IBM

IBM Power 795 (9119-FHB)

Technical Overview and Introduction

- Explores class-leading performance, energy efficiency, and scalability
- Shows how the design can offer maximum availability
- Features information about new I/O cards

An Ding Chen
Carlo Costantini
James Cruickshank
Dave Freeman
Volker Haug
Breno Henrique Leitão
Cesar Diniz Maciel
John T Schmidt

ibm.com/redbooks

Redpaper

International Technical Support Organization

IBM Power 795 (9119-FHB) Technical Overview and Introduction

February 2013

REDP-4640-01

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

Second Edition (February 2013)

This edition applies to the IBM Power 795 (9119-FHB) Power Systems server.

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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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:

Active Memory™
AIX 5L™
AIX®
AS/400®
DB2®
DS8000®
Electronic Service Agent™
EnergyScale™
Focal Point™
IBM Systems Director Active Energy Manager™
IBM®
Micro-Partitioning®
Power Architecture®
POWER Hypervisor™
Power Systems™
Power Systems Software™
POWER6+™
POWER6®
POWER7®
PowerHA®
PowerPC®
PowerVM®
POWER®
pSeries®
Rational Team Concert™
Rational®
Redbooks®
Redpaper™
Redbooks (logo) ®
RS/6000®
Storwize®
System i®
System p®
System Storage®
System z®
Tivoli®
WebSphere®
Workload Partitions Manager™
XIV®

The following terms are trademarks of other companies:

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

LTO, Ultrium, the LTO Logo and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

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

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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

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

Preface

This IBM® Redpaper™ publication is a comprehensive guide that covers the IBM Power 795 server that supports IBM AIX®, IBM i, and Linux operating systems. The goal of this paper is to introduce the innovative Power 795 offering and its major functions:

- IBM POWER7® processor, available at frequencies of 3.7 GHz and 4.0 GHz with TurboCore options of 4.25 GHz and 4.31 GHz
- Specialized POWER7 Level 3 cache that provides greater bandwidth, capacity, and reliability
- The latest IBM PowerVM® virtualization, including PowerVM Live Partition Mobility and PowerVM IBM Active Memory™ Sharing
- TurboCore mode that delivers the highest performance per core and can save on software costs for applications that are licensed per core
- Enhanced reliability, accessibility, and serviceability (RAS) features that are designed for maximum availability
- Active Memory Expansion technology that provides more usable memory than is physically installed in the system
- IBM EnergyScale™ technology that provides features such as power trending, power-saving, capping of power, and thermal measurement

Professionals who want to acquire a better understanding of IBM Power Systems™ products can benefit from reading this publication. The intended audience includes the following roles:

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

This paper complements the available set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power 795 system.

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

The team who wrote this paper

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization (ITSO), Poughkeepsie Center. The project that produced this publication was managed by Scott Vetter.

An Ding Chen is a Power systems Product Engineer in Shanghai China, who provides Level 3 hardware and firmware support in all Asia Pacific countries and Japan. He has ten years of experience with AIX/UNIX, IBM RS/6000®, IBM pSeries®, and Power Systems products. At his university, he passed the CATE certification on pSeries systems and IBM AIX 5L. He joined IBM in 2006.

Carlo Costantini is a Certified IT Specialist for IBM and has over 32 years of experience with IBM and IBM Business Partners. He currently works in Italy on 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.

James Cruickshank works on the Power Systems Field Technical Sales Support team for IBM in the UK. He holds an Honors degree in Mathematics from the University of Leeds. James has nine years of experience working on RS/6000, IBM System p®, and Power Systems products and is a member of the EMEA Power Champions team. James supports customers in the financial services sector in the UK.

Dave Freeman has worked for IBM since 1985. He is currently a Systems Service Representative (SSR) with IBM UK. He has worked extensively with IBM System p, IBM Power Systems, and IBM Storage for the last 10 years. Prior to this role, Dave was an IT Systems Engineer, providing presales technical support to IBM salesmen and IBM Business Partners in the small and medium business sector, primarily on IBM i (AS/400®). He has a degree in Information Technology from the Polytechnic of Central London.

Volker Haug is a certified Consulting IT Specialist in the IBM Systems and Technology Group, based in Ehningen, Germany. He holds a Bachelor's degree in Business Management from the University of Applied Studies in Stuttgart. His career has included more than 23 years working in the IBM PLM and Power Systems divisions as a RISC and AIX Systems Engineer. Volker is an expert in Power Systems hardware, AIX, and PowerVM virtualization. He is a member of the EMEA Power Champions team and also a member of the German Technical Expert Council, a affiliate of the IBM Academy of Technology. He has written various books and white papers about AIX, workstations, servers, and PowerVM virtualization.

Breno Henrique Leitão is an Advisory Software Engineer at the Linux Technology Center in Brazil. He has 14 years of experience with Linux. Breno is also a Master Inventor in Brazil and holds a degree in Computer Science from Universidade de Sao Paulo. His areas of expertise include operating systems performance, virtualization, and networking. He has written extensively on Linux, mainly about networking and debugging.

Cesar Diniz Maciel is a Senior Consulting IT Specialist with IBM in United States. He joined IBM in 1996 in Presales Technical Support for the RS/6000 family of UNIX servers in Brazil, and moved to IBM in the United States in 2005. He is part of the Global Techline team, working on presales consulting for Latin America. He holds a degree in Electrical Engineering from Universidade Federal de Minas Gerais (UFMG) in Brazil. His areas of expertise include Power Systems, AIX, and IBM POWER® Virtualization. He has written extensively about Power Systems and related products. This ITSO residency is his seventh.

John T Schmidt is an Accredited IT Specialist for IBM and has ten years of experience with IBM and Power Systems. He has a degree in Electrical Engineering from the University of Missouri - Rolla and an MBA from Washington University in St. Louis. In 2010, he completed an assignment with the IBM Corporate Service Corps in Hyderabad, India. He is currently working in the United States as a presales Field Technical Sales Specialist for Power Systems in St. Louis, MO.

The project that produced this deliverable was managed by:

Scott Vetter
Executive project manager, PMP

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

Salim Agha, Ray W. Anderson, Ron Arroyo, Richard W. Bishop, Joe Cahill, Sertac Cakici, Scott A. Carroll, Hsien-I Chang, Binh Chu, Charlie Cler, Michael Floyd, George Gaylord, Raymond J. Harrington, Daniel Henderson, John Hilburn, Bob Kovacs, Walter M. Lipp, Mary E. Magnuson, Bill McWaters, Duc T. Nguyen, Thoi Nguyen, Guy R. Paradise, David Pirnik, Pat O'Rourke, Amartey S. Pearson, Todd J. Rosedahl, Jeff Scheel, Joseph C. Schiavone, Timothy Stack, Vasu Vallabhaneni, Jacobo Vargas, Jesus G. Villarreal
IBM U.S.A.

Sabine Jordan, Stephen Lutz
IBM Germany

Tamikia Barrow, Brooke C Gardner
International Technical Support Organization, Poughkeepsie Center

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author - all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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® publications in one of the following ways:

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

  ibm.com/redbooks

- Send your comments in an email 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

Stay connected to IBM Redbooks publications

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks publications weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Summary of changes

This section describes the technical changes made in this edition of the paper and in previous editions. This edition might also include minor corrections and editorial changes that are not identified.

Summary of Changes
for IBM Power 795 (9119-FHB) Technical Overview and Introduction
as created or updated on February 8, 2013.

February 2013, Second Edition

This revision reflects the addition, deletion, or modification of new and changed information described below.

New information

- Added feature codes
- Added external storage options

Changed information

- A general freshening that is required to bring this paper up to date.

1

General description

The IBM Power 795 server is designed to help enterprises deploy cost-effective and flexible IT infrastructure while achieving exceptional application performance and increasing the speed of deployment of new applications and services. As the most powerful member of the IBM Power Systems family, this server provides exceptional performance, massive scalability, and energy-efficient processing for a full range of complex, mission-critical applications with the most demanding computing requirements.

The innovative IBM Power 795 server with POWER7 processors is a symmetric multiprocessing (SMP), rack-mounted server. Equipped with either eight 32-core or 24-core processor books, the Power 795 server can be deployed from 24-core to 256-core SMP configurations.

With up to 16 TB of buffered DDR3 memory and extensive I/O support, the Power 795 server can scale rapidly and seamlessly to address the changing needs of today's data center. With advanced PowerVM virtualization, EnergyScale technology, capacity on demand (CoD) options, and Active Memory Expansion options, the Power 795 helps businesses take control of their IT infrastructure and confidently consolidate multiple applications for the AIX, IBM i, and Linux operating systems onto a single system.

1.1 Overview of the Power 795

Figure 1-1 shows the Power 795 system.

Figure 1-1 The Power 795

The Power 795 server (9119-FHB) offers an expandable, high-end enterprise solution for managing the computing requirements to enable your business to become an On Demand Business. The Power 795 is a 24- to 256-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 16 TB of memory is supported on the Power 795 server.

The Power 795 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. The following CEC features are included:
  - 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 POWER7 processor books. Each processor book can contain either 24 or 32 cores that are packaged on four POWER7 processor chips. Each processor chip contains 2 MB of on-chip L2 cache and 32 MB of eDRAM L3 cache, and each core supports four hardware threads. Each processor book also provides these features:
    - Slots: 32 DDR3 memory DIMM slots
    - Cards: Support for up to four GX++ based I/O hub adapter cards (12X) for connection to system I/O drawers
    - Two node controllers (NC): One primary NC and a second one for redundancy
- One or two optional powered expansion racks, each with 32U of rack space for up to eight, 4U, I/O 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 thirty-two I/O drawers, which can be a combination of one of the following drawers (FC indicates feature code):
  - 12X I/O Drawer PCIe, SFF SAS disk (FC 5803; minimum of one)
  - 12X I/O Drawer PCIe, no disk (FC 5873; maximum of 31 drawers per system)
  - 12X I/O Drawer PCI-X, with repeater, SCSI disk (FC 5797; maximum of 30 drawers per system)

FC = feature code: Throughout this paper, all feature codes are referenced as FC *xxxx*, where *xxxx* is the feature code number.

- In addition to the 24 inch rack-mountable I/O drawers are standard, 2 meters high, 19-inch I/O racks for mounting 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 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 2,220 hot-swap SAS disks (999 TB).

Reusable components: When upgrading a Power 595 to a Power 795, some existing components are migrated and can be used in place of those we listed. See 1.6.1, “Upgrade considerations” on page 25 for details about reusable components.

What makes this IBM high-end server truly unique is the ability to switch between its standard throughput mode and its unique, optimized TurboCore mode, where performance per core is boosted with access to both additional L3 cache and additional clock speed. In standard mode (also known as MaxCore mode), each 8-core SCM operates at 4.0 GHz with 32 MB of L3 cache and can scale up to 256 cores. When operating in TurboCore mode, each 8-core single-chip module (SCM) operates with up to four active cores per SCM at 4.25 GHz and 32 MB of L3 cache, twice the L3 cache per core that is available in MaxCore mode. In TurboCore mode, the Power 795 server can scale up to 128 cores. Each 6-core SCM operates at 3.72 GHz only in standard mode with 4 MB of L3 cache per core. Both the 8-core and 6-core SCMs are supported by 256 KB of L2 cache per core.

Table 1-1 lists the major attributes of the Power 795 (9119-FHB) server.

Table 1-1 Attributes of the 9119-FHB

Attribute	**9119-FHB**
SMP processor configurations	24 to 256 POWER7 processor cores
Processor books	1 - 8
POWER7 processor clock rate	3.72 GHz (24-core books), 4.00 GHz (MaxCore mode) or 4.25 GHz (Turbocore mode)
L2 cache	256 KB per core
L3 cache	24 MB per 6-core POWER7 processor (shared by six cores) 32 MB per 8-core POWER7 processor (shared by eight cores)
RAM (memory)	Minimum of eight memory DIMMS per processor book (64 GB memory) supported. Minimum of 256 GB per processor book is best.
Processor packaging	Single-chip module (SCM)
Maximum memory configuration	16 TB
Rack space	42U 24-inch custom rack
I/O drawers	24-inch: 1 - 32 drawers
Internal disk bays	16 maximum per 24-inch PCI-X I/O drawer 26 maximum per 24-inch PCIe I/O drawer
Internal disk storage	Up to 11.7 TB per 24-inch PCIe I/O drawer
I/O adapter slots	Up to 640 PCIe slots when using FC 5803 or FC 5873 drawers, or up to 600 PCI-X slots when using FC 5797
I/O ports	4 GX++ adapter ports per processor book, 32 per system
PowerVM Standard Edition (optional)	IBM Micro-Partitioning® with up to 10 micropartitions 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 and Active Memory Sharing

Attribute	9119-FHB
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	IBM PowerHA® family
RAS features	Active Memory Mirroring for Hypervisor Processor instruction retry Alternate processor recovery Selective dynamic firmware updates Enhanced ECC DRAM, spare memory chip ECC L2 cache, L3 cache Redundant service processors with dynamic failover Redundant system clocks with dynamic failover Hot-swappable disk bays Hot-plug/blind-swap PCI-X and PCIe 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 and PCIe slots Redundant power supplies and cooling fans Redundant battery backup (optional)
Operating systems	AIX V5.3 or later IBM i 6.1 or later SUSE Linux Enterprise Server 10 for POWER SP3 or later Red Hat Enterprise Linux 5 Update 5 for POWER or later

Rack terminology: 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.

1.2 Installation planning

Complete installation instructions are included with each server. Comprehensive planning information is available at the following 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 795 server dimensions.

Table 1-2 Power 795 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)	127.3 cm (50.1 in)	148.6 cm (58.5 in)[a] 152.1 cm (61.3 in)[b]	148.6 cm (58.5 in)	180.6 cm (71.1 in)[c] 179.6 cm (70.7 in)[d]	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
c. Rack with acoustic front door and rear door, one or two frames
d. Rack with acoustic front door and rear door heat exchanger (RDHX), system rack only

Table 1-3 lists the Power 795 server full system weights without the covers.

Table 1-3 Power 795 server full system weights (no covers)

Frame	With integrated battery backup	Without integrated battery backup
System rack	1193 kg (2630 lb)	1102 kg (2430 lb)
Powered expansion rack	1291 kg (2845 lb)	1200 kg (2645 lb)
Power Expansion rack with bolt-on expansion rack	2381 kg (5250 lb)	2290 kg (5050 lb)

Table 1-4 lists the Power 795 cover weights.

