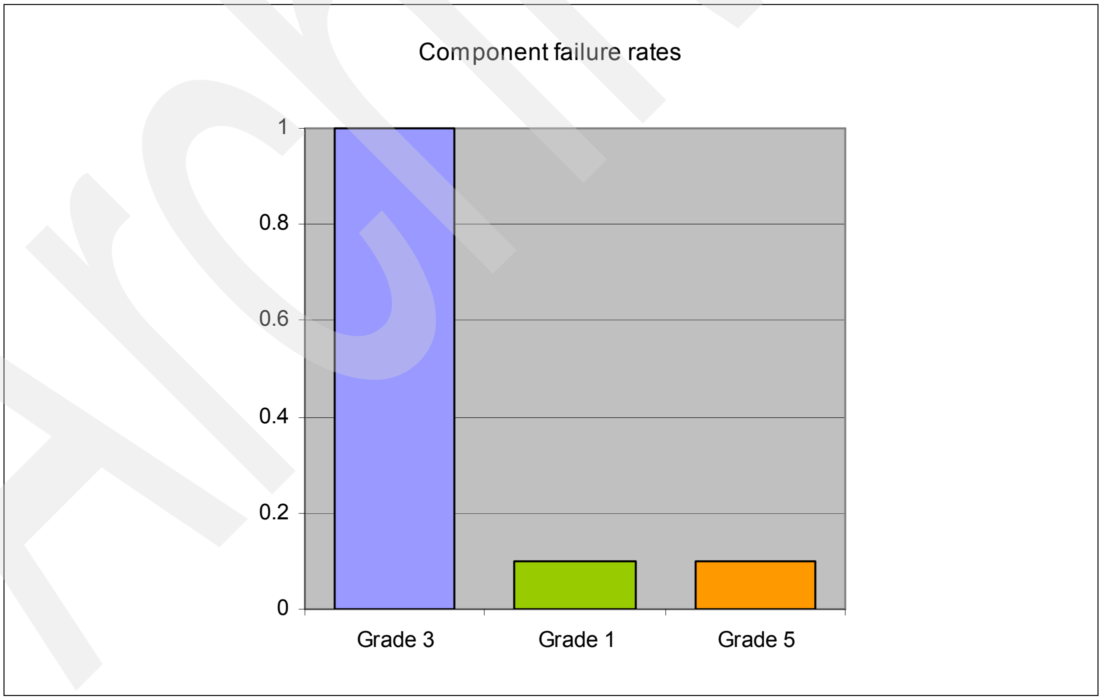


Figure 4-1 Component failure rates

4.1.2 Placement of components

Packaging is designed to deliver both high performance and high reliability. For example, the reliability of electronic components is directly related to their thermal environment, that is, large decreases in component reliability directly correlate with relatively small increases in temperature; POWER6 processor-based systems are carefully packaged to ensure adequate cooling. Critical system components such as the POWER6 processor chips are positioned on printed circuit cards so they receive cool air during operation. In addition, POWER6 processor-based systems are built with redundant, variable-speed fans that can automatically increase output to compensate for increased heat in the central electronic complex (CEC).

IBMs POWER6 chip was designed to save energy and cooling costs. Innovations include:

- A dramatic improvement in the way instructions are executed inside the chip. Performance was increased by keeping static the number of pipeline stages, but making each stage faster, removing unnecessary work and doing more in parallel. As a result, execution time is cut in half or energy consumption is reduced.
- Separating circuits that cannot support low voltage operation onto their own power supply *rails*, dramatically reducing power for the rest of the chip.
- Voltage/frequency *slewing*, which enables the chip to lower electricity consumption by up to 50%, with minimal performance impact.

Innovative and pioneering techniques allow the POWER6 chip to turn off its processor clocks when there's no useful work to be done, then turn them on when needed, reducing both system power consumption and cooling requirements. Power saving is also realized when the memory is not fully utilized, as power to parts of the memory not being utilized is dynamically turned off and then turned back on when needed. When coupled with other RAS improvements, these features can deliver a significant improvement in overall system availability.

4.1.3 Continuous field monitoring

Aided by the IBM First Failure Data Capture (FFDC) methodology and the associated error reporting strategy, product engineering builds an accurate profile of the types of failures that might occur, and initiates programs to enable corrective actions. Product engineering also continually analyzes critical system faults, testing to determine if system firmware and maintenance procedures and tools are effectively handling and recording faults as designed. See section 4.3.3, “Detecting errors” on page 152.

A system designed with the FFDC methodology includes an extensive array of error checkers and fault isolation registers (FIR) to detect, isolate, and identify faulty conditions in a server. This type of automated error capture and identification is especially useful in providing a quick recovery from unscheduled hardware outages. This data provides a basis for failure analysis of the component, and can also be used to improve the reliability of the part and as the starting point for design improvements in future systems.

IBM RAS engineers use specially designed logic circuitry to create faults that can be detected and stored in FIR bits, simulating internal chip failures. This technique, called *error injection*, is used to validate server RAS features and diagnostic functions in a variety of operating conditions (power-on, boot, and operational run-time phases). IBM traditionally classifies hardware error events in the following ways:

- Repair actions (RA) are related to the industry standard definition of mean time between failure (MTBF). An RA is any hardware event that requires service on a system. Repair actions include incidents that effect system availability and incidents that are concurrently repaired.

- Unscheduled incident repair action (UIRA) is a hardware event that causes a system or partition to be rebooted in full or degraded mode. The system or partition will experience an unscheduled outage. The restart can include some level of capability degradation, but remaining resources are made available for productive work.
- High impact outage (HIO) is a hardware failure that triggers a system crash, which is not recoverable by immediate reboot. This is usually caused by failure of a component that is critical to system operation and is, in some sense, a measure of system single points-of-failure. HIOs result in the most significant availability impact on the system, because repairs cannot be effected without a service call.

A consistent, architecture-driven focus on system RAS (using the techniques described in this document and deploying appropriate configurations for availability), has led to almost complete elimination of HIOs in currently available POWER processor servers.

The clear design goal for Power Systems is to prevent hardware faults from causing an outage: platform or partition. Part selection for reliability, redundancy, recovery and self-healing techniques, and degraded operational modes are used in a coherent, methodical strategy to avoid HIOs and UIRAs.

4.2 Availability

IBMs extensive system of FFDC error checkers also supports a strategy of Predictive Failure Analysis®, which is the ability to track intermittent correctable errors and to vary components off-line before they reach the point of hard failure causing a crash.

The FFDC methodology supports IBMs autonomic computing initiative. The primary RAS design goal of any POWER processor-based server is to prevent unexpected application loss due to unscheduled server hardware outages. To accomplish this goal, a system must have a quality design that includes critical attributes for:

- Self-diagnosing and self-correcting during run time
- Automatically reconfiguring to mitigate potential problems from suspect hardware
- Self-healing or automatically substituting good components for failing components

4.2.1 Detecting and deallocating failing components

Runtime correctable or recoverable errors are monitored to determine if there is a pattern of errors. If these components reach a predefined error limit, the service processor initiates an action to deconfigure the faulty hardware, helping to avoid a potential system outage and to enhance system availability.

Persistent deallocation

To enhance system availability, a component that is identified for deallocation or deconfiguration on a POWER6 processor-based system is flagged for persistent deallocation. Component removal can occur either dynamically (while the system is running) or at boot-time (initial program load (IPL)), depending both on the type of fault and when the fault is detected.

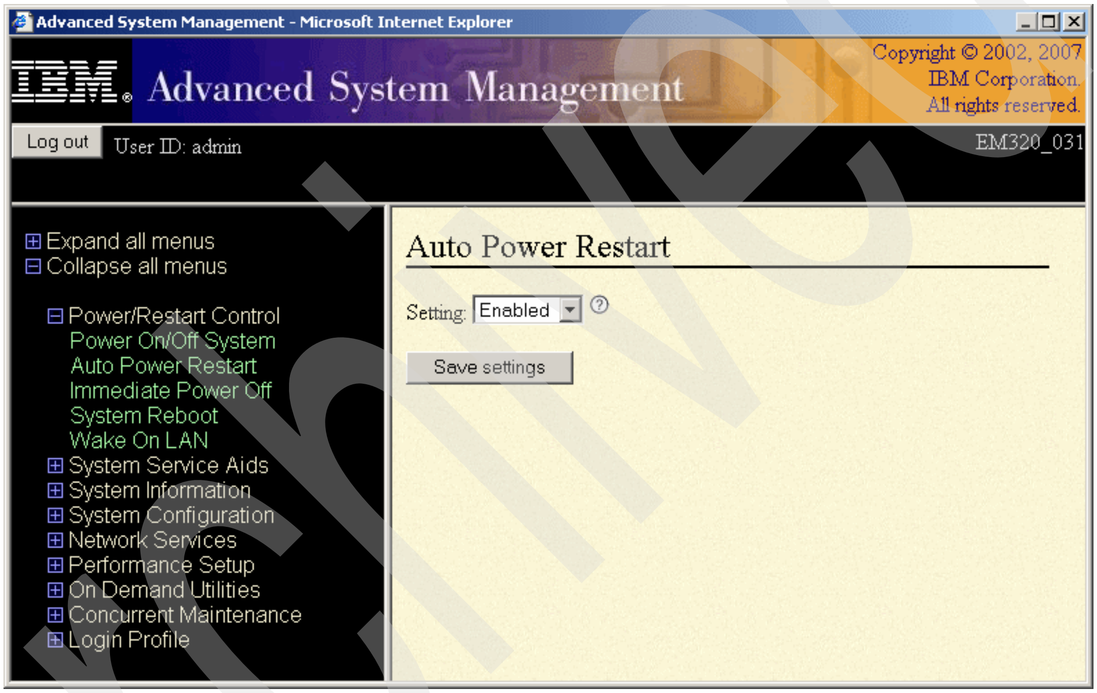
In addition, runtime unrecoverable hardware faults can be deconfigured from the system after the first occurrence. The system can be rebooted immediately after failure and resume operation on the remaining stable hardware. This prevents the same faulty hardware from affecting system operation again, and the repair action is deferred to a more convenient, less critical time.

Persistent deallocation functions include:

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

Note: The auto-restart (reboot) option has to be enabled from the Advanced System Manager interface (ASMI) or from the Operator Panel.

Figure 4-2 shows the ASMI window.



Setting: Enabled

Figure 4-2 ASMI Auto Power Restart setting

Dynamic processor deallocation

Dynamic processor deallocation enables automatic deconfiguration of processor cores when patterns of recoverable errors, for example correctable errors on processor caches, are detected. Dynamic processor deallocation prevents a recoverable error from escalating to an unrecoverable system error, which could result in an unscheduled server outage. Dynamic processor deallocation relies on the service processor's ability to use FFDC-generated recoverable error information to notify the POWER Hypervisor when a processor core reaches its predefined error limit. Then, the POWER Hypervisor, in conjunction with the operating system, redistributes the work to the remaining processor cores, deallocates the offending processor core, continues normal operation, and can revert from simultaneous multiprocessing to uniprocessor processing.

IBMs logical partitioning strategy allows any processor core to be shared with any partition on the system, thus enabling the following sequential scenarios for processor deallocation and sparing:

1. When available, an unlicensed Capacity on Demand (CoD) processor core is by default automatically used for dynamic processor sparing.
2. If no CoD processor core is available, the POWER Hypervisor attempts to locate an unallocated core somewhere in the system.
3. If no spare processor core is available, the POWER Hypervisor attempts to locate a total of 1.00 spare processing units from a shared processor pool and redistributes the workload.
4. If the requisite spare capacity is not available, the POWER Hypervisor determines how many processing units each partition is required to relinquish to create at least 1.00 processing unit shared processor pool.
5. When a full core equivalent is attained, the CPU deallocation event occurs.

Figure 4-3 shows CoD processor cores that are available for dynamic processor sparing.