Table 1-4 Power 795 cover weights

Covers	Weight
Side covers, pair	50 kg (110 lb)
Slimline covers, single	15 kg (33 lb)
Acoustic doors, single	25 kg (56 lb)

1.2.2 Operating environment

Table 1-5 lists the operating environment specifications for the Power 795 server.

Table 1-5 Power 795 server operating environment specifications

Description	Range
Preferable operating temperature	10 - 27 degrees C[a] (50 - 80.6 degrees F)
Relative humidity	20 - 80%
Sound power	► Declared A-weighted sound power level, per ISO 9296: 8.7 bels maximum (with slim line doors) ► Declared A-weighted sound power level, per ISO 9296: 7.8 bels maximum (with acoustical doors)
Sound pressure	► Declared A-weighted one-meter sound pressure level, per ISO 9296: 69 decibels maximum (with slim line doors) ► Declared A-weighted one-meter sound pressure level, per ISO 9296: 60 decibels maximum (with acoustical doors)

a. The maximum operating temperature decreases 2 degrees C per 305 m (1000 ft) when you install in an altitude over 2134 m (7000 ft).

1.2.3 Power requirements

All Power 795 configurations are designed with a fully redundant power system. To take full advantage of the power subsystem redundancy and reliability features, connect each of the two power cords to different distribution panels.

Table 1-6 lists the electrical and thermal characteristics for a new Power 795 server.

Table 1-6 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 to 510 VAC ► DC current: 380 to 520 VDC ► Rated current (A per phase): 48 A or 63 A or 80 A; 34 A or 43 A; 24 A or 34 A
Operating frequency	50 or 60, plus or minus 0.5 Hz
Maximum Power source loading (8 processor books, 3 I/O drawers)	► 200 to 240 V ac, 31.9 kw ► 380 to 440 V ac, 30.6 kw ► 480 to 510 V ac, 30.8 kw ► 380 to 520 V dc, 30.8 kw
Maximum thermal output (8 processor books, 3 I/O drawers)	► 200 to 240 V ac, 108,847 BTU/h ► 380 to 415 V ac, 104,412 BTU/h ► 480 to 510 V ac, 105,094 BTU/h ► 380 to 520 V dc, 105,094 BTU/h

Description	Range
Typical Power source loading (8 processor books, 3 I/O drawers)	► 200 to 240 V ac, 20.5 kw ► 380 to 440 V ac, 19.7kw ► 480 to 510 V ac, 19.8kw ► 380 to 520 V dc, 19.8 kw
Typical thermal output (8 processor books, 3 I/O drawers)	► 200 to 240 V ac, 69,949 BTU/h ► 380 to 415 V ac, 67,219 BTU/h ► 480 to 510 V ac, 67,560 BTU/h ► 380 to 520 V dc, 67,560 BTU/h

Power evaluation: Actual system power is strongly affected by memory configuration and system workload, and is typically less than the listed maximum.The following assumptions are made on typical evaluations:

- System in normal operating mode with no processors in TurboCore mode
- A “typical” commercial workload, typical processor sort, altitude less than 5000 feet and temperature less than 27 degrees C.
- 32 x 16 GB DIMMs per book

The IBM Systems Energy Estimator is a web-based tool for estimating power requirements for IBM Power Systems. You can use this tool to estimate typical power requirements (watts) for a specific system configuration under normal operating conditions.

For further information about IBM Systems Energy Estimator see the following location:

http://www-912.ibm.com/see/EnergyEstimator

1.3 System disks and media features

This topic focuses on the I/O device support within the system unit. The Power 795 servers have internal hot-swappable drives supported in I/O drawers. I/O drawers can be allocated in 24-inch or 19-inch rack. Specific client requirements can be satisfied with several external disk options that are supported by the Power 795 server.

For further information about IBM disk storage systems, including withdrawn products, go to the following location:

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

The 12X I/O drawers FC 5803, FC 5873, FC 5797, and FC 5798 may be used for additional I/O capacity, on both of the 795 expansion racks (FC 6953, FC 6954), and the CEC rack. Either the FC 5803, FC 5798, or FC 5797 may be used as the primary I/O drawer located at the EIA-5 position in the CEC rack. The primary I/O drawer provides power to the media drawer (FC 5720 or FC 5724), when present.

Additional DASD expansion is available in a 42U high, 19-inch unpowered expansion rack (FC 0553). Both the feature 5786 SCSI (4U) and feature 5886 or 5887 SAS drawers may be

mounted in this rack. These drawers are connected to adapters located in one of the 24-inch PCI I/O drawers FC 5803, FC 5873, FC 5798, and FC 5797.

Table 1-7 lists the Power 795 hard disk drive features that are available for I/O drawers.

Table 1-7 IBM Power 795 disk drive feature codes and descriptions

Feature code	Description	Support		
		AIX	IBM i	Linux
1888	139 15K RPM SAS Disk Drive	—	✓	—
1947	139 GB 15K RPM SAS SFF-2 Disk Drive	—	✓	—
1886	146 GB 15K RPM SAS Disk Drive	✓	—	✓
1917	146 GB 15K RPM SAS SFF-2 Disk Drive	✓	—	✓
1775	177 GB SFF-1 SSD with eMLC	✓	—	✓
1787	177 GB SFF-1 SSD with eMLC	—	✓	—
1793	177 GB SFF-2 SSD with eMLC	✓	—	✓
1794	177 GB SFF-2 SSD with eMLC	—	✓	—
1911	283 GB 10K RPM SAS SFF Disk Drive	—	✓	—
1956	283 GB 10K RPM SAS SFF-2 Disk Drive	—	✓	—
1948	283 GB 15K RPM SAS SFF-2 Disk Drive	—	✓	—
1885	300 GB 10K RPM SAS Disk Drive	✓	—	✓
1925	300 GB 10K RPM SAS SFF-2 Disk Drive	✓	—	✓
1953	300 GB 15K RPM SAS SFF-2 Disk Drive	✓	—	✓
ES0A	387 GB SFF-1 SSD	✓	—	✓
ES0B	387 GB SFF-1 SSD	—	✓	—
ES0C	387 GB SFF-2 SSD	✓	—	✓
ES0D	387 GB SFF-2 SSD	—	✓	—
1916	571 GB 10K RPM SAS SFF Disk Drive	—	✓	—
1962	571 GB 10K RPM SAS SFF-2 Disk Drive	—	✓	—
1790	600 GB 10K RPM SAS SFF Disk Drive	✓	—	✓
1964	600 GB 10K RPM SAS SFF-2 Disk Drive	✓	—	✓
1737	856 GB 10K RPM SAS SFF-1 Disk Drive	—	✓	—
1738	856 GB 10K RPM SAS SFF-2 Disk Drive in Gen2-S Carrier	—	✓	—
1751	900 GB 10K RPM SAS SFF-1 Disk Drive	✓	—	✓
1752	900 GB 10K RPM SAS SFF-2 Disk Drive in Gen2-S Carrier	✓	—	✓
EQ37	Quantity 150 of FC 1737	—	—	—

Feature code	Description	Support		
		AIX	IBM i	Linux
EQ38	Quantity 150 of FC 1738	—	—	—
EQ51	Quantity 150 of FC 1751	—	—	—
EQ52	Quantity 150 of FC 1752	—	—	—
7582	Quantity 150 of FC 1777	—	—	—
7578	Quantity 150 of FC 1785	—	—	—
7550	Quantity 150 of FC 1790	—	—	—
1887	Quantity 150 of FC 1793	—	—	—
1958	Quantity 150 of FC 1794	—	—	—
7547	Quantity 150 of FC 1885	—	—	—
7548	Quantity 150 of FC 1886	—	—	—
7544	Quantity 150 of FC 1888	—	—	—
7557	Quantity 150 of FC 1911	—	—	—
7566	Quantity 150 of FC 1916	—	—	—
1866	Quantity 150 of FC 1917	—	—	—
1869	Quantity 150 of FC 1925	—	—	—
1868	Quantity 150 of FC 1947	—	—	—
1844	Quantity 150 of FC 1956	—	—	—
1817	Quantity 150 of FC 1962	—	—	—
1818	Quantity 150 of FC 1964	—	—	—
7564	Quantity 150 of FC 3648	—	—	—
7565	Quantity 150 of FC 3649	—	—	—
EQ0A	Quantity 150 of FC ES0A	—	—	—
EQ0B	Quantity 150 of FC ES0B	—	—	—
EQ0C	Quantity 150 of FC ES0C	—	—	—
EQ0D	Quantity 150 of FC ES0D	—	—	—

The Power 795 server must have access to a device that is capable of reading CD/DVD media, or have access to a NIM server. The preferred devices for reading CD/DVD media are the Power 795 media drawer (FC 5720 or FC 5724), the external DVD device (7214-1U2 or 7216-103), or both the media drawer and DVD device. Ensure that a SAS adapter is 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 preferred NIM server connection is a PCI based Ethernet LAN adapter plugged in to 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 hard disks is required per 795 server. It is best that these disks be used as mirrored boot devices. Mount these disks in the first I/O drawer when 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, SAS, and Fibre Channel adapters, or from networks by Ethernet.

Placement of the operating system disks in the first I/O drawer allows the operating system to boot even if other I/O drawers are found offline during boot.

See 1.8.4, “Useful rack additions” on page 31 for more details about available media features.

1.4 Minimum configuration requirements

This section discusses the minimum configuration requirements for the Power 795. Also provided are the appropriate feature codes for each system component. The IBM configuration tool identifies the feature code for each component in your system configuration. Table 1-8 lists the required components for a minimum 9119-FHB configuration.

Table 1-8 Power 795 minimum system configuration

Quantity	Component description	Feature code
1	Power 795	9119-FHB
1	POWER7 processor node	FC 4702 24-core node FC 4700 32-core node
24	1-core, processor activations	24x FC 4714 when using 24-core node 24 x FC 4713 when using 32-core node
2	2 x identical memory features, 0/32 GB or larger	FC 5600 0/32 GB FC 5601 0/64 GB FC 5602 0/128 GB FC 5564 0/256 GB
48[a]	1 GB memory activations: minimum 50% of installed memory required to be active	FC 8212
1	Power Cable Group, first processor book	FC 6991
4	Bulk power regulators	FC 6332
2	Power distribution assemblies	FC 6352
2	Line cords, selected depending on country and voltage	—
1	Pair of doors (front/back), either slim line, acoustic, or heat exchanger	—
1	Universal lift tool/shelf/stool and adapter	FC 3759 and FC 3761
1	Language - specify one	FC 93xx (country dependent)
1	HMC (7042-C0x/CRx) attached with Ethernet cables	—
1	12X I/O Loop Adapter	FC 1816

Quantity	Component description	Feature code
Either 1	12X PCIe/DASD I/O drawer with: ► Two DDR 12X I/O cable 2.5 m ► One DDR 12X I/O cable 0.6 m ► PCIe SAS adapter to connect to I/O drawer for boot ► Cable from PCIe SAS adapter to drawer ► Two 73.4 GB SAS disks (AIX/Linux)	FC 5803 FC 1863 FC 1861 FC 3688 FC 1883
Or 1	12X PCI-X/DASD I/O drawer when upgrading from 9119-FHA with: ► Two Enhanced 12X I/O cable 2.5 m ► One Enhanced 12X I/O cable 0.6 m ► SCSI disk drives migrated from the 9119-FHA	FC 5797 or FC 5798 FC 1831 FC 1829
1	PCIe or PCI-X SAS adapter for attachment of a DVD drive	—
2	139.5 GB SAS disk drive or SCSI disk drive (IBM i)	—
1	Media drawer to hold the optical and tape drives	FC 5724 or FC 7216-1U2[b]

a. A minimum of 50% of installed memory capacity must be activated by using either memory feature: FC 5600 or FC 5602. Because there is a minimum of 24 cores activated in one processor book (FC 4700 or FC 4702), we suggest that a minimum of 48 GB of memory be activated when one processor book is installed.
b. Requires FC 0553 (19-inch rack).

If either the AIX or Linux operating system is specified as the primary operating system, see Table 1-9 for a list of the minimum required features.

Table 1-9 Minimum required feature 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	FC 2146 (for AIX) FC 2147 (for Linux)

If IBM i is specified as the primary operating system, see Table 1-10 for a list of the minimum required features.

Table 1-10 Minimum required features when IBM i is the primary operating system

Quantity	Component description	Feature code
1	Primary operating system indicator for IBM i	FC 2145
1	System console specify	—
1	Load source specify[a]	For example, FC 0726

a. If load source is a FC 5886, one 42 EIA unit 19-inch rack is required. If load source is SAN boot, then a PCIe Fibre Channel adapter is required.

1.4.1 Processor node features

Each of the processor nodes (also known as *processor books*) in the Power 795 contains four POWER7 processor chips. Each processor chip has six or eight cores with 256 KB L2 cache per core and 4 MB L3 cache per core. Figure 2-1 on page 38 shows the layout of a Power 795 node.

The two types of processor nodes available on the Power 795 offer the following features:

- Four 6-core POWER7 single chip glass ceramic modules with 24 MB of L3 cache (24 cores per processor node) at 3.7 GHz (FC 4702)
- Four 8-core POWER7 single chip glass ceramic modules with 32 MB of L3 cache (32 cores per processor node) at 4.0 GHz (FC 4700)

The 8-core POWER7 processor module can run in two processing modes

- MaxCore

  In standard or MaxCore mode, the Power 795 system uses all processor cores running at 4.0 GHz and has access to the entire 32 MB of L3 cache.

- TurboCore

  In TurboCore mode, only four of the eight processor cores are available but at a higher frequency (4.25 GHz) and these four cores have access to the entire 32 MB of L3 cache. Thus, in TurboCore mode there are fewer cores running at a higher frequency with a higher core-to-L3-cache ratio.

For a more detailed description of MaxCore and TurboCore modes, see 2.3.5, “Flexible POWER7 processor packaging and offerings” on page 63.

1.4.2 Summary of processor features

Table 1-11 summarizes the processor feature codes for the Power 795.

Table 1-11 Summary of processor features for the Power 795

Feature code	Description	Support		
		AIX	IBM i	Linux
4700	4.0 GHz/4.25 GHz TurboCore processor node, 0/32-core POWER7, 32 DDR3 memory slots. 32 x 4.0 GHz POWER7 processors (inactive) on one node or 16 x 4.25 GHz POWER7 processors (inactive) on one node when using TurboCore mode.	✓	✓	✓
4702	3.7 GHz processor node, 0/24-core POWER7, 32 DDR3 memory slots. 24 x 3.7 GHz processors (inactive) on one node. TurboCore is not supported.	✓	✓	✓
4704	100 Processor-day On/Off usage billing for FC 4700 or FC 7560: after an On/Off Processor Enablement feature is ordered and the associated enablement code is entered into the system, you must report your On/Off usage to IBM at least monthly. This information, used to compute your billing data, is provided to your sales channel. The sales channel will place an order on your behalf for the quantity of this feature that matches your reported use. One FC 4704 provides 100 day of On/Off processor billing for POWER7 CoD Processor Book FC 4700 or FC 7560.	✓	—	✓
4706	Billing for 100 processor minutes usage of processor book FC 4700: Utility billing for 100 processor minutes for 0/32-core POWER7 processor, FC 4700.	✓	—	✓
4707	Billing for 100 processor minutes usage of processor book FC 4702: Utility billing for 100 processor minutes for 0/32-core POWER7 processor, FC 4702.	✓	—	✓

Feature code	Description	Support		
		AIX	IBM i	Linux
4709	On/Off proc CoD billing, one processor day, for processor book FC 4700 or FC 7560: After the On/Off Processor function is enabled in a system you must report your On/Off usage to IBM at least monthly. This information, used to compute your billing data, is provided to your sales channel. The sales channel will place an order on your behalf for the quantity of this feature that matches your reported use. One FC 4709 provides one day of on/off processor billing for 0/32-core POWER7 CoD Processor Book FC 4700 or FC 7560	✓	—	✓
4711	On/Off processing CoD billing, 1 processing-day, for processor book FC 4702 or FC 7562: After the On/Off Processor function is enabled in a system, you must report your On/Off usage to IBM at least monthly. This information, used to compute your billing data, is provided to your sales channel. The sales channel will place an order on your behalf for the quantity of this feature that matches your reported use. One FC 4709 provides 1 day of On/Off processor billing for 0/32-core POWER7 CoD Processor Book FC 4702 or FC 7562.	✓	—	✓
4712	On/Off IBM i processing CoD billing, 100 processing-day, for processor book FC 4700 or FC 7560: After the On/Off Processor function is enabled in a system, you must report your On/Off usage to IBM at least monthly. This information, used to compute your billing data, is provided to your sales channel. The sales channel will place an order on your behalf for the quantity of this feature that matches your reported use. One FC 4709 provides one day of on/off processor billing for 0/32-core POWER7 CoD Processor Book FC 4700 or FC 7560.	—	✓	—
4713	Single core activation for POWER7 CoD processor book FC 4700: This feature permanently activates one processor core of POWER7 CoD processor book FC 4700.	✓	✓	✓
4714	Single core activation for POWER7 CoD processor book FC 4702: This feature permanently activates one processor core of POWER7 CoD processor book FC 4702.	✓	✓	✓
4717	Sixty four core activations for POWER7 CoD processor book FC 4700: This feature permanently activates 64 processor cores of POWER7 CoD processor book FC 4700 (64 x FC 4713).	✓	✓	✓
4718	Sixty four core activations for POWER7 CoD processor book FC 4702: This feature permanently activates 64 processor cores of POWER7 CoD processor book FC 4700 (64 x FC 4714)	✓	✓	✓
4719	Billing for 100 processor minutes usage of processor book FC 4700, IBM i: Utility billing for 100 processor minutes for Processor, FC 4700 with IBM i.	—	✓	—

Feature code	Description	Support		
		AIX	IBM i	Linux
4721	On/Off processor CoD billing, 1 proc-day for processor book FC 4700 or FC 7560, IBM i: After the On/Off Processor function is enabled in a system, you must report your on/off usage to IBM at least monthly. This information, used to compute your billing data, is provided to your sales channel. The sales channel will place an order on your behalf for the quantity of this feature that matches your reported use. One FC 4721 provides one day of on/off processer billing for Processor Book FC 4700 or FC 7560 with IBM i.	—	✓	—
4722	Billing for 100 processor minutes usage of processor book FC 4702 with IBM i: utility billing for 100 processor minutes for Processor, FC 4702 with IBM i.	—	✓	—
4724	On/Off processor CoD billing, 1 proc-day for processor book FC 4702 or FC 7562, IBM i: After the On/Off Processor function is enabled in a system, you must report your On/Off usage to IBM at least monthly. This information, used to compute your billing data, is provided to your sales channel. The sales channel will place an order on your behalf for the quantity of this feature that matches your reported use. One FC 4724 provides one day of on/off processer billing for Processor Book FC 4702 or FC 7562 with IBM i.	—	✓	—
EP9T	FC EP9T provides access to on/off, inactive CoD processor resources for 90 days. Access to these resources is measured in processor-days. For example, if there are 32-cores of inactive, CoD processors installed, this feature will enable using 2880 proc-days before disabling CoD access to the inactive CoD processors. On/Off CoD enablement code does not ship with FC EP9T because customer contracts and Sales Channel registration are required prior to processing a miscellaneous equipment specification (MES) order for FC EP9T. After Sales Channel registration is completed and customer contracts are in force, the Power CoD Project Office will send an enablement key to enable activation of inactive CoD processors. After usage of this feature is exhausted, additional temporary usage of on/off CoD processor resources may be activated by ordering another on/off CoD enablement code, FC EP9T. Before ordering a second enablement code, the current code must be deleted (recovery point objective, RPO) from the configuration file with a MES RPO delete order. Usage of on/off temporary CoD processors is billed by ordering either on/off processor CoD billing for 100 processor-days or 1 processor-day. A different on/off billing feature code is available for each processor feature that is installed in a 9117, 9119, or 9179 MT.	✓	✓	✓

Feature code	Description	Support		
		AIX	IBM i	Linux
EPJ0	FC EPJ0 provides access to 48 processor-days of on/off, inactive CoD processor resources per each processor feature code ordered (FC 4700 or FC EPH2). For example, if there are available 24 inactive, CoD processor cores, FC EPJ0 will enable 2 days of usage of the inactive CoD cores. FC EPJ0 is ordered with a maximum quantity of 10 per processor feature code ordered. This increases the available usage of on/off CoD processor-days for 24 inactive CoD cores 2 - 20 days before access to the inactive CoD processor cores are disabled. FC EPJ0 and EMJ0 (768 GB-Days of On/Off CoD Temporary Memory Resources) may be ordered together. On/Off CoD enablement code does not ship with FC EPJ0, because customer contracts and Sales Channel registration are required prior to enabling FC EPJ0. Prior to enabling FC EMJ0, customers must use MES to order FC EP9T - 90 Days On/Off CoD Temporary Processor Core Enablement. After usage of this feature is exhausted, additional temporary usage of on/off CoD processor resources may be activated by ordering another on/off CoD enablement code. Before ordering another enablement code, the current code must be deleted from the configuration file with an MES RPO delete order. Usage of on/off temporary CoD processor cores is billed by ordering on/off processing CoD billing for 100 processor-days or 1 processor-day. A different on/off billing feature code is available for each processor feature that is installed in a 9117, 9119 or 9179 MT.	✓	✓	✓
EPJ1	FC EPJ1 provides access to 36 processor-days of on/off, inactive CoD processor cores per each processor feature code ordered (FC 4702). For example, if 18 inactive, CoD processor cores are available, this feature will enable two days of usage of the inactive CoD cores before disabling access to the inactive CoD cores. FC EPJ1 is ordered with a maximum quantity of 10 per processor feature code ordered. This increases the available usage of on/off processor-days for 18 inactive CoD cores 2 - 20 days before access to the inactive CoD cores would be disabled. FC EPJ1 may be ordered with EMJ1 (576 GB-Days of On/Off CoD Temporary Memory Resources). On/Off CoD enablement code does is not included with FC EPJ1, because customer contracts and Sales Channel registration are required prior to enabling FC EPJ1. Prior to enabling FC EPJ1, the customer must use MES to order FC EP9T - 90 Days On/Off CoD Temporary Processor Core Enablement. After usage of this feature is exhausted, additional temporary usage of on/off CoD processor resources may be activated by ordering another on/off CoD enablement code. Before ordering another enablement code, the current code must be deleted from the configuration file with an MES RPO delete order. Usage of on/off temporary CoD processor cores is billed by ordering on/off processing CoD billing for 100 processor-days or 1 processor-day. A different on/off billing feature code is available for each processor feature that is installed in a 9117, 9119 or 9179 MT.	✓	✓	✓

1.4.3 Memory features

The Power 795 uses buffered DDR3 memory cards. Each processor book provides 32 memory card slots for a maximum of 256 memory cards per server. The minimum system memory is two memory features (8 DIMMs) per processor book.

Figure 1-2 outlines the general connectivity of the POWER7 processor and DDR3 memory DIMMS.

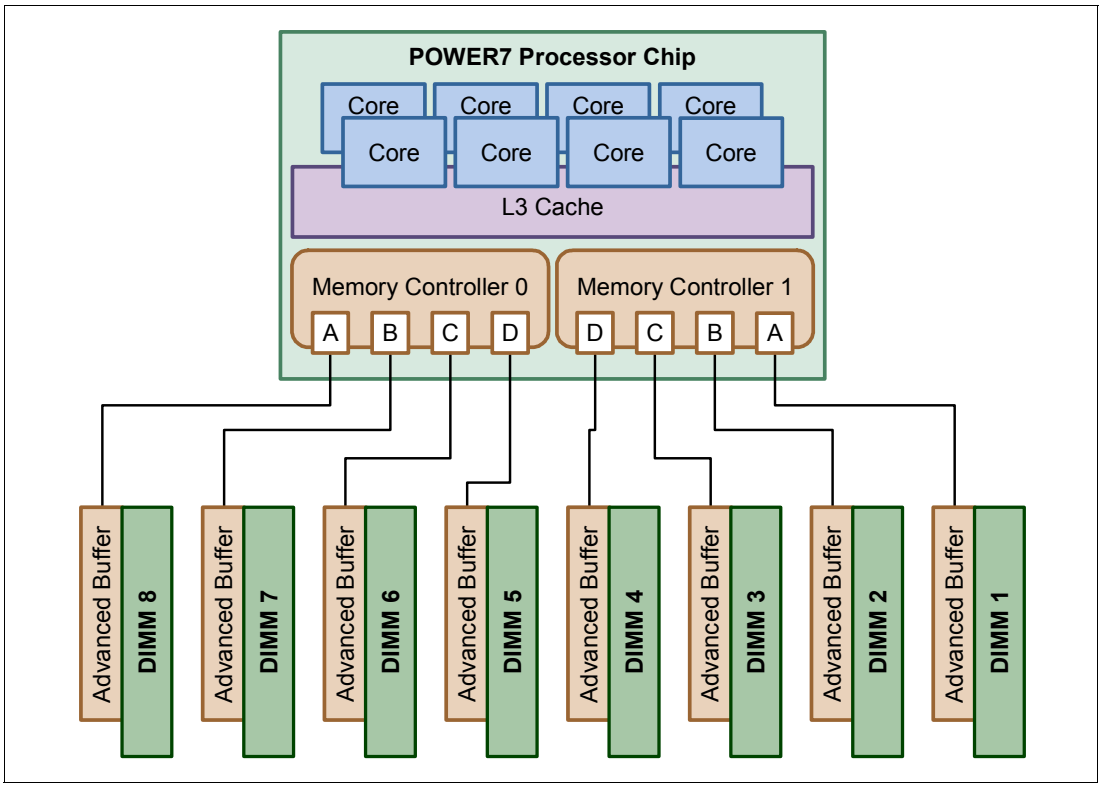

Figure 1-2 Outline of POWER7 memory connectivity to DDR3 DIMMs

On each processor book for the Power 795, there are four POWER7 single-chip modules (SCMs), which allow a total of 32 DDR3 memory DIMM slots to be connected (eight DIMM slots per processor chip).

The quad-high (96 mm) DIMM cards can have an 8 GB, 16 GB, 32 GB, or 64 GB capacity and are connected to the POWER7 processor memory controller through an advanced memory buffer. For each DIMM there is a corresponding memory buffer. Each memory channel into the POWER7 memory controllers is driven at 6.4 GHz.

Unsupported: DDR2 DIMMs (used in IBM POWER6® based systems) are not supported in POWER7 based systems.

The Power 795 has memory features in 32 GB, 64 GB, 128 GB, and 256 GB capacities. Table 1-12 summarizes the capacities of the memory features and highlights other characteristics.

Table 1-12 Summary of memory features

Feature code	Memory technology	Capacity	Access rate	DIMMs	DIMM slots used
5600	DDR3	32 GB	1066 MHz	4 x 8 GB DIMMs	4
5601	DDR3	64 GB	1066 MHz	4 x 16 GB DIMMs	4
5602	DDR3	128 GB	1066 MHz	4 x 32 GB DIMMs	4
5564	DDR3	256 GB	1066 MHz	4 x 64 GB DIMMs	4

Memory for the Power 795 can also be purchased in bundles. The available memory bundles are summarized in Table 1-13.

Table 1-13 Summary of memory features

Feature code	Capacity	Equivalent features	DIMMs	DIMM slots used
8211	256 GB	8 x FC 5600	32 x 8 GB DIMMs	32
8214	512 GB	16 x FC 5600	64 x 8 GB DIMMs	64
8218	256 GB	4 x FC 5601	16 x 16 GB DIMMs	16
8219	512 GB	8 x FC 5601	32 x 16 GB DIMMs	32
8221	1024 GB	8 x FC 5602	32 x 32 GB DIMMs	32

None of the memory in these features is active. Table 1-14 outlines the memory activation feature codes and corresponding memory capacity activations.