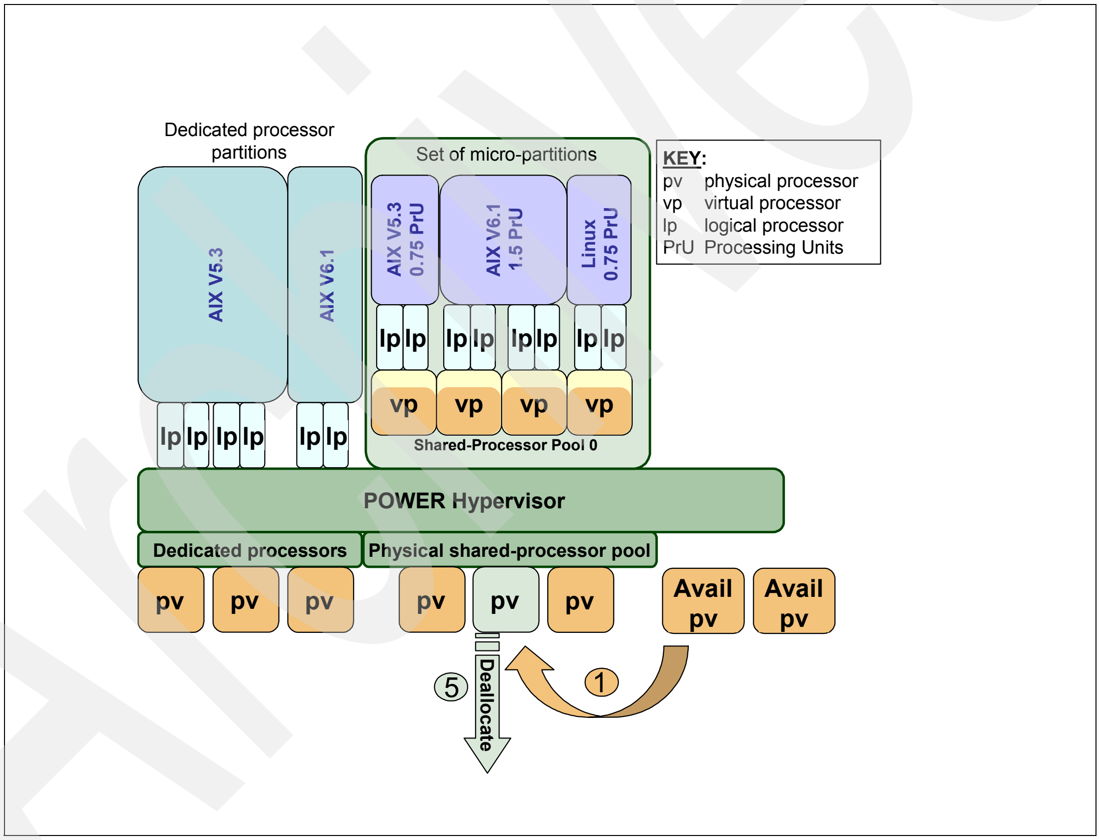


Figure 4-3 Dynamic processor deallocation and dynamic processor sparing

The deallocation event is not successful if the POWER Hypervisor and operating system cannot create a full core equivalent. This results in an error message and the requirement for a system administrator to take corrective action. In all cases, a log entry is made for each partition that could use the physical core in question.

POWER6 processor instruction retry

To achieve the highest levels of server availability and integrity, FFDC and recovery safeguards must protect the validity of user data anywhere in the server, including all the internal storage areas and the buses used to transport data. Equally important is to authenticate the correct operation of internal latches (registers), arrays, and logic within a processor core that comprise the system execution elements (branch unit, fixed instruction, floating point instruction unit and so forth) and to take appropriate action when a fault (*error*) is discovered.

The POWER6 microprocessor has incrementally improved the ability of a server to identify potential failure conditions by including enhanced error check logic, and has dramatically improved the capability to recover from core fault conditions. Each core in a POWER6 microprocessor includes an internal processing element known as the Recovery Unit (*r unit*). Using the Recovery Unit and associated logic circuits, the POWER6 microprocessor takes a *snap shot*, or *checkpoint*, of the architected core internal state before each instruction is processed by one of the core's nine-instruction execution units.

If a fault condition is detected during any cycle, the POWER6 microprocessor uses the saved state information from r unit to effectively *roll back* the internal state of the core to the start of instruction processing, allowing the instruction to be retried from a *known good* architectural state. This procedure is called *processor instruction retry*. In addition, using the POWER Hypervisor and service processor, architectural state information from one recovery unit can be loaded into a different processor core, allowing an entire instruction stream to be restarted on a substitute core. This is called *alternate processor recovery*.

POWER6 processor-based systems include a suite of mainframe-inspired processor instruction retry features that can significantly reduce situations that could result in checkstop:

- Processor instruction retry: Automatically retry a failed instruction and continue with the task. By combining enhanced error identification information with an integrated Recovery Unit, a POWER6 microprocessor can use processor instruction retry to transparently operate through (recover from) a wider variety of fault conditions (for example *non-predicted* fault conditions undiscovered through predictive failure techniques) than could be handled in earlier POWER processor cores. For transient faults, this mechanism allows the processor core to recover completely from what would otherwise have caused an application, partition, or system outage.
- Alternate processor recovery: For solid (hard) core faults, retrying the operation on the same processor core is not effective. For many such cases, the alternate processor recovery feature deallocates and deconfigures a failing core, moving the instruction stream to, and restarting it on, a spare core. The POWER Hypervisor and POWER6 processor-based hardware can accomplish these operations without application interruption, allowing processing to continue unimpeded, as follows:

  a. Identifying a spare processor core.

     Using an algorithm similar to that employed by dynamic processor deallocation, the POWER Hypervisor manages tss of acquiring a spare processor core.

  b. Using partition availability priority.

     Starting with POWER6 technology, partitions receive an integer rating with the lowest priority partition rated at 0 and the highest priority partition valued at 255. The default value is set at 127 for standard partitions and 192 for Virtual I/O Server (VIOS) partitions. Partition availability priorities are set for both dedicated and shared partitions. To initiate alternate processor recovery when a spare processor is not available, the POWER Hypervisor uses the partition availability priority to identify low priority partitions and keep high priority partitions running at full capacity.

c. Deallocating faulty core.

Upon completion of the Alternate Processor Recovery operation, the POWER Hypervisor will de-allocate the faulty core for deferred repair.

- Processor contained checkstop: If a specific processor detected fault cannot be recovered by processor instruction retry and alternate processor recovery is not an option, then the POWER Hypervisor will terminate (checkstop) the partition that was using the processor core when the fault was identified. In general, this limits the outage to a single partition. When all the previously mentioned mechanisms fail, in almost all cases (excepting the POWER Hypervisor) a termination will be contained to the single partition using the failing processor core.

Memory protection

Memory and cache arrays comprise data bit lines that feed into a memory word. A memory word is addressed by the system as a single element. Depending on the size and addressability of the memory element, each data bit line can include thousands of individual bits or memory cells. For example:

- A single memory module on a dual inline memory module (DIMM) can have a capacity of 1 Gb, and supply eight bit lines of data for an error correcting code (ECC) word. In this case, each bit line in the ECC word holds 128 Mb behind it, corresponding to more than 128 million memory cell addresses.
- A 32 KB L1 cache with a 16-byte memory word, alternatively would have only 2 KB behind each memory bit line.

A memory protection architecture that provides good error resilience for a relatively small L1 cache might be very inadequate for protecting the much larger system main storage. Therefore, a variety of different protection methods are used in POWER6 processor-based systems to avoid uncorrectable errors in memory.

Memory protection plans must take into account many factors, including:

- Size
- Desired performance
- Memory array manufacturing characteristics

POWER6 processor-based systems have a number of protection schemes to prevent, protect, or limit the effect of errors in main memory:

Hardware scrubbing This method deals with soft errors. IBM POWER6 processor systems periodically address all memory locations; any memory locations with an ECC error are rewritten with the correct data.

Error correcting code Error correcting code (ECC) allows a system to detect up to two errors in a memory word and correct one of them. However, without additional correction techniques, if more than one bit is corrupted a system can fail.

Chipkill This is an enhancement to ECC that enables a system to sustain the failure of an entire DRAM. Chipkill spreads the bit lines from a DRAM over multiple ECC words, so that a catastrophic DRAM failure would affect at most one bit in each word. Barring a future single bit error, the system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced.

Redundant bit steering This helps to avoid a situation in which multiple single-bit errors align to create a multi-bit error. In the event that an IBM POWER6 process-based system detects an abnormal number of errors on a bit line, it can dynamically steer the data stored at this bit line into one of a number of spare lines.

Figure 4-4 shows memory protection capabilities in action.

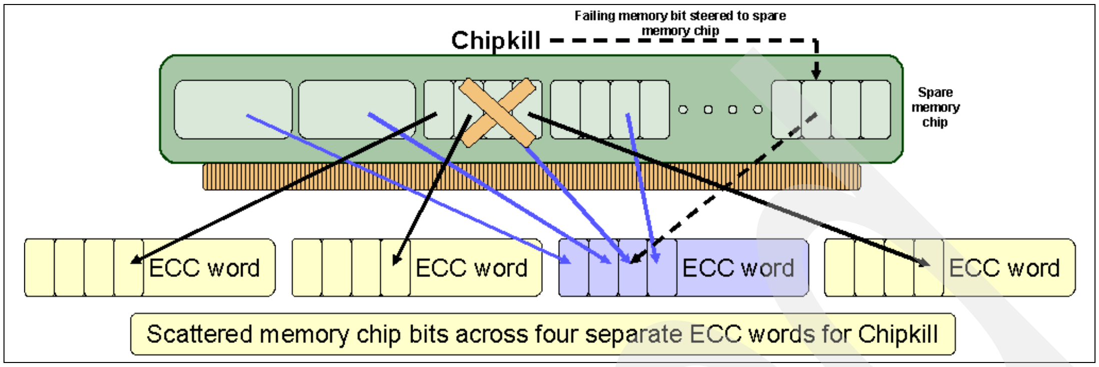

Failing memory bit steered to spare memory chip
Chipkill
Spare memory chip
ECC word
ECC word
ECC word
ECC word
Scattered memory chip bits across four separate ECC words for Chipkill

Figure 4-4 Memory protection capabilities in action

Memory page deallocation

Although coincident single cell errors in separate memory chips is a statistic rarity, POWER6 processor systems can contain these errors by using a memory page deallocation scheme for partitions running AIX and for memory pages owned by the POWER Hypervisor. If a memory address experiences an uncorrectable or repeated correctable single cell error, the service processor sends the memory page address to the POWER Hypervisor to be marked for deallocation.

The operating system performs memory page deallocation without any user intervention and is transparent to end users and applications.

The POWER Hypervisor maintains a list of pages marked for deallocation during the current platform IPL. During a partition IPL, the partition receives a list of all the bad pages in its address space.

In addition, if memory is dynamically added to a partition (through a Dynamic LPAR operation), the POWER Hypervisor warns the operating system if memory pages are included that must be deallocated.

Finally, if an uncorrectable error occurs, the system can deallocate the memory group associated with the error on all subsequent system reboots until the memory is s intended to guard against future uncorrectable errors while waiting for parts replacement.

Note: Although memory page deallocation handles single cell failures, because of the sheer size of data in a data bit line, it might be inadequate for dealing with more catastrophic failures. Redundant bit steering continues to be the preferred method for dealing with these types of problems.

Memory control hierarchy

While POWER6 processor systems maintain the same basic function as POWER5, including chipkill detection and correction, a redundant bit steering capability, and OS based memory page deallocation, the memory subsystem is structured differently.

The POWER6 chip includes two memory controllers (each with four ports) and two L3 cache controllers. Delivering exceptional performance for a wide variety of workloads, the Power 595 uses both POWER6 memory controllers and both L3 cache controllers for high memory performance. The other Power models deliver balanced performance using only a single memory controller. Some models also employ an L3 cache controller. A memory controller on a POWER6 process-based system is designed with four ports. Each port connects up to three DIMMS using a daisy-chained bus. The memory bus supports ECC checking on data, addresses, and command information. A spare line on the bus is also available for repair using a self-healing strategy. In addition, ECC checking on addresses and commands is extended to DIMMs on DRAMs. Because it uses a daisy-chained memory-access topology, this system can deconfigure a DIMM that encounters a DRAM fault, without deconfiguring the bus controller, even if the bus controller is contained on the DIMM.

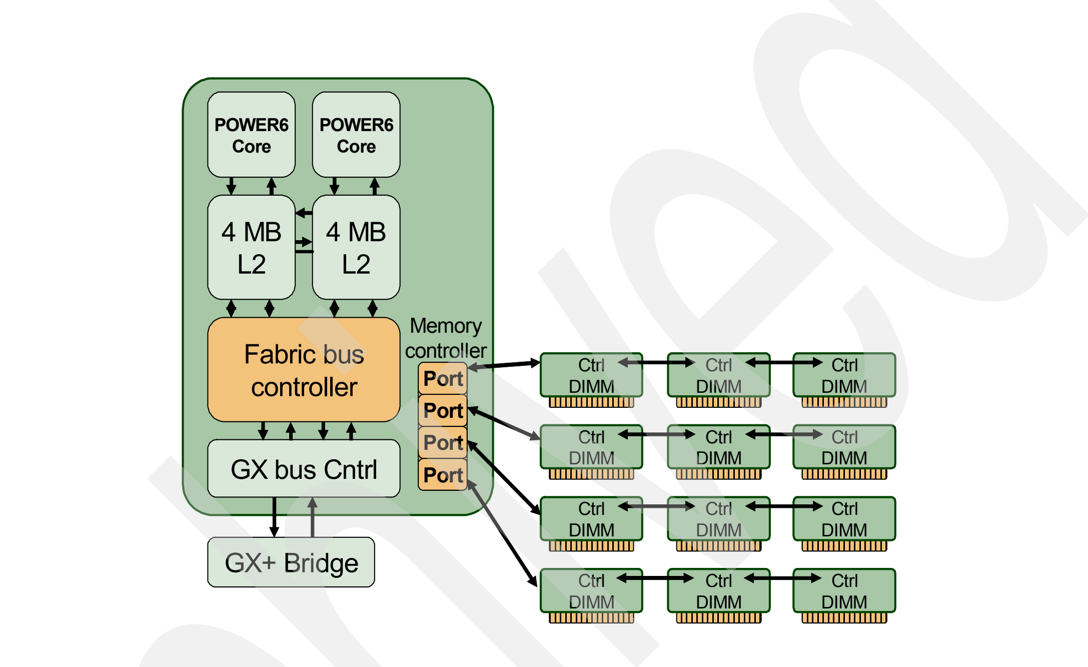


Figure 4-5 Memory control hierarchy

Memory deconfiguration

Defective memory discovered at boot time is automatically switched off, unless it is already the minimum amount required to boot. If the service processor detects a memory fault at boot time, it marks the affected memory as bad so it is not to be used on subsequent reboots (memory persistent deallocation).

If the service processor identifies faulty memory in a server that includes CoD memory, the POWER Hypervisor attempts to replace the faulty memory with available CoD memory. Faulty resources are marked as deallocated and working resources are included in the active memory space. Because these activities reduce the amount of CoD memory available for future use, repair of the faulty memory should be scheduled as soon as is convenient.

Upon reboot, if not enough memory is available to meet minimum partition requirements, the POWER Hypervisor reduces the capacity of one or more partitions. The HMC receives notification of the failed component, triggering a service call.

Defective memory discovered at IPL time is switched off by a server, as follows:

- If a memory fault is detected by the service processor at boot time, the affected memory will be marked as bad and will not be used on this or subsequent IPLs (memory persistent deallocation).

- As the manager of system memory, at boot time the POWER Hypervisor decides which memory to make available for server use and which to put in the unlicensed or spare pool, based upon system performance and availability considerations.
  - If the service processor identifies faulty memory in a server that includes CoD memory, the POWER Hypervisor attempts to replace the faulty memory with available CoD memory. As faulty resources on POWER6 or POWER5 process-based offerings are automatically *demoted* to the system's unlicensed resource pool, working resources are included in the active memory space.
  - On POWER5 midrange systems (p5-570, i5-570), only memory associated with the first card failure is spared to available CoD memory. If simultaneous failures occur on multiple memory cards, only the first memory failure found is spared.
  - Because these activities reduce the amount of CoD memory available for future use, repair of the faulty memory should be scheduled as soon as is convenient.
- Upon reboot, if not enough memory is available; the POWER Hypervisor will reduce the capacity of one or more partitions. The HMC receives notification of the failed component, triggering a service call.

4.2.2 Special uncorrectable error handling

Although it is rare, an uncorrectable data error can occur in memory or a cache. POWER6 processor systems attempt to limit, to the least possible disruption, the impact of an uncorrectable error using a well defined strategy that first considers the data source. Sometimes, an uncorrectable error is temporary in nature and occurs in data that can be recovered from another repository. For example:

- Data in the instruction L1 cache is never modified within the cache itself. Therefore, an uncorrectable error discovered in the cache is treated like an ordinary cache miss, and correct data is loaded from the L2 cache.
- The POWER6 processor's L3 cache can hold an unmodified copy of data in a portion of main memory. In this case, an uncorrectable error in the L3 cache would simply trigger a *reload* of a cache line from main memory. This capability is also available in the L2 cache.

In cases where the data cannot be recovered from another source, a technique called special uncorrectable error (SUE) handling is used to determine whether the corruption is truly a threat to the system. If, as is sometimes the case, the data is never actually used but is simply overwritten, then the error condition can safely be voided and the system will continue to operate normally.

When an uncorrectable error is detected, the system modifies the associated ECC word, thereby signaling to the rest of the system that the *standard* ECC is no longer valid. The service processor is then notified, and takes appropriate actions. When running AIX V5.2 or greater or Linux[1] and a process attempts to use the data, the operating system is informed of the error and terminates only the specific user program.

Critical data is dependant on the system type and the firmware level. For example, on POWER6 process-based servers, the POWER Hypervisor will in most cases, tolerate partition data uncorrectable errors without causing system termination. It is only in the case where the corrupt data is used by the POWER Hypervisor that the entire system must be rebooted, thereby preserving overall system integrity.

Depending on system configuration and source of the data, errors encountered during I/O operations might not result in a machine check. Instead, the incorrect data is handled by the

[1] SLES 10 SP1 or later, and in RHEL 4.5 or later (including RHEL 5.1)

processor host bridge (PHB) chip. When the PHB chip detects a problem. it rejects the data, preventing data being written to the I/O device. The PHB then enters a freeze mode halting normal operations. Depending on the model and type of I/O being used, the freeze can include the entire PHB chip, or simply a single bridge. This results in the loss of all I/O operations that use the frozen hardware until a power on reset of the PHB. The impact to a partition or partitions depends on how the I/O is configured for redundancy. In a server configured for failover availability, redundant adapters spanning multiple PHB chips could enable the system to recover transparently, without partition loss.

4.2.3 Cache protection mechanisms

In POWER5 processor-based servers, the L1 instruction cache (I-cache), directory, and instruction effective to real address translation (I-ERAT) are protected by parity. If a parity error is detected, it is reported as a cache miss or I-ERAT miss. The cache line with parity error is invalidated by hardware and the data is refetched from the L2 cache. If the error occurs again (the error is solid), or if the cache reaches its soft error limit, the processor core is dynamically deallocated and an error message for the FRU is generated.

Although the L1 data cache (D-cache) is also parity-checked, it gets special consideration when the threshold for correctable errors is exceeded. The error is reported as a synchronous machine check interrupt. The error handler for this event is executed in the POWER Hypervisor. If the error is recoverable, the POWER Hypervisor invalidates the cache (clearing the error). If additional soft errors occur, the POWER Hypervisor will disable the failing portion of the L1 D-cache when the system meets its error threshold. The processor core continues to run with degraded performance. A service action error log is created so that when the machine is booted, the failing part can be replaced. The data ERAT and translation look aside buffer (TLB) arrays are handled in a similar manner.

L1 instruction and data array protection

The POWER6 processor's instruction and data caches are protected against temporary errors by using the POWER6 processor instruction retry feature and against solid failures by alternate processor recovery, both mentioned earlier. In addition, faults in the Segment Lookaside Buffer (SLB) array are recoverable by the POWER Hypervisor.

L2 array protection

On a POWER6 processor system, the L2 cache is protected by ECC, which provides single-bit error correction and double-bit error detection. Single-bit errors are corrected before forwarding to the processor, and subsequently written back to L2. Like the other data caches and main memory, uncorrectable errors are handled during run-time by the special uncorrectable error handling mechanism. Correctable cache errors are logged and if the error reaches a threshold, a dynamic processor deallocation event is initiated.

Starting with POWER6 processor systems, the L2 cache is further protected by incorporating a dynamic cache line delete algorithm. Up to six L2 cache lines might be automatically deleted. Deletion of a few cache lines are unlikely to adversely affect server performance. When six cache lines have been repaired, the L2 is marked for persistent deconfiguration on subsequent system reboots until it can be replaced.

L3 cache

The L3 cache is protected by ECC and special uncorrectable error handling. The L3 cache also incorporates technology to handle memory cell errors.

During system runtime, a correctable error is reported as a recoverable error to the service processor. If an individual cache line reaches its predictive error threshold, the cache is purged, and the line is dynamically deleted (removed from further use). The state of L3 cache line delete is maintained in a *deallocation record*, so line delete persists through system IPL. This ensures that cache lines *varied offline* by the server can remain offline if the server is rebooted. These *error prone* lines cannot then cause system operational problems. A server can dynamically delete up to 10 cache lines in a POWER5 processor-based server and up to 14 cache lines in POWER6 processor-based models. Deletion of this many cache lines are unlikely to adversely affect server performance. If this total is reached, the L3 cache is marked for persistent deconfiguration on subsequent system reboots until repaired.

Furthermore, for POWER6 processor-based servers, the L3 cache includes a purge delete mechanism for cache errors that cannot be corrected by ECC. For unmodified data, purging the cache and deleting the line ensures that the data is read into a different cache line on reload, thus providing good data to the cache, preventing reoccurrence of the error, and avoiding an outage. For an uncorrectable error (UE) on modified data, the data is written to memory and marked as a SUE. Again, purging the cache and deleting the line allows avoidance of another UE.

In addition, POWER6 process-based servers introduce a hardware assisted cache memory scrubbing feature where all the L3 cache memory is periodically addressed and any address with an ECC error is rewritten with the faulty data corrected. In this way, soft errors are automatically removed from L3 cache memory, decreasing the chances of encountering multi-bit memory errors.

4.2.4 The input output subsystem

All IBM POWER6 processor-based servers use a unique *distributed switch* topology providing high bandwidth data busses for fast efficient operation. The high-end Power 595 server uses an 8-core building block. System interconnects scale with processor speed. Intra-MCM and Inter-MCM busses at half processor speed. Data movement on the fabric is protected by a full ECC strategy. The GX+ bus is the primary I/O connection path and operates at half the processor speed.

In this system topology, every node has a direct connection to every other node, improving bandwidth, reducing latency, and allowing for new availability options when compared to earlier IBM offerings. Offering further improvements that enhance the value of the simultaneous multithreading processor cores, these servers deliver exceptional performance in both transaction processing and numeric-intensive applications. The result is a higher level of SMP scaling. IBM POWER6 processor servers can support up to 64 physical processor cores.

I/O drawer and tower redundant connections and concurrent repair

Power System servers support a variety integrated I/O devices (disk drives, PCI cards). The standard server I/O capacity can be significantly expanded in the rack mounted offerings by attaching optional I/O drawers or I/O towers using IBM RIO-2 busses, or on POWER6 processor offerings, a 12x channel adapter for optional 12x channel I/O drawers. A remote I/O (RIO) loop or 12x cable loop includes two separate cables providing highspeed attachment. If an I/O cable becomes inoperative during normal system operation, the system can automatically reconfigure to use the second cable for all data transmission until a repair can be made. Selected servers also include facilities for I/O drawer or tower concurrent additions (while the system continues to operate) and allow the drawer or tower to be varied on or off-line. Using these features, a failure in an I/O drawer or tower that is configured for availability (I/O devices accessed through the drawer must not be defined as *required* for a partition boot or, for IBM i partitions, ring level or tower level mirroring has been implemented) can be repaired while the main server continues to operate.

GX+ bus adapters

The GX+ bus provides the primary high bandwidth path for RIO-2 or GX 12x Dual Channel adapter connection to the system CEC. Errors in a GX+ bus adapter, flagged by system *persistent deallocation* logic, cause the adapter to be varied offline upon a server reboot.

PCI error recovery

IBM estimates that PCI adapters can account for a significant portion of the hardware based errors on a large server. Although servers that rely on boot time diagnostics can identify failing components to be replaced by hot-swap and reconfiguration, run time errors pose a more significant problem.

PCI adapters are generally complex designs involving extensive onboard instruction processing, often on embedded microcontrollers. They tend to use industry standard grade components with an emphasis on product cost relative to high reliability. In some cases, they might be more likely to encounter internal microcode errors, and many of the hardware errors described for the rest of the server.

The traditional means of handling these problems is through adapter internal error reporting and recovery techniques in combination with operating system device driver management and diagnostics. In some cases, an error in the adapter might cause transmission of bad data on the PCI bus itself, resulting in a hardware detected parity error and causing a global machine check interrupt, eventually requiring a system reboot to continue.