Table 1-14 CoD system memory activation features

Feature code	Activation capacity	Additional information	Support		
			AIX	IBM i	Linux
8212	1 GB	Activation of 1 GB of DDR3 POWER7 memory.	✓	✓	✓
8213	100 GB	Activation of 100 GB of DDR3 POWER7 memory.	✓	✓	✓
7973	N/A	On/Off Memory Enablement: Can be ordered to enable your server for On/Off Capacity on Demand. After it is enabled, you can request memory on a temporary basis. You must sign an On/Off Capacity on Demand contract before you order this feature.	✓	✓	✓
7377	N/A	On/Off, 1 GB-1Day, Memory Billing POWER7: After the On/Off Memory function is enabled in a system, you must report the client's on/off usage to IBM on a monthly basis. This information is used to compute IBM billing data. Order one FC 7377 feature for each billable day for each 1 GB increment of POWER7 memory that was used. **Note:** inactive memory must be available in the system for temporary use.	✓	✓	✓

Feature code	Activation capacity	Additional information	Support		
			AIX	IBM i	Linux
EM9T		FC EM9T provides access to on/off, inactive CoD memory resources for 90 days. Access to these resources is measured in GB-days. For example, if there are 100 GB of inactive, CoD memory installed, this feature will enable usage of 9000 GB days (9 TB days) before disabling CoD access to the inactive memory. On/Off CoD enablement code is not included with FC EM9T because customer contracts and Sales Channel registration are required prior to processing an MES order for FC EM9T. After Sales Channel registration is completed and customer contracts are in force, the Power CoD Project Office will send an enablement key to enable activation of inactive CoD memory. After usage of this feature has been exhausted, additional temporary usage of on/off CoD memory resources may be activated by ordering another On/Off CoD enablement code, FC EM9T. Before ordering another enablement code, the current code must be RPO deleted from the configuration file with a MES delete order. Usage of on/off temporary CoD memory is billed by ordering FC 4710 On/Off 999 GB-Days, Memory Billing POWER7 or FC 7377 - On/Off, 1GB-1Day, Memory Billing POWER7.	✓	✓	✓
EMJ0	N/A	FC EMJ0 provides access to 768 GB-days of on/off, inactive CoD memory resources per each processor feature code ordered (FC 4700 or FC EPH2). For example, if 100 GB of inactive, CoD memory is available, FC EMJ0 will enable 7 days of usage of the inactive CoD memory. FC EMJ0 is ordered with a maximum quantity of 10 per processor feature code ordered. This increases the available usage of on/off CoD GB-Days for 100 GB of inactive CoD memory 7 - 70 days before access to the inactive CoD memory would be disabled. FC EMJ0 and EPJ0 (48 Proc-Days of On/Off CoD Temporary Processor Resources) may be ordered together. FC EM9T (90 Days On/Off CoD Temporary Memory Enablement) must be ordered through MES to enable FC EMJ0. The on/off CoD enablement code is not included with EMJ0, because customer contracts and Sales Channel registration are required to enable FC EMJ0. After usage of this feature is exhausted, additional temporary usage of on/off CoD memory resources may be activated by using MES to order another On/Off CoD enablement code, FC EM9T. Before ordering another FC EM9T, the current code must be RPO deleted from the configuration file with a MES delete order. Usage of on/off temporary CoD memory is billed by ordering FC 4710 (On/Off 999 GB-Days, Memory Billing POWER7) or FC 7377 (On/Off, 1GB-1Day, Memory Billing POWER7).	✓	✓	✓

Feature code	Activation capacity	Additional information	Support		
			AIX	IBM i	Linux
EMJ1	N/A	FC EMJ1 provides access to 576 GB-days of on/off, inactive CoD memory resources per each processor feature code ordered (FC 4702). For example, if 100 GB of inactive, CoD memory is available, this feature will enable 5 days of usage of the inactive CoD memory before disabling access to the inactive CoD memory. FC EMJ1 is ordered with a maximum quantity of 10 per processor feature code ordered. This increases the available usage of on/off CoD GB-days for 100 GB of inactive CoD memory 5 - 50 days before access to the inactive CoD memory is disabled. FC EMJ1 may be ordered with EPJ1 (36 Proc-Days of On/Off CoD Temporary CoD Processor Resources). Feature code EM9T (90 Days On/Off CoD Temporary Memory Enablement) must be ordered using MES to enable FC EMJ1. On/Off CoD enablement code is not included with EMJ1, because customer contracts and Sales Channel registration are required to enable FC EMJ1. After usage of this feature has been exhausted, additional temporary usage of on/off CoD memory resources may be activated by using MES to order another On/Off CoD enablement code, FC EM9T. Before ordering another FC EM9T, the current code must be RPO deleted from the configuration file with an MES delete order. Usage of on/off temporary CoD memory is billed by ordering FC 4710 (On/Off 999 GB-Days, Memory Billing POWER7) or FC 7377 (On/Off, 1GB-1Day, Memory Billing POWER7).	✓	✓	✓

1.4.4 Memory minimums and maximums

The Power 795 has the following minimum and maximum configurable memory resource allocation requirements:

- Minimum memory activation is the greatest of the following values:
  - 32 GB
  - 2 GB per active processor core
  - 50% of installed memory (FC 5600), 75% of installed memory (FC 5602, FC 8221)
- The Power 795 supports a maximum of 16 TB DDR3 memory.
- Memory must be configured with a minimum of two identical memory features per processor book.
- Memory features FC 5600, FC 5601, FC 5602 and FC 5564 can be mixed within the same POWER7 processor book.
- Memory features must be ordered in matched pairs.
- Memory can be activated in increments of 1 GB.
- Features FC 8214, FC 8218, FC 8219, and FC 8221 are memory bundles.

1.5 I/O drawers for the Power 795 server

The Power 795 server supports 12X attached I/O drawers, providing extensive capability to connect to network and storage devices. A Power 795 can be equipped with up to four expansion racks and up to 32 I/O drawers of the following options:

- 12X I/O Drawer PCIe, SFF Disk (FC 5803)
- 12X I/O Drawer PCIe, no Disk (FC 5873)
- 12X I/O Drawer PCI-X, with repeater (FC 5797), withdrawn
- 12X I/O Drawer PCI-X, no repeater (FC 5798), withdrawn

Disk-only I/O drawers are also supported, providing large storage capacity and multiple partition support:

- FC 5886 EXP 12S holds a 3.5-inch SAS disk or SSD
- FC 5887 EXP24S SFF Gen2-bay Drawer for high-density storage holds SAS Hard Disk drives.

The 12X I/O drawers are connected to the processor books by a GX Dual-port 12X HCA adapter (FC 1816). Each adapter has two ports and can support one 12X loop. Up to four of the adapters can be installed in each 8-core processor book. Each half of the foregoing 12X I/O drawers is powered separately for redundancy

The I/O drawers of a Power 795 are always connected to the processor books in loops to make the configuration more reliable. Systems with nonlooped configurations can also experience degraded performance and serviceability.

I/O drawers can be connected to the processor book in either single-loop or dual-loop mode.

- Single-loop mode connects an entire I/O drawer to the processor book 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 processor book 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.

Dual-loop mode is preferable when possible, because it provides the maximum bandwidth between the I/O drawer and the processor book and also independent paths to each of the I/O drawer planars.

Table 1-15 lists the maximum number of I/O drawers that can be attached to the Power 795.

Table 1-15 Maximum number of I/O drawers for feature codes

Server	Maximum of FC 5803	Maximum of FC 5873	Maximum of FC 5797	Maximum of FC 5798	Total maximum of all I/O drawers
Power 795	32	31	30	2	32

1.5.1 12X I/O Drawer PCIe, SFF disk

The 12X I/O Drawer with SFF disks (FC 5803) is a 4U-high I/O drawer that contains 20 PCIe slots and 26 SAS hot-swap small-form factor (SFF) disk bays. This drawer attaches to the CEC by 12X DDR cables (FC 1862, FC 1863, or FC 1864). When SFF disk are installed, they are driven by at least one SAS PCIe adapter and SAS AT cable (FC 3688). Using a mode switch, the 26 SFF slots can be configured as one group of 26 slots (AIX and Linux) or two groups of 13 slots (AIX, IBM i, and Linux) or four groups of six or seven slots (AIX and Linux).

1.5.2 12X I/O Drawer PCIe, no disk

The 12X I/O Drawer (FC 5873) is equivalent to the FC 5803 drawer and provides a 4U-high I/O drawer that contains 20 PCIe slots, but has no disk bays. This drawer attaches to the CEC by 12X DDR cables (FC 1862, FC 1863, or FC 1864).

1.5.3 12X I/O Drawer PCI-X, with repeater

The 12X I/O Drawer PCI-X with repeater (FC 5797) provides a 4U-high I/O drawer that contains 20 PCI-X slots and 16 hot-swap SCSI disk bays. This drawer attaches to the CEC by 12X attachment cables and includes an installed repeater card. The repeater card is designed to strengthen signal strength over the longer 12X cables used with the powered expansion rack (FC 6954) and nonpowered expansion rack (FC 6983).

Unsupported: IBM i does not support SCSI drives in the FC 5797 drawer.

1.5.4 12X I/O Drawer PCI-X, no repeater

The 12X I/O Drawer PCI-X (FC 5798) is equivalent to the FC 5797 drawer and provides a 4U-high I/O drawer that contains 20 PCI-X slots and 16 hot-swap SCSI disk bays. This drawer is configured without a repeater and can be used only in the primary system rack. It is not intended for use in the powered expansion rack (FC 6954) or bolt-on expansion rack (FC 6953), thus limiting the FC 5798 drawer to a maximum of two drawers per system. This drawer attaches to the CEC by 12X cables.

Unsupported: IBM i does not support SCSI drives in the FC 5798 drawer.

1.5.5 EXP 12S SAS Drawer

The EXP 12S SAS drawer (FC 5886) is a 2-EIA drawer and mounts in a 19-inch rack. The drawer can hold either SAS disk drives or SSD. The EXP 12S SAS drawer has twelve 3.5-inch SAS disk bays with redundant data paths to each bay. The SAS disk drives or SSDs that are contained in the EXP12S are controlled by one or two PCIe or PCI-X SAS adapters that are connected to the EXP12S by SAS cables.

FC 5886 can also be directly attached to the SAS port on the rear of the Power 720, providing a low cost disk storage solution. When used this way, the imbedded SAS controllers in the system unit control the disk drives in EXP12S. A second unit cannot be cascaded to FC 5886 attached in this way.

1.5.6 EXP24S SFF Gen2-bay Drawer

The EXP24S SFF Gen2-bay Drawer (FC 5887) is an expansion drawer with twenty-four 2.5-inch form factor (SFF) SAS bays. It supports up to 24 hot-swap SFF SAS hard disk drives (HDDs) on POWER6 or POWER7 servers in 2U of 19-inch rack space. The EXP24S bays are controlled by SAS adapters or controllers that attached to the I/O drawer by SAS X or Y cables.

The SFF bays of the EXP24S different from the SFF bays of the POWER7 system units or 12X PCIe I/O Drawers (FC 5802, FC 5803). The EXP24S uses Gen-2 or SFF-2 SAS drives that physically do not fit in the Gen-1 or SFF-1 bays of the POWER7 system unit or 12X PCIe I/O Drawers, or vice versa

1.5.7 Supported I/O drawers after migration from Power 595 to Power 795

The following drawers present on a Power 595 system can be moved to a Power 795 server:

- 12X I/O Drawer PCIe, SFF disk (FC 5803)
- 12X I/O Drawer PCIe, no disk (FC 5873)
- 12X I/O Drawer PCI-X, with repeater (FC 5797)
- 12X I/O Drawer PCI-X, without repeater (FC 5798)

Moving to an expansion rack: If the FC 5798 drawer is moved to an expansion rack, it must be converted to a FC 5797.

- EXP 12S SAS Disk Drawer (FC 5886)
- IBM TotalStorage EXP24 SCSI Disk Drawer (FC 5786)

Also, any PCI adapters that are contained in the previous remote I/O and high-speed link (RIO/HSL) drawers that are supported on the Power 795 must be moved to a 12X I/O drawer or be replaced by newer I/O.

Unsupported: All I/O drawers that are attached to previous Power 595, p5 595, and p5 590 servers through an RIO/HSL link cannot be reused on the Power 795. The following RIO/HSL I/O towers and drawers are not supported in a Power 795 server:

- 7040-61D I/O drawers
- PCI Expansion Drawer, 19-inch PCI-X (FC 5790)
- I/O Drawer, 20 Slots, 16 Disk Bays (FC 5791)
- 12X I/O Drawer PCI-X, no repeater (FC 5798)
- Model Upgrade Carry-Over Indicator for converted FC 4643 with DCA (FC 5809)
- PCI/SCSI Disk Expansion Drawer (FC 0595)
- PCI-X Expansion Unit (FC 5088)
- PCI Expansion Drawer (FC 5790)
- PCI-X Expansion Tower/Unit (FC 5094)
- PCI-X Expansion Tower, no disk (FC 5096, FC 5088)
- 1.8 m I/O Tower (FC 5294)
- 1.8 m I/O Tower (no disk) (FC 5296)
- PCI-X Expansion Unit in Rack (FC 0588)
- PCI-X Tower Unit in Rack (FC 0595)

1.6 Power 795 model upgrade

Customers with installed 9119-FHA servers can increase their computing power by ordering a model conversion to the Power 795 server. You can upgrade the IBM POWER6 9119-FHA server to an IBM POWER7 9119-FHB server with POWER7 processors and POWER7 DDR3 memory.

For upgrades from POWER6 processor-based systems, IBM will install new Power 795 processor books and memory in the 9119-FHA CEC frame along with other POWER7 components. The current POWER6 Bulk Power Distribution and Bulk Regulator Assemblies will not be replaced.

From the Power 595 CEC frame, the Bulk Power Regulators (FC 6333), Bulk Power Distributors (FC 6334), and the 12X PCI-X and PCIe I/O drawers are allowed to migrate to the upgraded Power 795 CEC.

Any older RIO/HSL I/O drawers (such as the FC 5790, FC 0595, FC 5094/5294, FC 0588 used in 9119-FHA) are not supported on the Power 795. Also, any PCI adapters that are contained in older RIO/HSL drawers, which are supported on the Power 795, must be moved to a 12X I/O drawer or be replaced by newer I/O adapters.

Along with PCI-X and PCIe I/O drawers that are migrating to the new upgraded 795, both the Power 595 24-inch self-powered expansion and bolt-on racks (FC 6953 and FC 6954) can be migrated.

The 19-inch, 2M expansion rack (FC 0553) can migrate with the foregoing model upgrade. The 19-inch, 2EIA SAS I/O drawer (FC 5886) and TotalStorage EXP24 disk drawer (FC 5786) can migrate with the 19-inch rack. Note that FC 5786 Disk Drawers are not available for new orders.

The ability to restore all Power 595 partitions on the Power 795 requires considerations:

- The number of cores that are available on the Power 795 system must be greater than or equal to the number of cores that are required for all LPARs on the Power 595 system.
- Memory requirements for the Power 795 are 15% (approximately) greater than the amount in use on the Power 595 system.
- Growth in memory requirements is driven by default enablement of SMM and increased use of memory to support new features that are typical in new releases (for example, hypervisor mirroring).
- Only hubs that assigned and connected to GX InfiniBand (IB) and that are in use on the Power 595 system are migrated from the Power 595 to the Power 795.
- There is no RIO support on the Power 795, and the RIO drawers must be removed.
- Partitions redefined to utilize only IB attached I/O and the Power 595 system must be IPLed and the partition profiles saved following removal of the RIO attached drawers before commencing upgrade to the Power 795.
- If the MES order results in fewer physical GX slots than GX IB adapters used on the Power 595 system, detailed review must be done.
- If equivalent amounts of memory, cores, I/O, or GX hubs are not available following the upgrade, the customer might still be able to migrate all partitions (for example, LPARs using shared processor pools, LPARs with lower minimum memory limits that fit within the available resources, and reconfiguration of I/O).
- Clients must be alerted to the possibility that part or all of their partitions might need to be manually recreated on the Power 795.