In 2001, IBM introduced a methodology that uses a combination of system firmware and Enhanced Error Handling (EEH) device drivers that allows recovery from intermittent PCI bus errors. This approach works by recovering and resetting the adapter, thereby initiating system recovery for a permanent PCI bus error. Rather than failing immediately, the faulty device is frozen and restarted, preventing a machine check. POWER6 and POWER5 processor servers extend the capabilities of the EEH methodology. Generally, all PCI adapters controlled by operating system device drivers are connected to a PCI secondary bus created through a PCI-to-PCI bridge, designed by IBM. This bridge isolates the PCI adapters and supports *hot-plug* by allowing program control of the *power state* of the I/O slot. PCI bus errors related to individual PCI adapters under partition control can be transformed into a PCI slot freeze condition and reported to the EEH device driver for error handling. Errors that occur on the interface between the PCI-to-PCI bridge chip and the Processor Host Bridge (the link between the processor remote I/O bus and the primary PCI bus) result in a *bridge freeze* condition, effectively stopping all of the PCI adapters attached to the bridge chip. An operating system can recover an adapter from a bridge freeze condition by using POWER Hypervisor functions to remove the bridge from freeze state and resetting or reinitializing the adapters. This same EEH technology will allow system recovery of PCIe bus errors in POWER6 processor-based servers.

Figure 4-6 on page 145 illustrates PCI error recovery in GX bus connections.

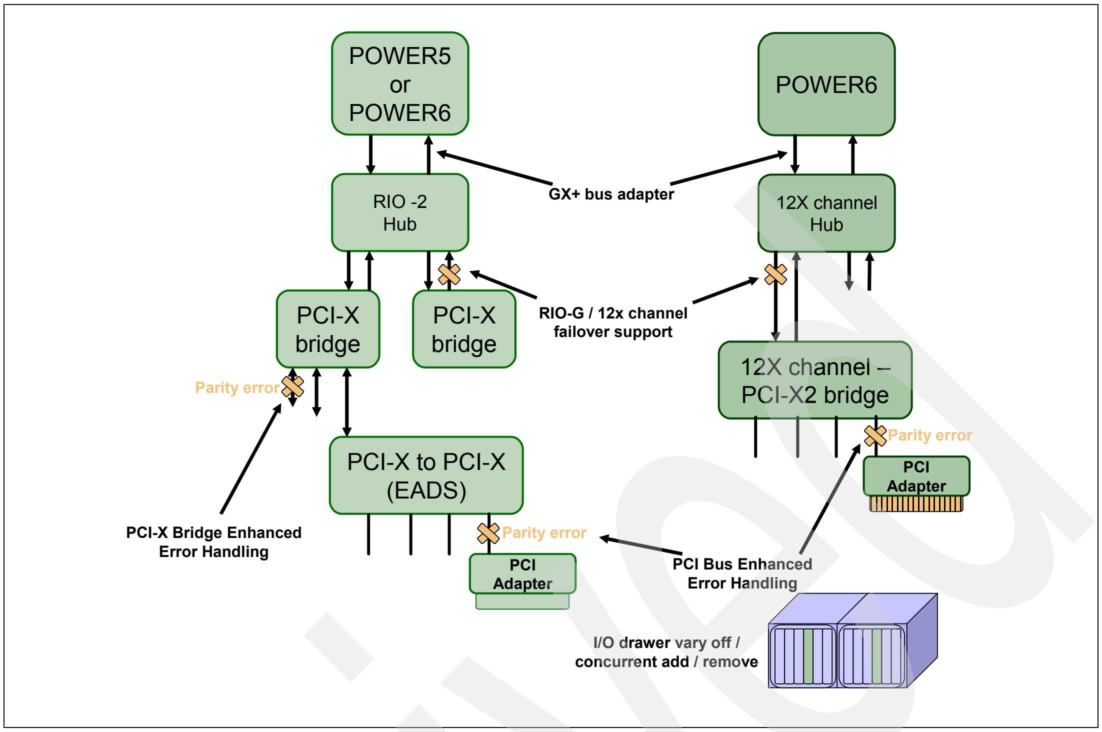


Figure 4-6 PCI error recovery

4.2.5 Redundant components and concurrent repair update

High opportunity components, or those that most affect system availability, are protected with redundancy and the ability to be repaired concurrently.

Power 595

Equipped with ultrahigh frequency IBM POWER6 processors in up to 64-core, multiprocessing (SMP) configurations, the Power 595 server can scale rapidly and seamlessly to address the changing needs of today's data center. With advanced PowerVM virtualization, EnergyScale technology, and Capacity on Demand (CoD) options, the Power 595 helps businesses take control of their IT infrastructure and confidently consolidate multiple UNIX, IBM i (formerly known as i5/OS), and Linux application workloads onto a single system.

Extensive mainframe inspired reliability, availability, and serviceability (RAS) features in the Power 595 help ensure that mission critical applications run reliably around the clock. The 595 is equipped with a broad range of standard redundancies for improved availability:

- Bulk power assemblies (BPA) and line cords. All components internal to the BPA, BPC, BPD, BPR, BPH, BPF are redundant. (active redundancy, hot replace)
- CEC cage:
  - System controller, oscillator card (hot failover)
  - VPD card (active redundancy
- Processor books:
  - Node controller (hot failover)
  - Hot RIO/GX adapter add

  - Vital Product Data and CoD modules (active redundancy)
  - DCAs (active redundancy)
  - Voltage regulator modules (active redundancy, hot replace)
- Blowers (active redundancy, hot replace)
- All out of band service interfaces (active redundancy)
- System Ethernet hubs (active redundancy)
- All LED indicator drive circuitry (active redundancy)
- Thermal sensors (active redundancy)

Additional features can support enhanced availability:

- Concurrent firmware update
- I/O drawers with dual internal controllers
- Hot add/repair of I/O drawers.
- Light strip with redundant, active failover circuitry.
- Hot RIO/GX adapter add

> **Note:** Statement of general direction.
>
> IBM is committed to enhancing client investments in Power Systems servers. Based on this commitment, IBM plans to provide the following enhancements:
>
> - The capability to add additional processor books to POWER6 process-based 595 systems without powering down (hot node *add*)
> - The capability to reactivate a POWER6 595 processor book that has been repaired without powering down (cold node *repair*)
> - The capability to deactivate, repair components or add memory, and then reactivate a POWER6 595 processor book without powering down (hot node *repair*).
>
> These capabilities are planned to be provided at no additional charge to POWER6 Power 595 and Power 570 clients via a system firmware upgrade by the end of 2008.
>
> All statements regarding IBMs plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

POWER Hypervisor

Because the availability of the POWER Hypervisor is crucial to overall system availability, great care has been taken to design high quality, well-tested code. In general, a hardware system sees a higher than normal error rate when first introduced and when first installed in production. These types of errors are mitigated by strenuous engineering and manufacturing verification testing and by using methodologies such as *burn in*, designed to catch the fault before the server is shipped. At this point, hardware failures typically even out at relatively low, but constant, error rates. This phase can last for many years. At some point, however, hardware failures can again increase as parts begin to *wear out*. Clearly, the *design for availability* techniques discussed here will help mitigate these problems.

Coding errors are significantly different from hardware errors. Unlike hardware, code can display a variable rate of failure. New code typically has a higher failure rate and older more seasoned code a very low rate of failure. Code quality will continue to improve as bugs are discovered and fixes installed. Although the POWER Hypervisor provides important system functions, it is limited in size and complexity when compared to a full operating system implementation, and therefore can be considered better *contained* from a design and quality assurance viewpoint. As with any software development project, the IBM firmware development team writes code to strict guidelines using well-defined software engineering methods. The overall code architecture is reviewed and approved and each developer schedules a variety of peer code reviews. In addition, all code is strenuously tested, first by visual inspections, looking for logic errors, then by simulation and operation in actual test and production servers. Using this structured approach, most coding error are caught and fixed early in the design process.

An inherent feature of the POWER Hypervisor is that the majority of the code runs in the protection domain of a hidden system partition. Failures in this code are limited to this system partition. Supporting a very robust tasking model, the code in the system partition is segmented into critical and noncritical tasks. If a noncritical task fails, the system partition is designed to continue to operate, albeit without the function provided by the failed task. Only in a rare instance of a failure to a critical task in the system partition would the entire POWER Hypervisor fail.

The resulting code provides not only advanced features but also superb reliability. It is used in IBM Power Systems and in the IBM TotalStorage DS8000 series products. It has therefore been strenuously tested under a wide ranging set of system environments and configurations. This process has delivered a quality implementation that includes enhanced error isolation and recovery support when compared to POWER4 process-based offerings.

Service processor and clocks

A number of availability improvements have been included in the service processor in the POWER6 and POWER5 processor-based servers. Separate copies of service processor microcode and the POWER Hypervisor code are stored in discrete flash memory storage areas. Code access is CRC protected. The service processor performs low level hardware initialization and configuration of all processors. The POWER Hypervisor performs higher level configuration for features like the virtualization support required to run up to 254 partitions concurrently on the POWER6 595, p5-590, p5-595, and i5-595 servers. The POWER Hypervisor enables many advanced functions; including sharing of processor cores, virtual I/O, and high speed communications between partitions using Virtual LAN. AIX, Linux, and IBM i are supported. The servers also support dynamic firmware updates, in which applications remain operational while IBM system firmware is updated for many operations. Maintaining two copies ensures that the service processor can run even if a Flash memory copy becomes corrupted, and allows for redundancy in the event of a problem during the upgrade of the firmware.

In addition, if the service processor encounters an error during runtime, it can reboot itself while the server system stays up and running. No server application impact exists for service processor transient errors. If the service processor encounters a code *hang* condition, the POWER Hypervisor can detect the error and direct the service processor to reboot, avoiding other outage

Each POWER6 processor chip is designed to receive two oscillator signals (clocks) and can be enabled to switch dynamically from one signal to the other. POWER6 595 servers are equipped with two clock cards. For the POWER6 595, failure of a clock card will result in an automatic (runtime) failover to the secondary clock card. No reboot is required. For other multiclock offerings, an IPL time failover occurs if a system clock fails.

4.2.6 Availability in a partitioned environment

IBMs dynamic logical partitioning architecture has been extended with micro-partitioning technology capabilities. These new features are provided by the POWER Hypervisor and are configured using management interfaces on the HMC. This very powerful approach to partitioning maximizes partitioning flexibility and maintenance. It supports a consistent partitioning management interface just as applicable to single (full server) partitions as to systems with hundreds of partitions.

In addition to enabling fine-grained resource allocation, these LPAR capabilities provide all the servers in the POWER6 and POWER5 processor models the underlying capability to individually assign any resource (processor core, memory segment, I/O slot) to any partition in any combination. Not only does this allow exceptional configuration flexibility, it enables many high availability functions like:

- Resource sparing (dynamic processor deallocation and dynamic processor sparing).
- Automatic redistribution of capacity on N+1 configurations (automated shared pool redistribution of partition entitled capacities for dynamic processor sparing).
- LPAR configurations with redundant I/O (across separate processor host bridges or even physical drawers) allowing system designers to build configurations with improved redundancy for automated recovery.
- The ability to reconfigure a server *on the fly*. Because any I/O slot can be assigned to any partition, a system administrator can *vary off* a faulty I/O adapter and *back fill* with another available adapter, without waiting for a spare part to be delivered for service.
- Live Partition Mobility provides the ability to move running partitions from one POWER6 process-based server to another (refer to section “PowerVM Live Partition Mobility” on page 125).
- Automated scaleup of high availability backup servers as required (through dynamic LPAR).
- Serialized sharing of devices (optical, tape) allowing *limited* use devices to be made available to all the partitions.
- Shared I/O devices through I/O server partitions. A single I/O slot can carry transactions on behalf of several partitions, potentially reducing the cost of deployment and improving the speed of provisioning of new partitions (new applications). Multiple I/O server partitions can be deployed for redundancy, giving partitions multiple paths to access data and improved availability in case of an adapter or I/O server partition outage.

In a logically partitioning architecture, all server memory is physically accessible to all processor cores and all I/O devices in the system, regardless of physical placement of the memory or where the logical partition operates. The POWER Hypervisor mode with real memory offset facilities enables the POWER Hypervisor to ensure that any code running in a partition (operating systems and firmware) only has access to the physical memory allocated to the dynamic logical partition. POWER6 and POWER5 processor systems also have IBM-designed PCI-to-PCI bridges that enable the POWER Hypervisor to restrict direct memory access (DMA) from I/O devices to memory owned by the partition using the device. The single memory cache coherency domain design is a key requirement for delivering the highest levels of SMP performance. Because it is IBM’s strategy to deliver hundreds of dynamically configurable logical partitions, allowing improved system utilization and reducing overall computing costs, these servers must be designed to avoid or minimize conditions that would cause a full server outage.

IBM’s availability architecture provides a high level of protection to the individual components making up the memory coherence domain, including the memory, caches, and fabric bus. It also offers advanced techniques designed to help contain failures in the coherency domain to a subset of the server. Through careful design, in many cases, failures are contained to a component or to a partition, despite the shared hardware system design. Many of these techniques have been described in this document.

System-level availability (in any server, no matter how partitioned) is a function of the reliability of the underlying hardware and the techniques used to mitigate the faults that do occur. The availability design of these systems minimizes system failures and localizes potential hardware faults to single partitions in multi-partition systems. In this design, although some hardware errors might cause a full system crash (causing loss of all partitions), because the rate of system crashes is very low, the rate of partition crashes is also very low.

The reliability and availability characteristics described in this document show how this *design for availability* approach is consistently applied throughout the system design. IBM believes this is the best approach to achieving partition level availability while supporting a truly flexible and manageable partitioning environment.

In addition, to achieve the highest levels of system availability, IBM and third party software vendors offer clustering solutions (such as HACMP™), which allow for failover from one system to another, even for geographically dispersed systems.

4.2.7 Operating system availability

The focus of this section is a discussion of RAS attributes in the POWER6 and POWER5 hardware to provide for availability and serviceability of the hardware itself. Operating systems, middleware, and applications provide additional key features concerning their own availability that is outside the scope of this hardware discussion.

A worthwhile note, however, is that hardware and firmware RAS features can provide key enablement for selected software availability features. As can be seen in section 4.4, “Operating system support for RAS features” on page 160, many RAS features described in this document are applicable to all supported operating systems.

The AIX, IBM i, and Linux operating systems include many reliability features inspired by IBMs mainframe technology designed for robust operation. In fact, clients in survey, have selected AIX as the highest quality UNIX operating system. In addition, IBM i offers a highly scalable and virus resistant architecture with a proven reputation for exceptional business resiliency. IBM i integrates a trusted combination of relational database, security, Web services, networking and storage management capabilities. It provides a broad and highly stable database and middleware foundation. All core middleware components are developed, tested, and preloaded together with the operating system.

AIX 6 introduces unprecedented continuous availability features to the UNIX market designed to extend its leadership continuous availability features.

POWER6 servers support a variety of enhanced features:

- POWER6 storage protection keys

  These keys provide hardware-enforced access mechanisms for memory regions. Only programs that use the correct key are allowed to read or write to protected memory locations. This new hardware allows programmers to restrict memory access within well defined, hardware enforced boundaries, protecting critical portions of AIX 6 and applications software from inadvertent memory overlay.

  Storage protection keys can reduce the number of intermittent outages associated with undetected memory overlays inside the AIX kernel. Programmers can also use the

POWER6 memory protection key feature to increase the reliability of large, complex applications running under the AIX V5.3 or AIX 6 releases.

- Concurrent AIX kernel update

  These updates allow installation of some kernel patches without rebooting the system. This can reduce the number of unplanned outages required to maintain a secure, reliable system.

- Dynamic tracing

  This facility can simplify debugging of complex system or application code. Using a new tracing command, **probevue**, developers or system administrators can dynamically insert trace breakpoints in existing code without having to recompile which allows them to more easily troubleshoot application and system problems.

- Enhanced software first failure data capture (FFDC)

  AIX V5.3 introduced FFDC technology to gather diagnostic information about an error at the time the problem occurs. Like hardware-generated FFDC data, this allows AIX to quickly and efficiently diagnose, isolate, and in many cases, recover from problems and reducing the need to recreate the problem (and impact performance and availability) simply to generate diagnostic information. AIX 6 extends the FFDC capabilities, introducing more instrumentation to provide real time diagnostic information.

4.3 Serviceability

The IBM POWER6 Serviceability strategy evolves from, and improves upon, the service architecture deployed on the POWER5 processor systems. The IBM service team has enhanced the base service capabilities and continues to implement a strategy that incorporates best-of-breed service characteristics from IBMs diverse System x, System i, System p, and high-end System z™ offerings.

The goal of the IBM Serviceability Team is to design and provide the most efficient system service environment that incorporates:

- Easy access to service components
- On demand service education
- An automated guided repair strategy that uses common service interfaces for a converged service approach across multiple IBM server platforms

By delivering on these goals, POWER6 processor systems enable faster and more accurate repair while reducing the possibility of human error.

Client control of the service environment extends to firmware maintenance on all of the POWER6 process-based systems. This strategy contributes to higher systems availability with reduced maintenance costs.

This section provides an overview of the progressive steps of error detection, analysis, reporting, and repairing found in all POWER6 process-based systems.

4.3.1 Service environments

The IBM POWER5 and POWER6 processor platforms support three main service environments:

- Servers that do not include a Hardware Management Console (HMC). This is the manufacturing default configuration for from two operational environments:
  - Standalone full system partition: The server can be configured with a single partition that owns all server resources and has only one operating system installed.
  - Non-HMC partitioned system: For selected Power Systems servers, the optional PowerVM feature includes Integrated Virtualization Manager (IVM), which is a browser-based system interface for managing servers without an attached HMC. Multiple logical partitions can be created, each with its own operating environment. All I/O is virtualized and shared.

    An analogous feature, the Virtual Partition Manager (VPM), is included with IBM i (5.3 and later), and supports the needs of small and medium clients who want to add simple Linux workloads to their System i5 or Power server. The VPM introduces the capability to create and manage Linux partitions without the use of the HMC. With the VPM, a server can support one i partition and up to four Linux partitions. The Linux partitions must use virtual I/O resources that are owned by the i partition.
- Server configurations that include attachment to one or multiple HMCs. This is the default configuration for high-end systems and servers supporting logical partitions with dedicated I/O. In this case, all servers have at least one logical partition.
- Mixed environments of POWER6 and POWER5 processor-based systems controlled by one or multiple HMCs for POWER6 technologies. This HMC can simultaneously manage POWER6 and POWER5 process-based systems. An HMC for a POWER5 process-based server, with a firmware upgrade, can support this environment.

4.3.2 Service processor

The service processor is a separately powered microprocessor, separate from the main instruction processing complex. The serv selected remote power control, environmental monitoring, reset and boot features, remote maintenance and diagnostic activities, including console mirroring. On systems without a HMC, the service processor can place calls to report surveillance failures with the P and critical processing faults even when the main processing unit is inoperable. The service processor provides services common to modern computers such as:

- Environmental monitoring
  - The service processor monitors the server’s built-in temperature sensors, sending instructions to the system fans to increase rotational speed when the ambient temperature is above the normal operating range.
  - Using an architected operating system interface, the service processor notifies the operating system of potential environmental related problems (for example, air conditioning and air circulation around the system) so that the system administrator can take appropriate corrective actions before a critical failure threshold is reached.
  - The service processor can also post a warning and initiate an orderly system shutdown for a variety of other conditions:
    - When the operating temperature exceeds the critical level (for example failure of air conditioning or air circulation around the system).

        - When the system fan speed is out of operational specification, for example because of a fan failure, the system can increase speed on the redundant fans to compensate for this failure or take other actions
        - When the server input voltages are out of operational specification.

- Mutual Surveillance
    - The service processor monitors the operation of the POWER Hypervisor firmware during the boot process and watches for loss of control during system operation. It also allows the POWER Hypervisor to monitor service processor activity. The service processor can take appropriate action, including calling for service, when it detects the POWER Hypervisor firmware has lost control. Likewise, the POWER Hypervisor can request a service processor repair action if necessary.
- Availability
    - The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced (ac power) failure.
- Fault Monitoring
    - Built-in self-test (BIST) checks processor, memory, and associated hardware required for proper booting of the operating system, when the system is powered on at the initial install or after a hardware configuration change (such as an upgrade). If a non-critical error is detected or if the error occurs in a resource that can be removed from the system configuration, the booting process is designed to proceed to completion. The errors are logged in the system nonvolatile random access memory (NVRAM). When the operating system completes booting, the information is passed from the NVRAM into the system error log where it is analyzed by error log analysis (ELA) routines. Appropriate actions are taken to report the boot time error for subsequent service if required.

One important service processor improvement allows the system administrator or service representative dynamic access to the Advanced Systems Management Interface (ASMI) menus. In previous generations of servers, these menus were only accessible when the system was in standby power mode. Now, the menus are available from any Web browser enabled console attached to the Ethernet service network concurrent with normal system operation. A user with the proper access authority and credentials can now dynamically modify service defaults, interrogate service processor progress and error logs, set and reset Guiding Light LEDs, and access all service processor functions without having to power down the system to the standby state.

The service processor also manages the interfaces for connecting uninterruptible power source systems to the POWER6 process-based systems, performing Timed Power On (TPO) sequences, and interfacing with the power and cooling subsystem.

4.3.3 Detecting errors

The first and most crucial component of a solid serviceability strategy is the ability to accurately and effectively detect errors when they occur. While not all errors are a guaranteed threat to system availability, those that go undetected can cause problems because the system does not have the opportunity to evaluate and act if necessary. POWER6 process-based systems employ System z server inspired error detection mechanisms that extend from processor cores and memory to power supplies and hard drives.

Error checkers

IBM POWER6 process-based systems contain specialized hardware detection circuitry that can detect erroneous hardware operations. Error checking hardware ranges from parity error detection coupled with processor instruction retry and bus retry, to ECC correction on caches and system buses. All IBM hardware error checkers have distinct attributes, as follows:

- Continually monitoring system operations to detect potential calculation errors.
- Attempting to isolate physical faults based on runtime detection of each unique failure.
- Initiating a wide variety of recovery mechanisms designed to correct the problem. The POWER6 process-based systems include extensive hardware and firmware recovery logic.

Fault isolation registers

Error checker signals are captured and stored in hardware Fault Isolation Registers (FIRs). The associated *who's on first* logic circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, runtime error diagnostics can be deterministic so that for every check station, the unique error domain for that checker is defined and documented. Ultimately, the error domain becomes the field replaceable unit (FRU) call, and manual interpretation of the data is not normally required.

First failure data capture (FFDC)

First failure data capture (FFDC) is an error isolation technique, which ensures that when a fault is detected in a system through error checkers or other types of detection methods, the root cause of the fault gets captured without the need to recreate the problem or run an extended tracing or diagnostics program.

For the vast majority of faults, a good FFDC design means that the root cause can be detected automatically without intervention of a service representative. Pertinent error data related to the fault is captured and saved for analysis. In hardware, FFDC data is collected from the fault isolation registers and *who's on first* logic. In firmware, this data consists of return codes, function calls, and others.

FFDC *check stations* are carefully positioned within the server logic and data paths to ensure that potential errors can be quickly identified and accurately tracked to an FRU.

This proactive diagnostic strategy is a significant improvement over the classic, less accurate *reboot and diagnose* service approaches.

Figure 4-7 on page 154 shows a schematic of a fault isolation register implementation.

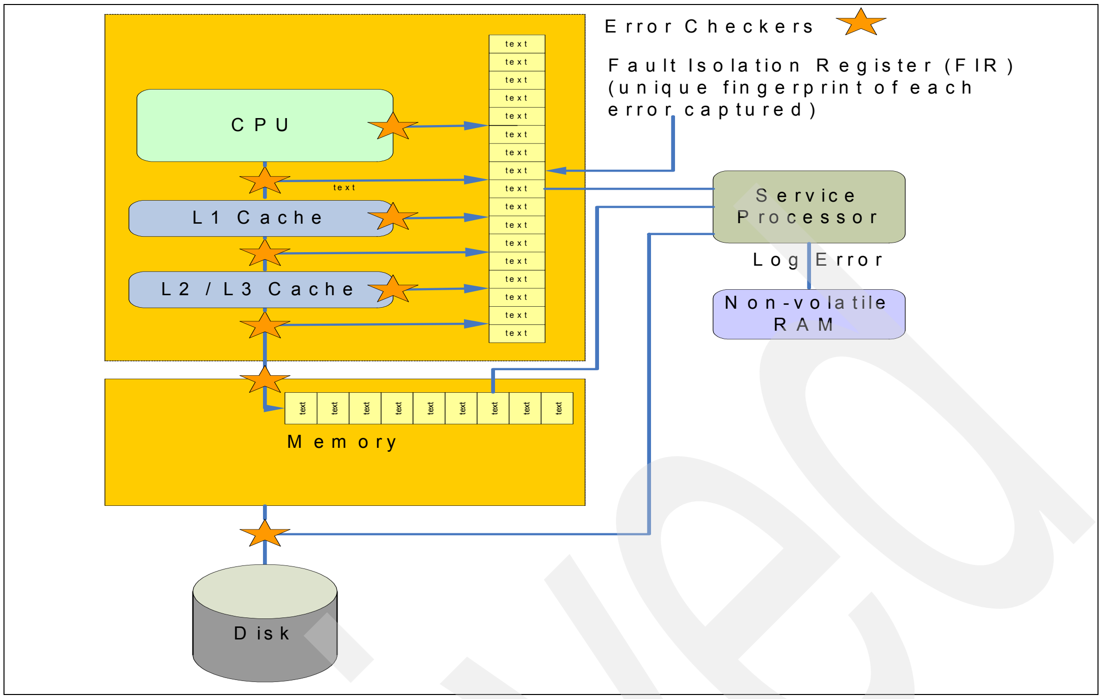


Figure 4-7 Schematic of an FIR implementation

Fault isolation

The service processor interprets error data captured by the FFDC checkers, which is saved in the FIRs and *who's on first* logic or other firmware-related data capture methods, to determine the root cause of the error event.