1.6.1 Upgrade considerations

Each model conversion order also requires feature conversions for the following purposes:

- Update machine configuration records
- Ship system components as necessary
- Return converted features to IBM

Table 1-16 lists the feature conversions that are set up for the upgrade.

Table 1-16 Features conversions

From	Feature code	To	Feature code
9119-FHA	4694	9119-FHB	4700
9119-FHA	4695	9119-FHB	4700
9119-FHA	4705	9119-FHB	4700
9119-FHA	4694	9119-FHB	4702
9119-FHA	4754	9119-FHB	4713
9119-FHA	4755	9119-FHB	4713
9119-FHA	4754	9119-FHB	4714
9119-FHA	4694	9119-FHB	4700
9119-FHA	5693	9119-FHB	5600
9119-FHA	5694	9119-FHB	5600
9119-FHA	5695	9119-FHB	5600
9119-FHA	5694	9119-FHB	5601
9119-FHA	5695	9119-FHB	5601
9119-FHA	5696	9119-FHB	5601
9119-FHA	5695	9119-FHB	5602
9119-FHA	5696	9119-FHB	5602
9119-FHA	5697	9119-FHB	5602
9119-FHA	5680	9119-FHB	8212
9119-FHA	5681	9119-FHB	8213
9119-FHA	5684	9119-FHB	8213
9119-FHA	8201	9119-FHB	8211
9119-FHA	8202	9119-FHB	8211
9119-FHA	8203	9119-FHB	8214
9119-FHA	8202	9119-FHB	8218
9119-FHA	8203	9119-FHB	8219
9119-FHA	8204	9119-FHB	8219
9119-FHA	8203	9119-FHB	8221

From	Feature code	To	Feature code
9119-FHA	8204	9119-FHB	8221
9119-FHA	6865	9119-FHB	6867
9119-FHA	6870	9119-FHB	6888
9119-FHA	6869	9119-FHB	6868
9119-FHA	6866	9119-FHB	6887
9119-FHA	6941	9119-FHB	6821
9119-FHA	6331	9119-FHB	6259
9119-FHA	6961	9119-FHB	6991
9119-FHA	6962	9119-FHB	6992
9119-FHA	6963	9119-FHB	6993
9119-FHA	6964	9119-FHB	6994
9119-FHA	6965	9119-FHB	6995
9119-FHA	6966	9119-FHB	6996
9119-FHA	6967	9119-FHB	6997
9119-FHA	6968	9119-FHB	6998
9119-FHA	6865	9119-FHB	6867
9119-FHA	6870	9119-FHB	6888

1.6.2 Moving features

The following features that are present on a Power 595 system can be moved to Power 795:

- PCIe, PCI-X adapters with cables when supported on Power 595
- SCSI 15K RPM drives (FC 3278, FC 3279, FC 4328, FC 4329, FC 3585, and FC 4327)
- SCSI 10K RPM drives (FC 1882, FC 3159)
- Line cords, keyboards, and displays
- PowerVM (FC 7943 and FC 8002)
- Racks (FC 6953, FC 6954, FC 0553, FC 0551, FC 0554, and FC 0555)
- Bolt-on Rack Doors (FC 6880 and FC 6878)
- 1.8m Rack Doors (FC 6248)
- IBM 4.7 GB IDE Slimline DVD-RAM Drive (FC 5757)

The following drawers that are present on Power 595 system can be moved to Power 795:

- FC 5803 12X I/O Drawer PCIe, SFF disk
- FC 5873 12X I/O Drawer PCIe, no disk
- FC 5797 12X I/O Drawer PCI-X, with repeater
- FC 5798 12X I/O Drawer PCI-X, without repeater
- FC 5886 EXP 12S SAS Disk Drawer
- FC 5786 TotalStorage EXP24 SCSI Disk Drawer

Moving FC 5798 to an expansion rack: If moved to an expansion rack, the PCI-X expansion drawer type FC 5798 must be converted to a type FC 5797.

The following 24-inch racks and drawers are *not* supported with Power 795 upgrades:

- 7040-61D I/O drawers
- FC 5791: I/O Drawer, 20 Slots, 16 Disk Bays
- FC 5809: Model Upgrade Carry-Over Indicator for converted FC 4643 with DCA

The following additional 19-inch racks are supported with 795 upgrades:

- FC 0551 36 EIA unit 19-inch rack (7014-T00)
- FC 0555 25 EIA Unit Mini Rack (7014-S25)

The following RIO/HSL-attached I/O towers and drawers are *not* supported:

- FC 0595 PCI/SCSI Disk Expansion Drawer
- FC 0588 PCI-X Expansion Unit in Rack
- FC 5088 PCI-X Expansion Unit
- FC 5094 PCI-X Expansion Tower
- FC 5096 PCI-X Expansion Tower (no disk)
- FC 5294 PCI-X 1.8 m Expansion Tower
- FC 5296 PCI-X 1.8 m Expansion Tower
- FC 5790: PCI Expansion Drawer

Notes:

- For Power 595 servers that have the on/off CoD function enabled, you must reorder the on/off enablement features (FC 7971 and FC 7973) when placing the upgrade MES order for the new Power 795 server to keep the On/Off CoD function active.

  Remove the On/Off enablement features from the configuration file before the MES order is started to initiate the model upgrade.
- Any temporary use of processors or memory owed to IBM on the existing system must be paid before installing the new Power 795. Use up an existing credit days/minutes before upgrading because there is no product structure to convert Power 595 processor or memory days or processor minutes to Power 795 units.

1.7 Management consoles

This section describes the available management consoles.

1.7.1 Hardware Management Console models

A Hardware Management Console (HMC) is required to manage a Power 795 system. Multiple POWER6 and POWER7 processor-based servers can be supported by a single HMC. The HMC must be either a rack-mount CR3 (or later), or deskside C05 (or later).

IBM announced an HMC model, with machine type 7042-CR7. Hardware features on the CR7 model include a second hard disk drive (HDD), FC 1998, for RAID 1 data mirroring, and the option of a redundant power supply. At the time of writing, the latest version of HMC code is V7.6.0. This code level is required for new LPAR function support that allows the HMC to manage more LPARs per processor core; a core can now be partitioned into up to 20 LPARs (0.05 of a core).

The HMC provides a set of functions that can be used to manage the system:

- Creating and maintaining a multiple partition environment
- Displaying a virtual operating system session terminal for each partition

- Displaying a virtual operator panel for each partition
- Detecting, reporting, and storing changes in hardware conditions
- Powering managed systems on and off
- Acting as IBM Service Focal Point™ for service representatives to determine an appropriate service strategy

Several HMC models are supported to manage POWER7 processor-based systems. Licensed Machine Code Version 7 Revision 710 (FC 0962) is required to support POWER7 processor-based servers, in addition to IBM POWER5, IBM POWER5+, and IBM POWER6 processor-based servers. HMC model 7042-CR7 is available for ordering, but you can also use one of the withdrawn models listed in Table 1-17.

Table 1-17 HMC models supporting POWER7 processor technology-based servers

Type-model	Availability	Description
7310-C05	Withdrawn	IBM 7310 Model C05 Desktop Hardware Management Console
7310-C06	Withdrawn	IBM 7310 Model C06 Deskside Hardware Management Console
7042-C06	Withdrawn	IBM 7042 Model C06 Deskside Hardware Management Console
7042-C07	Withdrawn	IBM 7042 Model C07 Deskside Hardware Management Console
7042-C08	Withdrawn	IBM 7042 Model C08 Deskside Hardware Management Console
7310-CR3	Withdrawn	IBM 7310 Model CR3 Rack-mounted Hardware Management Console
7042-CR4	Withdrawn	IBM 7042 Model CR4 Rack-mounted Hardware Management Console
7042-CR5	Withdrawn	IBM 7042 Model CR5 Rack-mounted Hardware Management Console
7042-CR6	Withdrawn	IBM 7042 Model CR6 Rack-mounted Hardware Management Console
7042-CR7	Available	IBM 7042 Model CR7 Rack-mounted Hardware Management Console

For the IBM Power 795, the Licensed Machine Code Version 7 Revision 720 is required, and an HMC is required, even if you plan to use a full system partition.

HMC code: You can download or order the latest HMC code from the Fix Central website:

`http://www.ibm.com/support/fixcentral`

Existing HMC models 7310 can be upgraded to Licensed Machine Code Version 7 to support environments that can include POWER5, POWER5+, POWER6, and POWER7 processor-based servers. Licensed Machine Code Version 6 (FC 0961) is not available for 7042 HMCs.

When the IBM Systems Director is used to manage an HMC or if the HMC manages more than 254 partitions, the HMC must have 3 GB of RAM minimum and be a rack-mount CR3 model, or later, or deskside C06, or later.

1.7.2 IBM SDMC

IBM withdrew the SDMC product. Customers should migrate to the HMC platform by upgrading their 7042-CR6 server to the latest HMC code. See *IBM Power Systems: SDMC to HMC Migration Guide (RAID1)*, REDP-4872.

1.8 System racks

The Power 795 is designed to mount in a primary system rack with a redundant integrated power subsystem. It provides 24-inch 42U of rack space and houses the processors, memory, media drawer, and up to three I/O drawers.

A self-powered expansion rack (FC 6954) is available for large system configurations that require more 24-inch I/O drawer expansion than is available in the primary system rack. It provides an identical redundant power subsystem as what is available in the primary rack. Several types of I/O drawers can be used with rack FC 6954.

If additional I/O drawers are required, a nonpowered expansion rack (FC 6953) is available. Installing the nonpowered expansion rack is accomplished as follows:

1. The side cover of the powered expansion rack (FC 6954) is removed
2. The nonpowered expansion rack (FC 6953) is bolted to the side
3. The side cover is placed on the exposed side of the nonpowered expansion rack (FC 6953).

Power for components in the expansion rack is provided from the bulk power assemblies in the powered expansion rack.

Up to two powered expansion racks (FC 6954) and nonpowered expansion racks (FC 6953) are available to support up to 32 I/O drawers in a Power 795 server.

Nonpowered rack: One nonpowered expansion rack (FC 6953) can be attached to each powered expansion rack (FC 6954). The nonpowered expansion rack cannot be attached to the primary system rack of the Power 795.

If additional external communication and storage devices are required, a 19-inch, 42U nonpowered expansion rack (FC 0533) can be ordered. The FC 0533 is equivalent to the 7014-T42 rack, which is supported for use with a 9119-FHA server. Some features that are delivered as part of the 7014-T42 must be ordered separately with the FC 0553.

1.8.1 Door kit

All 24-inch system racks and expansion racks must have door assemblies installed. Door kits containing front and rear doors are available in either slimline, acoustic, or acoustic rear heat exchanger styles:

- The slimline door kit (FC 6868) provides a smaller footprint alternative to the acoustical doors for customers who have less floor space. The doors are slimmer because they do not contain acoustical treatment to attenuate the noise. This door option is the default.
- The acoustical door kit (FC 6867) provides specially designed front and rear acoustical doors that 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.
- Acoustical front door and rear door heat exchanger kit (FC 6887) 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 (RDHX) 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.

- CEC acoustic doors, front and rear (FC ERG1) provide front and rear acoustical doors for use with the CEC rack. This feature provides specially designed 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. The front door is steel and includes the Power 795 logo. The steel front door has a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack.
- Self-powered expansion rack acoustic doors, front and rear (FC ERG2) provides front and rear acoustical doors for use with the FC 6954 24-inch, self-powered expansion rack. This feature provides specially designed 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. The front door is steel and includes the Power 795 logo. The steel front door has a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack.
- Bolt-on expansion rack acoustic doors, front and rear (FC ERG3) provides front and rear acoustical doors for use with FC 6953 24-inch, nonpowered, bolt-on expansion rack. This feature provides specially designed 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. The front door is steel and has a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack.
- Front acoustic door only for CEC rack (FC ERG4) provides the front door only of the front and rear door set shipped with the CEC rack acoustical door set FC ERG1. This feature provides a specially designed front door that greatly reduces the noise emissions from the system and thereby lower the noise levels in the data center. The front door includes a unique air inlet and exhaust ducts to attenuate the noise. The front door is steel and includes the Power 795 logo. The steel front door has a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack.
- Front acoustic door only for the self-powered expansion rack (FC ERG5) provides the front door only of the FC ERG2 acoustical front and rear door set that is included with the FC 6954 self-powered expansion rack. This feature provides a specially designed front door that greatly reduces the noise emissions from the system and thereby lower the noise levels in the data center. The front door includes a unique air inlet and exhaust ducts to attenuate the noise. The front door is steel and includes the Power 795 logo. The steel front door has a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack.
- Front acoustic door only for the bolt-on expansion rack (FC ERG6) provides the front door only of the FC ERG3 acoustical front and rear door set shipped with the FC 6953 bolt-on expansion rack. This feature provides a specially designed front door that greatly reduces the noise emissions from the system and thereby lower the noise levels in the data center. The front door includes a unique air inlet and exhaust ducts to attenuate the noise.

1.8.2 Power supplies

The primary system rack and powered expansion rack always incorporate two bulk power assemblies for redundancy. These provide power for devices located in those racks and associated nonpowered expansion racks. Both the primary system CEC and self-powered expansion racks can be serviced by either AC or DC power. A Power 795 rack, which is upgraded from a Power 595, can be serviced only by AC power.

These bulk power assemblies are mounted in front and rear positions and occupy the top eight EIA positions of the rack. To help provide optimum system availability, these bulk power assemblies must be powered from separate power sources with separate line cords.

1.8.3 Rear heat door exchanger

The Power 795 systems support the rear door heat exchanger (FC 6877) similar to the one used in the p5 590, p5 595, and Power 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 was 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 benefit 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.

1.8.4 Useful rack additions

This section highlights several available rack additions for the Power 795.

Media drawers

Tape and DVD support is provided though the use of a media drawer. As listed in Table 1-18, three media drawers are available for the Power 795 server. Only one internal FC 5720 or FC 5724 media drawer is supported per Power 795 system.