Root cause analysis can indicate that the event is recoverable, meaning that a service action point or need for repair has not been reached. Alternatively, it could indicate that a service action point has been reached, where the event exceeded a predetermined threshold, or was unrecoverable. Based on the isolation analysis, recoverable error threshold counts can be incremented. No specific service action is necessary when the event is recoverable.

When the event requires a service action, additional required information is collected to service the fault. For unrecoverable errors or for recoverable events that meet or exceed their service threshold—meaning a service action point has been reached—a request for service is initiated through an error logging component.

4.3.4 Diagnosing problems

Using the extensive network of advanced and complementary error detection logic built directly into hardware, firmware, and operating systems, the IBM POWER6 processor systems can perform considerable self diagnosis.

Boot time

When an IBM POWER6 processor system powers up, the service processor initializes system hardware. Boot-time diagnostic testing uses a multi-tier approach for system validation, starting with managed low-level diagnostics supplemented with system firmware initialization and configuration of I/O hardware, followed by OS initiated software test routines. Boot-time diagnostic routines include:

- BISTs for both logic components and arrays ensure the internal integrity of components. Because the service processor assist in performing these tests, the system is enabled to perform fault determination and isolation whether system processors are operational or not. Boot-time BISTs can also find faults undetectable by process-based power-on self-test (POST) or diagnostics.
- Wire tests discover and precisely identify connection faults between components such as processors, memory, or I/O hub chips.
- Initialization of components such as ECC memory, typically by writing patterns of data and allowing the server to store valid ECC data for each location, can help isolate errors.

To minimize boot time, the system determines which of the diagnostics are required to be started to ensure correct operation based on the way the system was powered off, or on the boot-time selection menu.

Runtime

All POWER6 processor systems can monitor critical system components during runtime, and they can take corrective actions when recoverable faults occur. IBM's hardware error check architecture provides the ability to report non-critical errors in an *out-of-band* communications path to the service processor without affecting system performance.

A significant part of IBM's runtime diagnostic capabilities originate with the POWER6 service processor. Extensive diagnostic and fault analysis routines have been developed and improved over many generations of POWER process-based servers, and enable quick and accurate predefined responses to both actual and potential system problems.

The service processor correlates and processes runtime error information, using logic derived from IBM's engineering expertise, to count recoverable errors (called *thresholding*) and predict when corrective actions must be automatically initiated by the system. These actions can include:

- Requests for a part to be replaced
- Dynamic (online) invocation of built-in redundancy for automatic replacement of a failing part
- Dynamic deallocation of failing components so that system availability is maintained

Device drivers

In certain cases, diagnostics are best performed by operating system-specific drivers, most notably I/O devices that are owned directly by a logical partition. In these cases, the operating system device driver often works in conjunction with I/O device microcode to isolate and recover from problems. Potential problems are reported to an operating system device driver, which logs the error. I/O devices can also include specific exercisers that can be invoked by the diagnostic facilities for problem recreation if required by service procedures.

4.3.5 Reporting problems

In the unlikely event that a system hardware failure or an environmentally induced failure is diagnosed, POWER6 processor systems report the error through a number of mechanisms. This ensures that appropriate entities are aware that the system can be operating in an error state. However, a crucial piece of a solid reporting strategy is ensuring that a single error communicated through multiple error paths is correctly aggregated, so that later notifications are not accidently duplicated.

Error logging and analysis

After the root cause of an error has been identified by a fault isolation component, an error log entry is created and that includes basic data such as:

- An error code uniquely describing the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data like engineering and manufacturing levels
- Return codes
- Resource identifiers
- FFDC data

Data containing information about the effecte repair can have on the system is also included. Error log routines in the operating system can tthis information and decide to call home to contact service and support, send a notification message, or continue without an alert.

Remote support

The Remote Management and Control (RMC) application is delivered as part of the base operating system, including the operating system running on the HMC. The RMC provides a secure transport mechanism across the LAN interface between the operating system and the HMC and is used by the operating system diagnostic application for transmitting error information. It performs a number of other functions as well, but these are not used for the service infrastructure.

Manage serviceable events

A critical requirement in a logically partitioned environment is to ensure that errors are not lost before being reported for service, and that an error should only be reported once, regardless of how many logical partitions experience the potential effect of the error. The **Manage Serviceable Events** task on the HMC is responsible for aggregating duplicate error reports, and ensuring that all errors are recorded for review and management.

When a local or globally-reported service request is made to the operating system, the operating system diagnostic subsystem uses the RMC subsystem to relay error information to the HMC. For global events (platform unrecoverable errors, for example) the service processor will also forward error notification of these events to the HMC, providing a redundant error-reporting path in case of errors in the RMC network.

The first occurrence of each failure type will be recorded in the **Manage Serviceable Events** task on the HMC. This task then filters and maintains a history of duplicate reports from other logical partitions or the service processor. It then looks across all active service event requests, analyzes the failure to ascertain the root cause, and, if enabled, initiates a call home for service. This method ensures that all platform errors will be reported through at least one functional path, ultimately resulting in a single notification for a single problem.

Extended error data (EED)

Extended error data (EED) is additional data collected either automatically at the time of a failure or manually at a later time. The data collected depends on the invocation method but includes information like firmware levels, operating system levels, additional fault isolation register values, recoverable error threshold register values, system status, and any other pertinent data.

The data is formatted and prepared for transmission back to IBM to assist the service support organization with preparing a service action plan for the service representative or for additional analysis.

System dump handling

In some circumstances, an error might require a dump to be automatically or manually created. In this event, it is offloaded to the HMC during reboot. Specific HMC information is included as part of the information that can optionally be sent to IBM support for analysis. If additional information relating to the dump is required, or if it beew the dump remotely, the HMC dump record notifies the IBM support center regarding on which HMC the dump is located.

4.3.6 Notifying the appropriate contacts

After a POWER6 processor-based system has detected, diagnosed, and reported an error to an appropriate aggregation point, it then takes steps to notify you, and if necessary the IBM Support Organization. Depending on the assessed severity of the error and support agreement, this notification could range from a simple notification to having field service personnel automatically dispatched to the client site with the correct replacement part.

Customer notify

When an event is important enough to report, but does not indicate the need for a repair action or the need to call home to IBM service and support, it is classified as *customer notify.* Customers are notified because these events can be of interest to an administrator. The event might be a symptom of an expected systemic change, such as a network reconfiguration or failover testing of redundant power or cooling systems. Examples of these events include:

- Network events such as the loss of contact over a LAN
- Environmental events such as ambient temperature warnings
- Events that need further examination by the client. These events, however, do not necessarily require a part replacement or repair action

Customer notify events are serviceable events by definition because they indicate that something has happened, which requires you to be aware in case you want to take further action. These events can always be reported back to IBM at your discretion.

t location to the IBM service and support organization with error data, server status, or other service related information. *Call home* invokes the service organization so that the appropriate service action can begin, automatically opening a problem report and in some cases also dispatching field support. This automated reporting provides faster and potentially more accurate transmittal of error information. Although configuring call home is optional, clients are strongly encouraged to configure this feature to obtain the full value of IBM service enhancements.

Vital Product Data (VPD) and inventory management

POWER6 process-based systems store vital product data (VPD) internally, which keeps a record of how much memory is installed, how many proare installed, manufacturing level of the parts, and so on. These records provide valuable information that can be used by remote support and service representatives, enabling them to provide assistance in keeping the firmware and software on the server up to date.

IBM problem management database

At the IBM Support Center, historical problem data is entered into the IBM Service and Support Problem Management database. All information related to the error and any service actions taken by the service representative are recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

4.3.7 Locating and repairing the problem

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER6 process-based systems utilize a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Guiding Light LEDs

Guiding Light uses a series of flashing LEDs, allowing a service provider to quickly and easily identify the location of system components. Guiding Light can also handle multiple error conditions simultaneously, which could be necessary in some very complex high-end configurations.

In the Guiding Light LED implementation, when a fault condition is detected on a POWER6 processor system, an amber System Attention LED is illuminated. Upon arrival, the service provider engages the *identify* mode by selecting a specific problem. The Guiding Light system then identifies the part that needs to be replaced by flashing the amber *identify* LED.

Datacenters can be complex places, and Guiding Light is designed to do more than identify visible components. When a component might be hidden from view, Guiding Light can flash a sequence of LEDs that extend to the frame exterior, clearly “guiding” the service representative to the correct rack, system, enclosure, drawer, and component.

The operator panel

The operator panel on a POWER6 process-based system is a four-row by 16-element LCD display used to present boot progress codes, indicating advancement through the system power-on and initialization processes. The operator panel is also used to display error and location codes when an error occurs that prevents the system from booting. The operator panel includes several buttons allowing a service representative or the client to change various boot-time options, and perform a subset of the service functions that are available on the ASMI.

Concurrent maintenance

The IBM POWER6 processor-based systems are designed with the understanding that certain components have higher intrinsic failure rates than others. The movement of fans, power supplies, and physical storage devices naturally make them more susceptible to wear or burnout, while other devices such as I/O adapters might begin to wear from repeated plugging or unplugging. For this reason, these devices are specifically designed to be concurrently maintainable, when properly configured.

In other cases, you might be in the process of moving or redesigning a datacenter, or planning a major upgrade. At times like these, flexibility is crucial. The IBM POWER6 process-based systems are designed for redundant or concurrently maintainable power, fans, physical storage, and I/O towers.

Blind-swap PCI adapters

Blind-swap PCI adapters represent significant service and ease-of-use enhancements in I/O subsystem design while maintaining high PCI adapter density.

Standard PCI designs supporting *hot-add* and *hot-replace* require top access so that adapters can be slid into the PCI I/O slots vertically. *Blind-swap* allows PCI adapters to be concurrently replaced without having to put the I/O drawer into a service position.

Firmware updates

Firmware updates for POWER6 processor-based systems are released in a cumulative sequential fix format, packaged as an RPM for concurrent application and activation. Administrators can install and activate many firmware patches without cycling power or rebooting the server.

The new firmware image is loaded on the HMC using any of the following methods:

- IBM distributed media such as a CD-ROM
- A Problem Fix distribution from the IBM service and support repository
- Download from the IBM Microcode Web site:

  `http://www14.sosn`
- FTP from another server

IBM supports multiple firmware releaso under expected circumstances, a server can operate on an existing firmware release, using concurrent firmware fixes to remain updated with the current patch level. Because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during system operation, a patch in this area will require a system reboot for activation. Under normal operating conditions, IBM intends to provide patches for an individual firmware release level for up to two years after first making the release code generally availability. After this period, you should plan to update so that you can remain on a supported firmware release.

Activation of new firmware functions, as opposed to patches, will require installation of a new firmware release level. This process is disruptive to server operations because it requires a scheduled outage and full server reboot.

In addition to concurrent and disruptive firmware updates, IBM also offers cohat include functions which are not activated until a subsequent server reboot. A server with these patches operates normally. The additional concurrent fixes are installed and activated when the system reboots after the next scheduled outage.

> **Note:** Additional capability is being added to the POWER6 firmware to be able to view the status of a system power control network background firmware update. This subsystem will update as necessary, as migrated nodes or I/O drawers are added to the configuration. The new firmware will not only provide an interface to be able to view the progress of the update, but also control starting and stopping of the background update if a more convenient time becomes available.

Repair and Verify

Repair and Verify (R&V) is a system used to guide a service provider, step by step, through the process of repairing a system and verifying that the problem has been repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being repaired. Repair scenarios covered by repair and verify include:

- Replacing a defective FRU
- Reattaching a loose or disconnected component
- Correcting a configuration error
- Removing or replacing an incompatible FRU
- Updating firmware, device drivers, operating systems, middleware components, and IBM applications after replacing a part
- Installing a new part

R&V procedures are designed to be used by service representative providers who are familiar with the task and those who are not. On demand education content is placed in the procedure at the appropriate locations. Throughout the repair and verify procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event, to ensure that the guided maintenance procedures are operating correctly.