Table 1-18 Media drawers

Feature code or type model	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
7214-1U2[a]	Tape and DVD Enclosure Express	1	✓	✓	✓
7216-1U2	Multimedia Enclosure	1	✓	✓	✓
5720[b]	DVD/Tape SAS External Storage Unit	1	✓	✓	✓
5724	DVD/Tape SAS External Storage Unit with SATA Media	1	✓	✓	✓

a. Supported, but no longer orderable
b. Replaced in new Power 795 proposals by FC 5724

A DVD media device is preferable for use in OS installation, maintenance, problem determination, and service actions such as maintaining system firmware and I/O microcode. AIX, Linux, and IBM i (starting with version 6.1) support network installation of an operating system, but having access to an optical device for systems diagnostics and software installation is useful.

DVD/Tape SAS External Storage Unit

A 24-inch DVD/Tape SAS External Storage Unit (FC 5720) is available to mount in the primary system rack. It is 1U high and is installed at the 12U or 34U location of the system rack. If the Media Drawer Placement Indicator - U12 (FC 8449) is ordered, the media drawer will be installed in the 12U location, below the processor books. Otherwise, it is positioned in the 34U location. When the drawer is installed in the 12U location, one I/O drawer in the primary system rack is eliminated.

Figure 1-3 shows the DVD/Tape SAS External Storage Unit.

Figure 1-3 DVD/Tape SAS External Storage Unit

The available media devices supported on the FC 5720 media drawer are listed in Table 1-19.

Table 1-19 Available tape and DVD media devices on the FC 5720 media drawer

Feature code	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
5619	80/160 GB DAT160 SAS Tape Drive	1	✓	✓	✓
5746	Half High 800 GB / 1.6 TB LTO4 SAS Tape Drive	1	✓	✓	✓
5756[a]	Slimline DVD-ROM Drive	2	✓	✓	✓
5757[a]	IBM 4.7 GB Slimline DVD-RAM Drive	2	✓	✓	✓

a. Supported, but no longer orderable

The media drawer provides the system with two bays for media devices. One SAS controller drives both bays. Only one of the two bays can be populated by a tape drive, which defines a maximum of one tape drive per media drawer. Only one of the two bays can be populated by DVD media. The DVD bay can contain either one or two DVD drives. When the DVD bay is populated by two DVD drives, one tape drive will be available in the other media bay. Both the tape and DVD drives are orderable as feature codes.

DVD/Tape SAS External Storage Unit with SATA Media

The DVD/Tape SAS External Storage Unit (FC 5724) is a 1U-I/O drawer that is used in the Power 795 CEC rack. It holds HH DAT160 tape drives, HH LTO5 tape drives, or DVD drives. In new Power 795 proposals, it replaces the use of the existing 1U I/ O drawer (FC 5720).

The FC 5724 media drawer occupies 1U of rack space in either the 12U or 34U positions in the CEC rack. FC 8449 specifies that it is located in the 12U location, below the processor books. Otherwise, it will be positioned in the 34U location. When positioned in the 12U location, one I/O drawer in the CEC rack is eliminated.

The media drawer provides the system with two bays for media devices. One SAS controller in either a FC 5803, FC 5797, or FC 5798 I/O drawer drives both bays. Only one of the two bays may be populated by a tape drive, which defines a maximum of one tape drive per media drawer. Only one of the two bays may be populated by SATA DVD media, which is driven by a SAS to SATA converter. The DVD bay may contain either one or two DVD drives.

Both the tape (either the DAT160 FC 5619 or the LT0-5 FC 5638) and DVD (FC 5771) drives are orderable with FC 5724.

Figure 1-4 shows the DVD/Tape SAS External Storage Unit with SATA Media.

Figure 1-4 DVD/Tape SAS External Storage Unit with SATA Media

The available media devices that are supported on the FC 5724 media drawer are listed in Table 1-20.

Table 1-20 Available tape and DVD media devices on the FC 5724 media drawer

Feature code	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
5619	80/160 GB DAT160 SAS Tape Drive	1	✓	✓	✓
5638	1.5TB/3.0TB LTO-5 SAS Tape Drive	1	✓	✓	✓
5762[a]	IBM 4.7 GB Slimline DVD-RAM drive	2	✓	✓	✓
5771	IBM 4.7 GB Slimline SATA DVD-RAM Drive	2	✓	✓	✓

a. Supported, but no longer orderable

IBM System Storage 7214 Tape and DVD Enclosure Express

The Tape and DVD Enclosure Express (7214-1U2) is a 19-inch rack-mounted media drawer with two media bays, as shown in Figure 1-5.

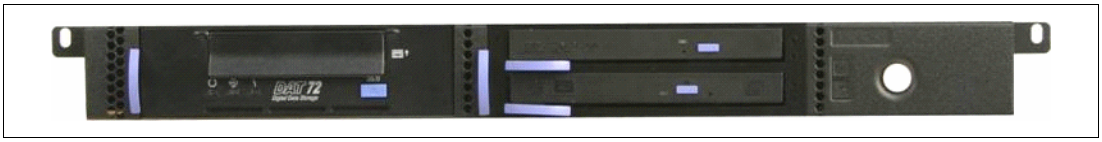

Figure 1-5 19-inch Media Drawer, 19-inch (7214-1U2)

The available media devices supported on the FC 5720 media drawer are listed in Table 1-21.

Table 1-21 Available tape and DVD media devices on the 7214-1U2 drawer

Feature code	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
1401	DAT160 80 GB Tape Drive	2	✓	✓	✓
1402	DAT320 160 GB Tape Drive	2	✓	✓	✓
1404	Half High LTO Ultrium 4 800 GB Tape Drive	2	✓	✓	✓
1420	DVD Sled with 1DVD-RAM Drive	2	✓	✓	✓
1421	DVD Sled with 1DVD-ROM Drive	2	✓	✓	✓
1422	DVD-RAM Slim Optical Drive	1	✓	✓	✓
1423	DVD-ROM Slim Optical Drive	1	✓	✓	✓

The first bay in the 7214-1U2 media drawer supports a tape drive. The second bay can support either a tape device, or two slimline DVD devices. Media devices are connected by a single SAS controller that drives all of the devices. The 7214-1U2 media drawer must be mounted in a 19-inch rack with 1U of available space.

IBM System Storage 7216 Multi-Media Enclosure

The IBM System Storage® 7216 Multi-Media Enclosure (Model 1U2) is designed to attach to the Power 795 or through a PCIe SAS adapter and a PCIe USB adapters, depending on the devices installed on the enclosure. The 7216 has two bays to accommodate external tape, removable disk drive, or DVD-RAM drive options. See Table 1-22 for these features.

Table 1-22 Available tape and DVD media devices on the 7216-1U2 drawer

Feature code	Description	Maximum allowed	Support		
			AIX	IBM i	Linux
1103	RDX Removable Disk Drive Docking Station	2	✓	—	✓
1402	DAT320 160 GB SAS Tape Drive	2	✓	✓	✓
1420	DVD-RAM Sled with one DVD-RAM SAS Drive	2	✓	✓	✓
1422	DVD-RAM SAS Drive Only	2	✓	✓	✓
5619	DAT160 80 GB SAS Tape Drive	2	✓	✓	✓
5673	DAT320 160 GB USB Tape Drive	2	✓	✓	✓
8247	LTO Ultrium 5 Half-Height 1.5 TB SAS Tape Drive	2	✓	—	✓

To attach a 7216 Multi-Media Enclosure to the Power 795, consider the following cabling procedures:

- Attachment by an SAS adapter

  A PCIe Dual-X4 SAS adapter (FC 5901) must be installed in the Power 795 server to attach to a 7216 Model 1U2 Multi-Media Storage Enclosure.

  For each SAS tape drive and DVD-RAM drive feature that is installed in the 7216, the appropriate external SAS cable is included.

  An optional Quad External SAS cable is available by specifying (FC 5544) with each FC 7216 order. The Quad External Cable allows up to four FC 7216 SAS tape or DVD-RAM features to attach to a single System SAS adapter.

  Up to two 7216 storage enclosure SAS features can be attached per PCIe Dual-X4 SAS adapter (FC 5901).

- Attachment by a USB adapter

  The DAT320 USB Tape Drive and Removable RDX HDD Docking Station features on 7216 only support the USB cable that is provided as part of the feature code. Additional USB hubs, add-on USB cables, or USB cable extenders are not supported.

  For each DAT320 USB tape drive and RDX Docking Station feature installed in the 7216, the appropriate external USB cable is included. The 7216 DAT320 USB tape drive or RDX Docking Station features can be connected to a 4-Port USB PCIe adapter (FC 2728).

The two drive slots of the 7216 enclosure can hold the following drive combinations:

- One tape drive (DAT160 SAS or LTO Ultrium 5 Half-Height SAS) with second bay empty
- Two tape drives (DAT160 SAS or LTO Ultrium 5 Half-Height SAS) in any combination
- One tape drive (DAT160 SAS or LTO Ultrium 5 Half-Height SAS) and one DVD-RAM SAS drive sled with one or two DVD-RAM SAS drives
- Up to four DVD-RAM drives
- One tape drive (DAT160 SAS or LTO Ultrium 5 Half-Height SAS) in one bay, and one RDX Removable HDD Docking Station in the other drive bay
- One RDX Removable HDD Docking Station and one DVD-RAM SAS drive sled with one or two DVD-RAM SAS drives in the right-side bay
- Two RDX Removable HDD Docking Stations

Figure 1-6 shows the 7216 Multi-Media Enclosure.

Figure 1-6 7216 Multi-Media Enclosure

In general, the 7216-1U2 is supported by the AIX, IBM i, and Linux operating systems.

2

Architecture and technical overview

This chapter discusses the overall system architecture, represented by Figure 2-1 on page 38, with its major components described in the following sections. The bandwidths that are provided throughout the section are theoretical maximums that are used for reference.

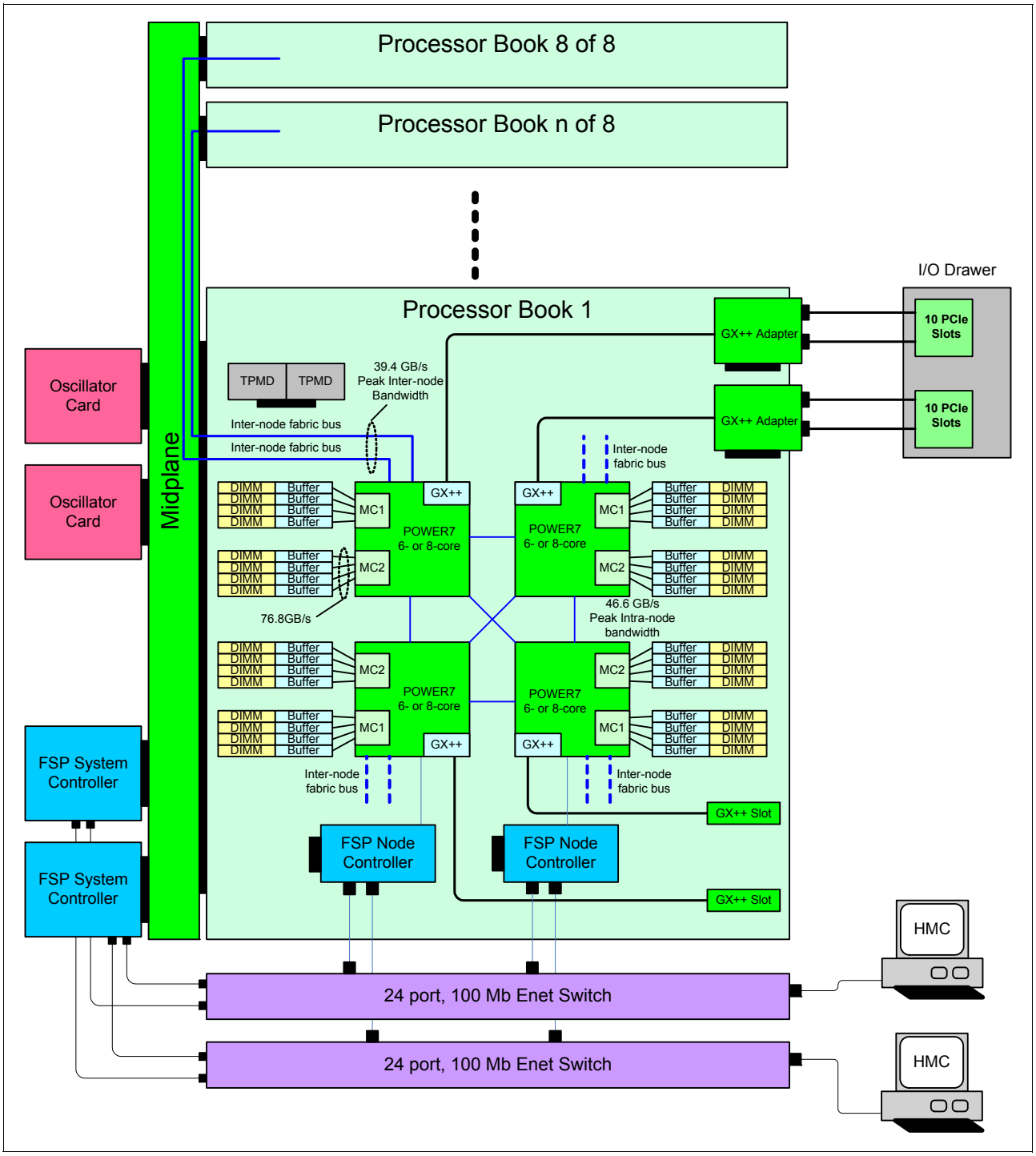

Figure 2-1 Power 795 logic data flow

The speeds shown are at an individual component level. Multiple components and application implementation are key to achieving the best performance.

Always do performance sizing at the application-workload environment level and evaluate performance by using real-world performance measurements using production workloads.

2.1 System design

The IBM Power 795 enterprise-class structure is the result of the continuous evolution of Power 795 and the enterprise-class machines before it. 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 Power 795 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 IBM WebSphere® software, Java, and Linux for integration and deployment in heterogeneous business solutions. What makes the IBM high-end server truly unique is the ability to switch between its standard throughput mode and its unique, optimized TurboCore mode, where performance per core is boosted with access to both additional L3 cache and additional clock speed.

The Power 795 enterprise server is designed to deliver outstanding price for performance, mainframe-inspired reliability and availability features, flexible capacity upgrades, and innovative virtualization technologies.

The Power 795 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. The advantages of this configuration are discussed in the following sections.

2.1.1 Design highlights

The IBM Power 795 (9119-FHB) is a high-end POWER7 processor-based symmetric multiprocessing (SMP) system. As mentioned, the Power 795 has 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 offers advantages in optimization of floor space usage.

A primary feature of the Power 795 system is that it allows simple processor book upgrades of the Power 595 system. Note that new build Power 795 systems will not be identical to upgraded systems. New build Power 795 systems use new power components, such as bulk power enclosures (BPEs) bulk power regulators (BPRs), and bulk power distributors (BPDs).

Conceptually, the Power 795 is similar to the previous server generations and can be configured in one rack (primary) or multiple racks (primary plus expansions).

The primary Power 795 frame is a 42U, 24-inch primary rack, containing the following major subsystems, as shown in Figure 2-2 on page 40, from top to bottom:

- A 42U-tall, 24-inch system rack (primary) houses the major subsystems.
- A redundant power subsystem, housed in the bulk power assemblies (BPAs), located in the 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 POWER7 technology-based processor books are mounted in the CEC. Each processor book contains four 6-core or 8-core single-chip modules (SCMs). The 6-core SCMs operate at 3.72 GHz. The 8-core SCMs operate at either 4.00 GHz or 4.25 GHz. Each 8-core SCM has the ability to switch between standard 4.00 GHz throughput mode and 4.25 GHz TurboCore mode. Each processor book incorporates 32 dual in-line memory module (DIMM) slots.

- The Integrated battery feature (IBF) for backup is optional.
- The media drawer is optional.
- There are up to 32 12X I/O drawers, each with 20 PCIe adapters and 26 SAS, SFF DASD bays with 26 HDD or 18 SSD DASD drives.

In addition, depending on the configuration, it can have the following features:

- Additional 19-inch expansion rack including DASD/SSD, tape, and DVD,
- 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 Power 795 I/O expansion rules.
- One to two nonpowered expansion racks, each supporting up to seven I/O drawers, 4U high.

> **Nonpowered rack:** 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 shows the primary rack major subsystems.

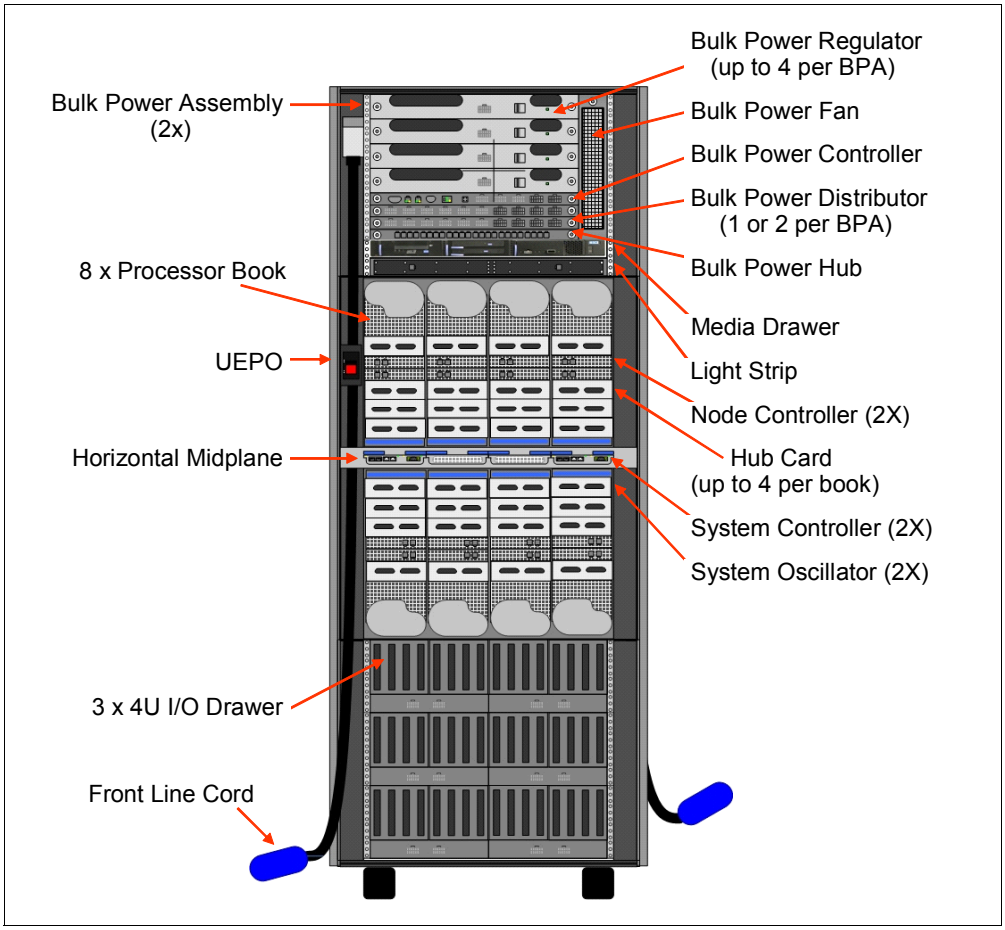

Figure 2-2 Power 795 front view

Figure 2-3 shows the four air-moving devices at the rear of the Power 795.

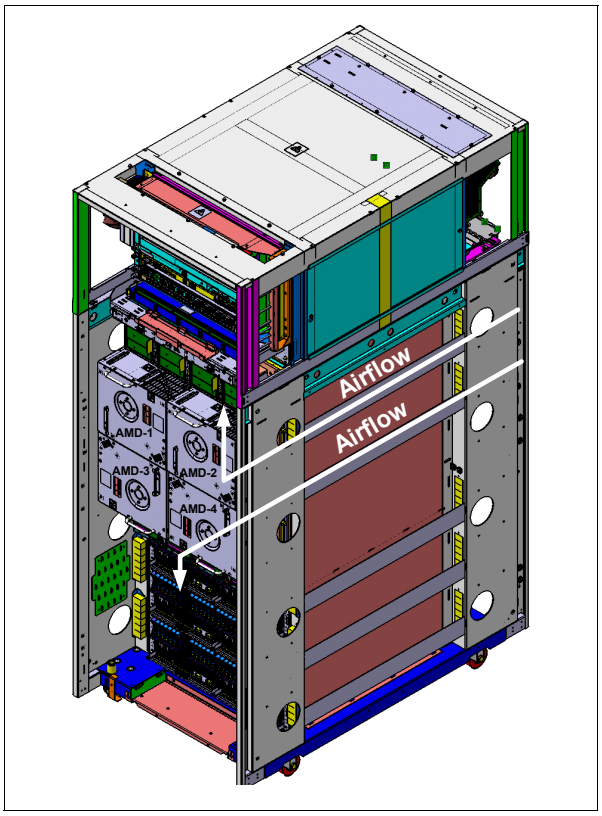

Figure 2-3 Power 795 rear view

2.1.2 Central electronic complex (CEC)

The Power 795 central electronic complex (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 are referred to as *processor books*. The processor books are located in the CEC, which is mounted in the primary rack.

Processor book components

Each processor book assembly contains many components:

- The processor book planar provides support for four single-chip modules (SCM), 32 memory DIMM slots, a connector to the mid-plane, four I/O adapter connectors, two node controller connectors, connector for the thermal power management device (TPMD), and one VPD card connector.
- Each processor book incorporates 32 memory dual in-line memory module (DIMM) slots. Each processor book contains four 6-core or 8-core SCMs. Each SCM is supported by 2048 KB of L2 cache and 32 MB of L3 cache. The 6-core SCMs operate at 3.72 GHz. The 8-core SCMs operate at either 4.00 GHz or 4.25 GHz. Each 8-core SCM has the ability to switch between standard, 4.00 GHz throughput mode and 4.25 GHz TurboCore mode. In TurboCore mode, each 8-core SCM operates with up to four active cores and 32 MB of L3 cache, twice the L3 cache per core available in standard mode.

- The minimum number of processor books supported in TurboCore mode is three, 96-cores. TurboCore mode is specified by FC 9982.
- Switching in and out of TurboCore requires a system reboot. The entire system must operate in either standard or TurboCore mode.
- 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. Route sense and control signals pass between the DCAs and processor book planar (DIMM LED Control). The VPS card plugs into the processor book planar and the node power distribution planar.
- Two Distributed Converter Assemblies (DCAs) are located at the back of each processor book.
- Four 12X I/O hub adapter slots are located at the front of the processor book.
- Two embedded node controller service processor cards (NCs) are located in the front of the processor book. The node controller cards communicate to the HMC through the bulk power hub (BPH) and are connected to both front and rear Ethernet ports on the BPH.
- Two TPMD cards are included. The TPMD function is composed of a risk processor and data acquisition capability, which allows the TPMD to monitor power usage and temperatures in real time and to adjust the processor node power and performance in real time. There are redundant TPMD implementations in each processor node.

Figure 2-4 shows the layout of a processor book and its components.

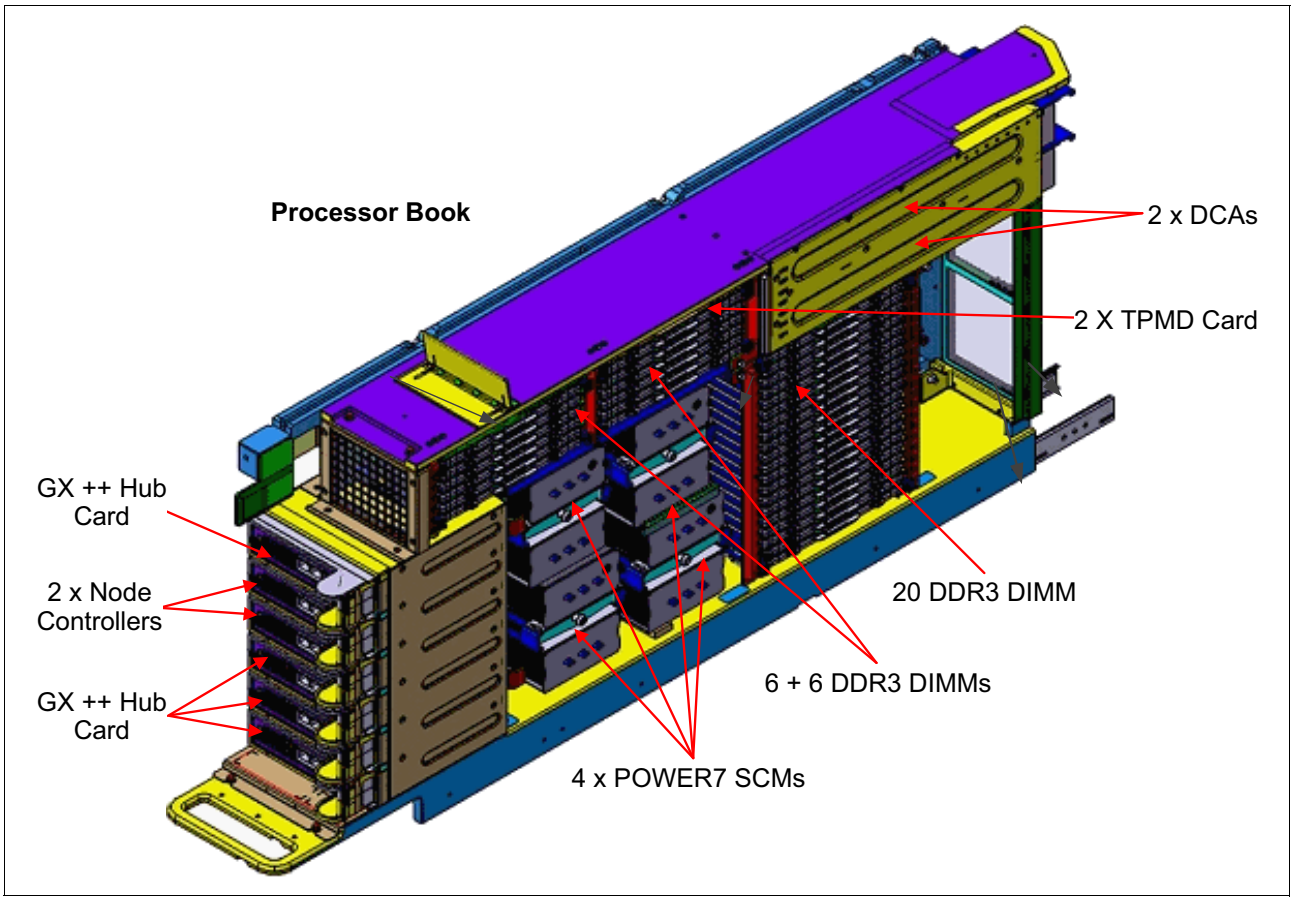

Figure 2-4 Processor book layout

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.

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 fails, the backup card detects the failure and continues to provide the clock signal so that no outage occurs because of 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.

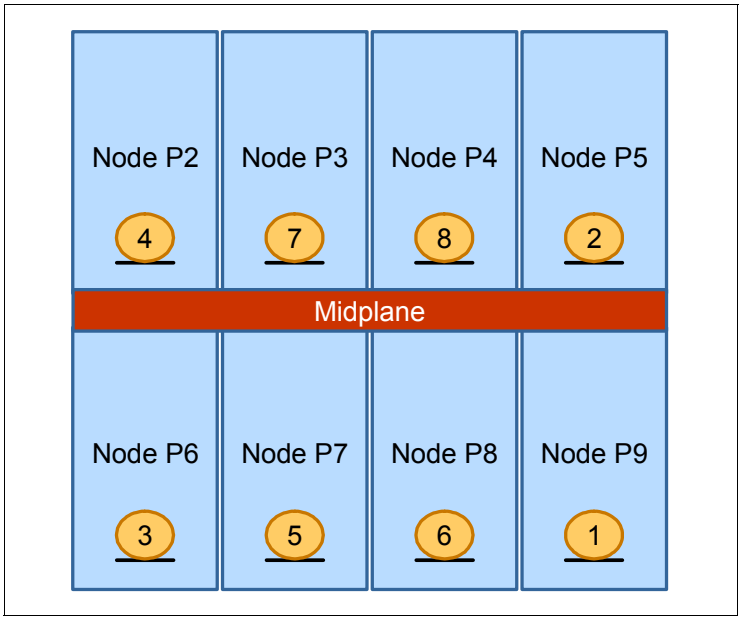

Figure 2-5 Processor book installation sequence

2.1.3 CEC midplane

The POWER7 795 CEC midplane holds the processor books in an ingenious way: four books 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 the following items:

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

Figure 2-6 shows the CEC midplane layout.