Service documentation on the support for IBM System p

The support for IBM System p Web site is an electronic information repository for POWER6 processor-based systems. This Web site provides online training, educational material, documentation, and service procedures that are not handled by the automated R&V guided component.

The Support for System p Web site is located at:

http://www.ibm.com/systems/support/p

You may subscribe through the Subscription Services to obtain the notifications on the latest updates available for service related documentation. The latest version of the documentation is accessible through the Internet, and a CD-ROM based version is also available.

4.4 Operating system support for RAS features

Table 4-1 lists a number of features for continuous availability supported by the different operating systems runnin POWER6 process-based systems.

Table 4-1 Operating system support for selected RAS features

RAS feature	AIX V5.3	AIX V6.1	IBM i 5.4 LIC 5.4.5	IBM i 6.1	RHEL V5.1	SLES V10
System Deallocation of Failing Components						
Dynamic processor deallocation	✓	✓	✓	✓	✓[a]	✓
Dynamic processor sparing	✓	✓	✓	✓	✓	✓
Processor instruction retry	✓	✓	✓	✓	✓	✓
Alternate processor recovery	✓	✓	✓	✓	✓	✓
Partition contained checkstop	✓	✓	✓	✓	✓	✓
Persistent processor deallocation	✓	✓	✓	✓	✓	✓

RAS feature	AIX V5.3	AIX V6.1	IBM i 5.4 LIC 5.4.5	IBM i 6.1	RHEL V5.1	SLES V10
GX+ bus persistent deallocation	✓	✓	✓	✓	—	—
PCI bus extended error detection	✓	✓	✓	✓	✓	✓
PCI bus extended error recovery	✓	✓	✓	✓	Limited[a]	Limited
PCI-PCI bridge extended error handling	✓	✓	✓	✓	✓	✓
Redundant RIO Link	✓	✓	✓	✓	✓	✓
PCI card hot swap	✓	✓	✓	✓	✓[a]	✓
Dynamic SP failover at runtime	✓	✓	✓	✓		
Clock failover at IPL	✓	✓	✓	✓	✓	✓
Memory Availability						
ECC Memory, L2 cache	✓	✓	✓	✓	✓	✓
Dynamic bit-steering (spare memory)	✓	✓	✓	✓	✓	✓
Memory scrubbing	✓	✓	✓	✓	✓	✓
Chipkill memory	✓	✓	✓	✓	✓	✓
Memory page deallocation	✓	✓	✓	✓	—	—
L1 parity check plus retry	✓	✓	✓	✓	✓	✓
L2 cache line delete	✓	✓	✓	✓	✓	✓
Array recovery & Array persistent deallocation (spare bits in L1 & L2 cache; L1, L2 directory)	✓	✓	✓	✓	✓	✓
Special uncorrectable error handling	✓	✓	✓	✓	✓	✓
Fault Detection and Isolation						
Platform FFDC diagnostics	✓	✓	✓	✓	✓	✓
I/O FFDC diagnostics	✓	✓	✓	✓	—	✓
Runtime diagnostics	✓	✓	✓	✓	Limited	Limited
Storage protection keys	✓	✓	✓	✓	—	—
Dynamic trace	—	✓	✓	✓	—	—
Operating system FFDC	✓	✓	✓	✓	—	—
Error log analysis	✓	✓	✓	✓	✓	✓
Service processor support for BIST for logic & arrays, wire tests, and component initialization	✓	✓	✓	✓	✓	✓
Serviceability						
Boot time progress indicator	✓	✓	✓	✓	Limited	Limited
Firmware error codes	✓	✓	✓	✓	✓	✓
Operating system error codes	✓	✓	✓	✓	Limited	Limited
Inventory collection	✓	✓	✓	✓	✓	✓
Environmental and power warnings	✓	✓	✓	✓	✓	✓
Hot plug fans, power supplies	✓	✓	✓	✓	✓	✓

RAS feature	AIX V5.3	AIX V6.1	IBM i 5.4 LIC 5.4.5	IBM i 6.1	RHEL V5.1	SLES V10
Extended error data collection	✓	✓	✓	✓	✓	✓
SP call home on non-HMC configurations	✓	✓	✓	✓	✓	✓
I/O drawer redundant connections	✓	✓	✓	✓	✓	✓
I/O drawer hot-add and concurrent repair	✓	✓	✓	✓	✓	✓
SP mutual surveillance with POWER Hypervisor	✓	✓	✓	✓	✓	✓
Dynamic firmware update with the HMC	✓	✓	✓	✓	✓	✓
Service agent call home application	✓	✓	✓	✓	✓	✓
Guiding Light LEDs	✓	✓	✓	✓	✓	✓
System dump for memory, POWER Hypervisor, SP	✓	✓	✓	✓	✓	✓
Operating system error reporting to HMC SFP application	✓	✓	✓	✓	✓	✓
RMC secure error transmission subsystem	✓	✓	✓	✓	✓	✓
Health check scheduled operations with HMC	✓	✓	✓	✓	✓	✓
Operator panel (virtual or real)	✓	✓	✓	✓	✓	✓
Redundant HMCs	✓	✓	✓	✓	✓	✓
Automated recovery and restart	✓	✓	✓	✓	✓	✓
Repair and verify guided maintenance	✓	✓	✓	✓	Limited	Limited
Concurrent kernel update	—	✓	✓	✓	—	—

a. This feature is not supported on Version 4 of RHEL.

4.5 Manageability

Several functions and tools help you to efficiently and effectively manage your system.

4.5.1 Service processor

The service processor is a controller running its own operating system. It is a component of the service interface card.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface connected to the POWER6 processor. The service processor is always working, regardless of the main system unit's state. The system unit can be in the following states:

- Standby, power off
- Operating, ready to start partitions
- Operating, with running logical partitions

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring the connection to the HMC for manageability purposes and accepting ASMI Secure Sockets Layer (SSL) network connections. The service processor provides the ability to view and manage the machine wide settings using the ASMI, and allows complete system and partition management from the HMC.

> **Note:** The service processor enables a system that does not boot to be analyzed. The error log analysis can be performed from either the ASMI or the HMC.

The service processor uses two Ethernet 10/100 Mbps ports that have:

- Visibility to the service processor and can be used to attach the server to an HMC or to access the ASMI. The ASMI options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- A default IP address:
  - Service processor Eth0 or HMC1 port is configured as 169.254.2.147
  - Service processor Eth1 or HMC2 port is configured as 169.254.3.147

4.5.2 System diagnostics

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

- Online diagnostics, when installed, are a part of the AIX operating system on the disk or server. They can be booted in single user mode (service mode), run in maintenance mode, or run concurrently (concurrent mode) with other applications. They have access to the AIX error log and the AIX configuration data.
  - Service mode, which requires a service mode boot of the system, enables the checking of system devices and features. Service mode provides the most complete checkout of the system resources. All system resources, except the SCSI adapter and the disk drives used for paging, can be tested.
  - Concurrent mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, some devices might require additional actions by the user or diagnostic application before testing can be done.
  - Maintenance mode enables the checking of most system resources. Maintenance mode provides the same test coverage as service mode. The difference between the two modes is the way they are invoked. Maintenance mode requires that all activity on the operating system be stopped. The **`shutdown -m`** command is used to stop all activity on the operating system and put the operating system into maintenance mode.
- The system management services (SMS) error log is accessible on the SMS menus. This error log contains errors that are found by partition firmware when the system or partition is booting.
- The service processor's error log can be accessed on the ASMI menus.
- You can also access the system diagnostics from a Network Installation Management (NIM) server.

4.5.3 Electronic Service Agent

Electronic Service Agent and the IBM Electronic Services Web portal comprise the IBM Electronic Service solution. IBM Electronic Service Agent is a no-charge tool that proactively monitors and reports hardware events, such as system errors, performance issues, and inventory. Electronic Service Agent can help focus on your company's strategic business initiatives, save time, and spend less effort managing day to day IT maintenance issues.

Now integrated in AIX 5L V5.3 TL6 in addition to the HMC, Electronic Service Agent automatically and electronically reports system failures and issues to IBM, and that can result in faster problem resolution and increased availability. System configuration and inventory information collected by Electronic Service Agent also can be viewed on the secure Electronic Service Web portal, and used to improve problem determination and resolution between the client and the IBM support team. As part of an increased focus to provide better service to IBM clients, Electronic Service Agent tool configuration and activation is standard with the system. In support of this effort, an HMC external connectivity security whitepaper has been published, which describes data exchanges between the HMC and the IBM Service Delivery Center (SDC) and the methods and protocols for this exchange.

To access Electronic Service Agent user guides, perform the following steps:

1. Go to the IBM Electronic Services Web site at:

   `https://www-304.ibm.com/jct03004c/support/electronic/portal`

2. Select your country or region.
3. Click Electronic Service Agent

> **Note:** To receive maximum coverage, activate Electronic Service Agent on every platform, partition, and HMC in your network. If your IBM System p server is managed by an HMC, the HMC can report all hardware problems, and the AIX operating system can report only software problems and system information. You must configure the Electronic Service Agent on the HMC. The AIX operating system will not report hardware problems for a system managed by an HMC.

IBM Electronic Service provide the following benefits:

- Increased uptime

  Electronic Service Agent enhances the warranty and maintenance service by providing faster hardware error reporting and uploading system information to IBM support. This can optimize the time monitoring the symptoms, diagnosing the error, and manually calling IBM support to open a problem record. 24x7 monitoring and reporting means no more dependency on human intervention or off-hours client personnel when errors are encountered in the middle of the night.

- Security

  Electronic Service Agent is secure in monitoring, reporting, and storing the data at IBM. Electronic Service Agent securely transmits through the Internet (HTTPS or VPN) and can be configured to communicate securely through gateways to provide clients a single point of exit from their site. Communication between the client and IBM only flows one way. Activating Service Agent does not permit IBM to call into a client's system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. Your business applications or business data is never transmitted to IBM.

- More accurate reporting

  Because system information and error logs are automatically uploaded to the IBM support center in conjunction with the service request, your are not required to find and send system information, decreasing the risk of misreported or misdiagnosed errors. When inside IBM, problem error data is run through a data knowledge management system and knowledge articles are appended to the problem record.

- Customized support

  Using the IBM ID entered during activation, you may view system and support information in the My Systems and Premium Search sections of the Electronic Services Web site.

The Electronic Services Web portal is a single Internet entry point that replaces the multiple entry points traditionally used to access IBM Internet services and support. This Web portal enables you to more easily gain access to IBM resources for assistance in resolving technical problems.

The Service Agent provides the following additional services:

- My Systems

  Client and IBM employees authorized by the client can view hardware and software information and error messages that are gathered by Service Agent on Electronic Services.

- Premium Search

  A search service using information gathered by Service Agents (this is a paid service that requires a special contract).

For more information about using the power of IBM Electronic Services, visit the following Web site or contact an IBM Systems Services Representative.

`https://www-304.ibm.com/jct03004c/support/electronic/portal`

4.5.4 Manage serviceable events with the HMC

Service strategies become more complicated in a partitioned environment. The **Manage Serviceable Events** task in the HMC can help streamline this process.

Each logical partition reports errors it detects, without determining whether other logical partitions also detect and report the errors. For example, if one logical partition reports an error for a shared resource, such as a managed system power supply, other active logical partitions might report the same error.

By using the Manage Serviceable Events task in the HMC, you can:

- Avoid long lists of repetitive *call home* information by recognizing that these are repeated errors and consolidating them into one error.
- Initiate service functions on systems and logical partitions including the exchanging of parts, configuring connectivity, and managing dumps.

4.5.5 Hardware user interfaces

Two interfaces are discussed in this section. The ASMI and the graphics terminal.

Advanced System Management Interface (ASMI)

The ASMI interfaces to the service processor enables you to manage the operation of the server, such as auto power restart. You can also view information about the server, such as the error log and vital product data. Use the ASMI to change the service processor IP addresses or to apply some security policies and avoid the access from undesired IP addresses or range. Various repair procedures require connection to the ASMI.

If you are able to use the service processor’s default settings, accessing the ASMI is not necessary.

The ASMI is accessible through:

- The HMC. For details, see section the next section, or see section 2.14, “Advanced System Management Interface” on page 103.
- A Web browser on a system that is connected directly to the service processor (in this case, either a standard Ethernet cable or a crossed cable) or through an Ethernet network.
- An ASCII terminal.

Accessing the ASMI using an HMC

If configured to do so, the HMC connects directly to the ASMI for a selected system from this task.

To connect to the ASMI from an HMC:

1. Open Systems Management from the navigation pane.
2. From the work pane, select one or more managed systems to work with.
3. From the System Management tasks list, select Operations.
4. From the Operations task list, select Advanced System Management (ASM).

Accessing the ASMI through a Web browser

The Web interface to the ASMI is accessible through Microsoft Internet Explorer 6.0, Microsoft Internet Explorer 7, Netscape 7.1, Mozilla Firefox, or Opera 7.23 running on a PC or mobile computer connected to the service processor. The Web interface is available during all phases of system operation, including the initial program load (IPL) and runtime. However, some menu options in the Web interface are unavailable during IPL or runtime to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides a Secure Sockets Layer (SSL) Web connection to the service processor. To establish an SSL connection, open your browser using `https://` format.

> **Note:** To make the connection through Internet Explorer, click **Tools** → **Internet Options**. Uncheck **Use TLS 1.0**, and click **OK**.

Accessing the ASMI using an ASCII terminal:

The ASMI on an ASCII terminal supports a subset of the functions provided by the Web interface and is available only when the system is in the platform standby state. The ASMI on an ASCII console is not available during some phases of system operation, such as the IPL and run time

Graphics terminal

The graphics terminal is available to users who want a graphical user interface (GUI) to their AIX or Linux systems. To use the graphics terminal, plug the graphics adapter into a PCI slot in the back of the server. You can connect a standard monitor, keyboard, and mouse to the adapter to use the terminal. This connection allows you to access the SMS menus, as well as an operating system console.

4.5.6 IBM System p firmware maintenance

The IBM Power Systems client managed microcode is a methodology that enables you to manage and install microcode updates on Power Systems, IBM System p, IBM System p5, pSeries, and RS/6000® systems and associated I/O adapters. The IBM System p microcode can be installed either from an HMC or from a running partition in case that system is not managed by an HMC. For update details, see the Microcode Web page:

http://www14.software.ibm.com/webapp/set2/firmware/gjsn

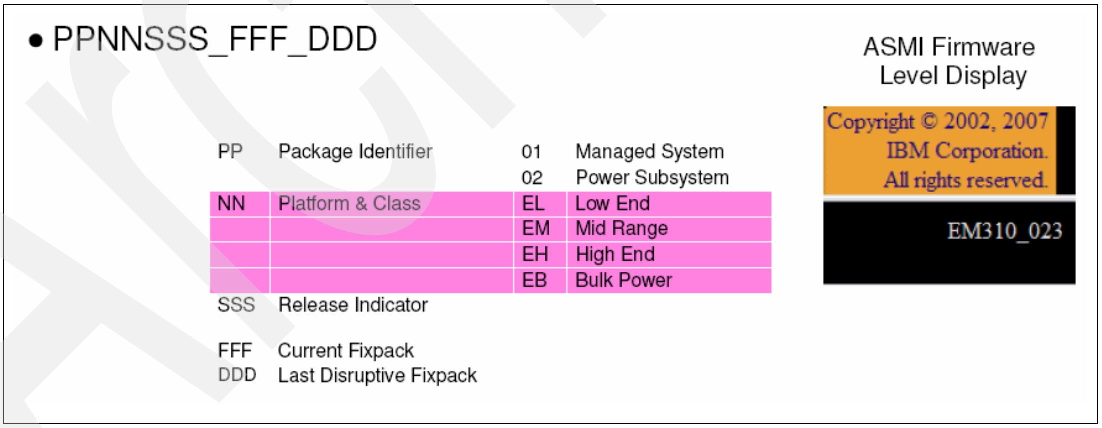
If you use an HMC to manage your server, use it also to view the levels of server firmware and power subsystem firmware that are installed on your server and are available to download and install.

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

- Installed level: This is the level of server firmware or power subsystem firmware that has been installed and will be installed into memory after the managed system is powered off and powered on. It is installed on the *t* side of system firmware.
- Activated level: This is the level of server firmware or power subsystem firmware that is active and running in memory.
- Accepted level: This is the backup level of server or power subsystem firmware. You can return to this level of server or power subsystem firmware if you decide to remove the installed level. It is installed on the *p* side of system firmware.

IBM provides the concurrent firmware maintenance (CFM) function on System p systems. This function supports applying nondisruptive system firmware service packs to the system concurrently (without requiring a reboot to activate changes). For systems that are not managed by an HMC, the installation of system firmware is always disruptive.

The concurrent levels of system firmware can, on occasion, contain fixes known as *deferred*. These deferred fixes can be installed concurrently but are not activated until the next IPL. For deferred fixes within a service pack, only the fixes in the service pack, which cannot be concurrently activated, are deferred. Figure 4-8 shows the system firmware file naming convention.



- PPNNSSS_FFF_DDD

PP	Package Identifier	01	Managed System
		02	Power Subsystem
NN	Platform & Class	EL	Low End
		EM	Mid Range
		EH	High End
		EB	Bulk Power
SSS	Release Indicator		
FFF	Current Fixpack		
DDD	Last Disruptive Fixpack		

ASMI Firmware Level Display

Copyright © 2002, 2007
IBM Corporation.
All rights reserved.

EM310_023

Figure 4-8 Firmware file naming convention

A file naming sample is shown in Example 4-1.

Example 4-1

```
01EM310_026_026 = Managed System Firmware for 9117-MMA Release 310 Fixpack 026
```

An installation is disruptive if:

- The release levels (*SSS*) of currently installed and new firmware are different.

- The service pack level (*FFF*) and the last disruptive service pack level (*DDD*) are equal in new firmware.

An installation is concurrent if the service pack level (*FFF*) of the new firmware is higher than the service pack level currently installed on the system and the previous conditions for disruptive installation are not met.

4.5.7 Management Edition for AIX

IBM Management Edition for AIX (ME for AIX) provides robust monitoring and quick time to value by incorporating out-of-the box best practice solutions that were created by AIX and PowerVM Virtual I/O Server developers. These best practice solutions include predefined thresholds for alerting on key metrics, expert advice that provides an explanation of the alert and recommends potential actions to take to resolve the issue, and the ability to take resolution actions directly from the Tivoli® Enterprise Portal or set up automated actions. Users have the ability to visualize the monitoring data in the Tivoli Enterprise Portal determine the current state of the AIX, LPAR, CEC, HMC and VIOS resources.

ME for AIX is an integrated systems management offering created specifically for the System p platform that provides as primary functions:

- Monitoring of the health and availability of the System p.
- Discovery of configurations and relationships between System p service and application components.
- Usage and accounting of System p IT resources.

For information regarding the ME for AIX, visit the following link:

http://www-03.ibm.com/systems/p/os/aix/sysmgmt/me/index.html

4.5.8 IBM Director

IBM Director is an integrated suite of tools that provides flexible system management capabilities to help realize maximum systems availability and lower IT costs.

IBM Director provides the following benefits:

- Easy to use and has a consistent look and feel, and single point of management to help simplify IT tasks.
- Automated, proactive capabilities to help reduce IT costs and maximize system availability
- Streamlined, intuitive interface to help you get started faster and accomplish more in a shorter period of time
- Open standards-based design, broad platform, and operating support to help you manage heterogeneous environments from a central point
- Extensible to provide more choice of tools from the same user interface

For more information about IBM Director, go to:

http://www-03.ibm.com/systems/management/director/

4.6 Cluster solution

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

Clusters offer the following advantages:

- High processing capacity
- Resource consolidation
- Optimal use of resources
- Geographic server consolidation
- 24x7 availability with failover protection
- Disaster recovery
- Scale-out and scale-up without downtime
- Centralized system management

The POWER process-based AIX and Linux cluster target scientific and technical computing, large-scale databases, and workload consolidation. IBM Cluster Systems Management (CSM) can help reduce the overall cost and complexity of IT management by simplifying the tasks of installing, configuring, operating, and maintaining clusters of servers or logical partitions (LPARs). CSM offers a single consistent interface for managing both AIX and Linux nodes, with capabilities for remote parallel network installation, remote hardware control, distributed command execution, file collection and distribution, cluster-wide monitoring capabilities, and integration with High Performance Computing (HPC) applications.

CSM V1.7 which is need to support POWER6 process-based HMC include highly available management server (HA MS) at no additional charge. CSM HA MS is positioned for enterprises that need the HA MS. The CSM HA MS is designed to remove the management server as a single point of failure in the cluster.

For information regarding the IBM CSM for AIX, HMC control, cluster building block servers, and cluster software available, see the following items:

- IBM System Cluster 1600 at:

  http://www-03.ibm.com/systems/clusters/hardware/1600/index.html

- IBM System Cluster 1350™ at:

  http://www-03.ibm.com/systems/clusters/hardware/1350/index.html

Related publications

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

IBM Redbooks

For information about ordering these publications, see "How to get Redbooks" on page 173. Note that some of the documents referenced here might be available in softcopy only.

- *PowerVM Virtualization on IBM System p: Introduction and Configuration Fourth Edition*, SG24-7940
- *PowerVM Virtualization on IBM System p: Managing and Monitoring*, SG24-7590
- *Getting Started with PowerVM Lx86*, REDP-4298
- *PowerVM Live Partition Mobility on IBM System p*, SG24-7460
- *Integrated Virtualization Manager on IBM System p5*, REDP-4061
- *Introduction to Workload Partition Management in IBM AIX Version 6.1*, SG24-7431
- *Hardware Management Console V7 Handbook*, SG24-7491
- *LPAR Simplification Tools Handbook*, SG24-7231

Other publications

These publications are also relevant as further information sources:

- *Logical Partitioning Guide*, SA76-0098
- *Site and Hardware Planning Guide*, SA76-0091
- *Site Preparation and Physical Planning Guide*, SA76-0103

These publications are also relevant as further information sources for installing:

- *Installation and Configuration Guide for the HMC*, SA76-0084
- *PCI Adapter Placement*, SA76-0090

These publications are also relevant as further information sources for using your system:

- *Introduction to Virtualization*, SA76-0145
- *Operations Guide for the ASMI and for Nonpartitioned Systems*, SA76-0094
- *Operations Guide for the HMC and Managed Systems*, SA76-0085
- *Virtual I/O Server and Integrated Virtualization Manager Command Reference*, SA76-0101

This publication is also relevant as further information sources for troubleshooting:

- *AIX Diagnostics and Service Aids*, SA76-0106

Online resources

These Web sites are relevant as further information sources:

- IBM EnergyScale for POWER6 Processor-Based Systems

  http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=SA&subtype=WH&attachment=POW03002USEN.PDF&appname=STGE_PO_PO_USEN&htmlfid=POW03002USEN
- IBM Systems Hardware Information Center

  http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp
- Disk systems

  http://www-03.ibm.com/systems/storage/disk/
- Fix Central

  http://www-933.ibm.com/eserver/support/fixes/fixcentral
- IBM Linux

  http://www-03.ibm.com/systems/power/software/linux/index.html
- IBM System Planning Tool

  http://www-304.ibm.com/systems/support/tools/systemplanningtool/
- IBM Prerequisite

  https://www-912.ibm.com/e_dir/eServerPrereq.nsf
- Support for Systems and Servers

  http://www-304.ibm.com/systems/support/
- Upgrade Planning

  http://www-304.ibm.com/systems/support/i/planning/upgrade/index.html
- Support for Network attached storage (NAS) & iSCSI

  http://www-304.ibm.com/systems/support/supportsite.wss/allproducts?brandind=5000029&taskind=1
- Support for Hardware Management Console for Power Systems

  https://www14.software.ibm.com/webapp/set2/sas/f/hmc/home.html
- Microcode downloads

  http://www14.software.ibm.com/webapp/set2/firmware/gjsn
- PowerVM Editions

  http://www-03.ibm.com/systems/power/software/virtualization/editions/index.html
- Capacity on Demand, Power Systems Capacity on Demand solutions

  http://www-03.ibm.com/systems/power/hardware/cod/
- IBM Capacity Backup for Power Systems

  http://www-03.ibm.com/systems/power/hardware/cbu/

How to get Redbooks

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

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM Power 595

Technical Overview and Introduction

Redpaper™

PowerVM virtualization technology including Live Partition Mobility

World-class performance and flexibility

Mainframe-inspired continuous availability

This IBM Redpaper is a comprehensive guide describing the IBM Power 595 (9119-FHA) enterprise-class server. The goal of this paper is to introduce several technical aspects of this innovative server. The major hardware offerings and prominent functions include:

- The POWER6 processor available at frequencies of 4.2 and 5.0 GHz
- Specialized POWER6 DDR2 memory that provides improved bandwidth, capacity, and reliability
- Support for AIX, IBM i, and Linux for Power operating systems
- EnergyScale technology that provides features such as power trending, power-saving, thermal measurement, and processor napping
- PowerVM virtualization
- Mainframe levels of continuous availability

This Redpaper is intended for professionals who want to acquire a better understanding of Power Systems products, including:

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

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

REDP-4440-00