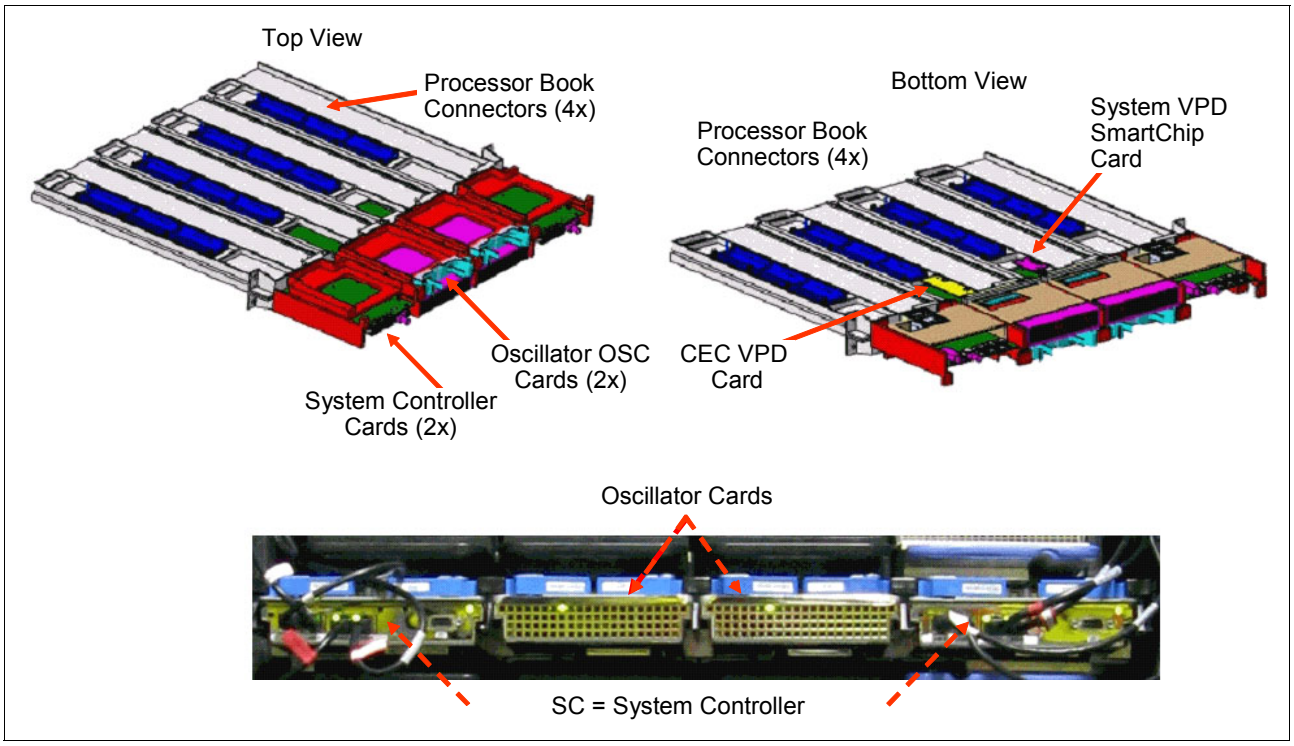

Figure 2-6 CEC midplane

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	CEC System controller (SC) card 1

2.1.4 Service and control system

The Power 795 can 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, consider what happens if a control task, in the course of recovering from an I/O error, decides 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 feature, which provides detailed problem analysis, pointing to the 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 (SCs), nodes, redundant node controllers (NCs), are in the BPC.

Figure 2-7 shows a high-level view of a Power 795, together with its associated control structure. The system depicted to the right is composed of the CEC with many processor books and I/O drawers.

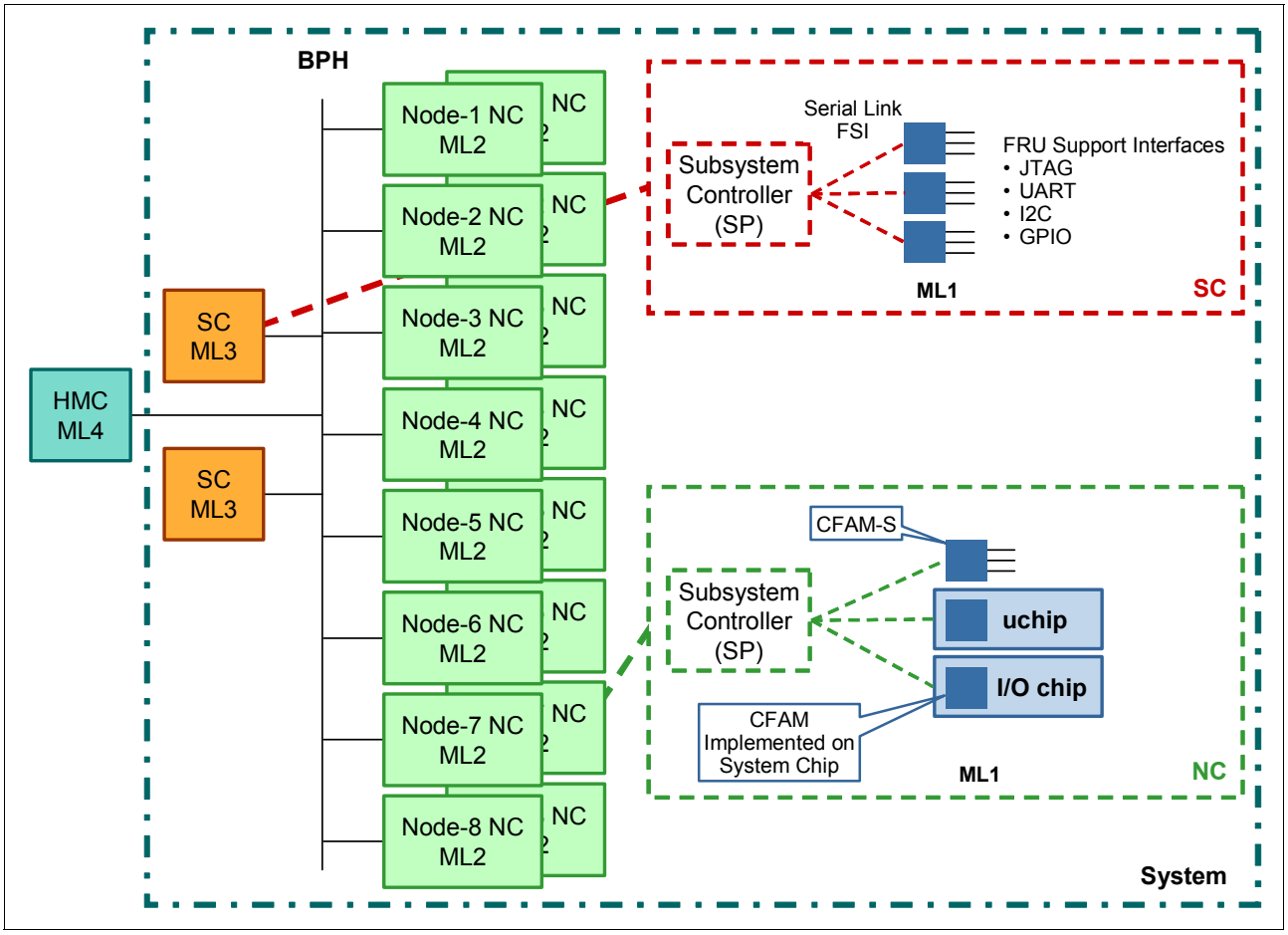

Figure 2-7 System Control Structure (SCS) architecture

The scope of the system controllers 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. This approach requires 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

Figure 2-7 on page 46 also shows that SCS is divided into management domains or management levels:

- Management Level 1 domain (ML1): This domain 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 domain is management of a single subsystem (for example, a processor book) or a 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 requires 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 superset 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): This set of functions can manage multiple systems and can be located apart from the system to be controlled (HMC level).

The Power Systems 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 scope of the system controllers 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 by 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 (NCs). 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 46, 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 46, 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 Macros (CFAMs). 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 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 four 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 use this functionality.

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

2.1.5 System controller card

Two service processor cards are on the CEC midplane. These cards are referred to as system controller (SC) cards, as shown in Figure 2-8.

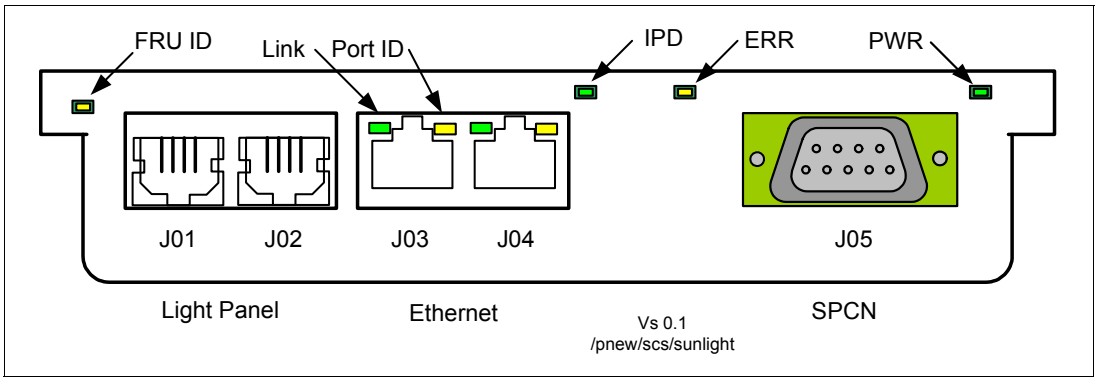

Figure 2-8 System controller card

The SC card provides connectors for the following ports:

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

System Power Control Network (SPCN)

The 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 interfaces and use standard RS485 9-pin female connector (DB9).

Network diagram of SPCN

Figure 2-9 details the SPCN control network.

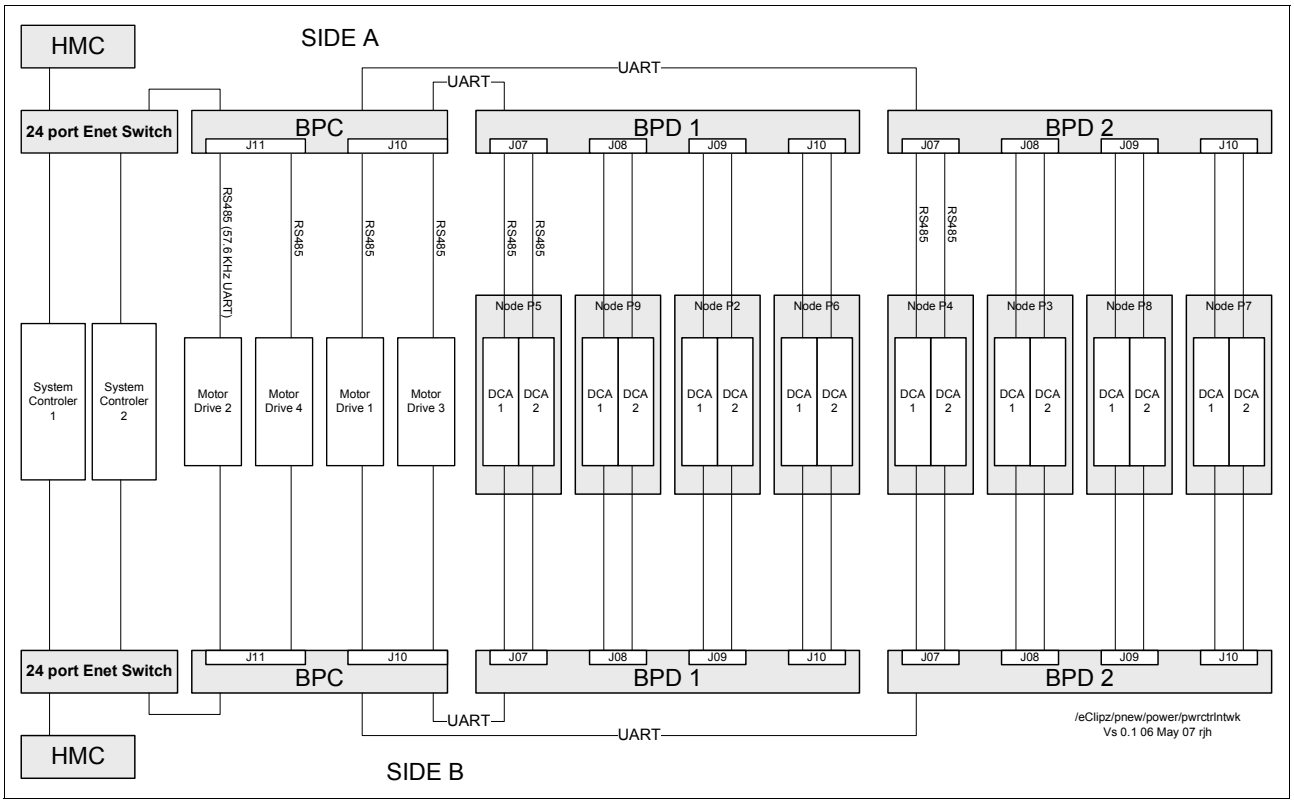

Figure 2-9 SPCN control network

2.1.6 System VPD cards

Two types of vital product data (VPD) cards are available: VPD and smartchip VPD (SVPD). VPDs 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 44) and on a VPD daughter card, part of the 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 the 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 44) and on the 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 is 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 only at processor book level.

SCM 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 SCM serial number is stored in the SEEPROMs on the SCM.

Figure 2-6 on page 44 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.

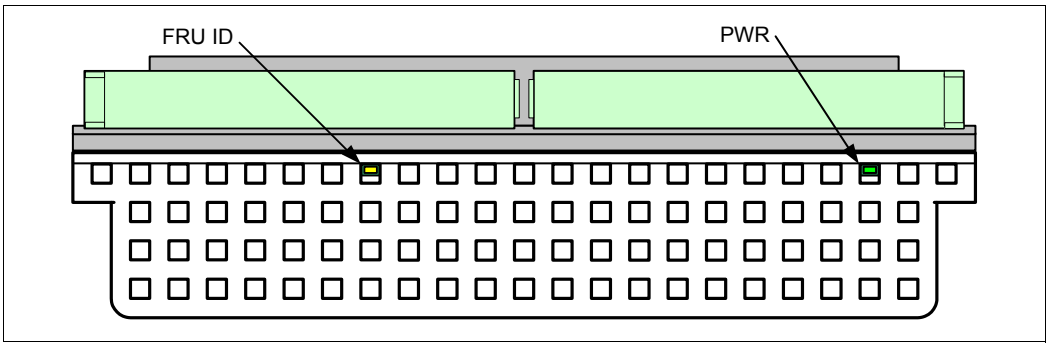

Figure 2-10 Oscillator card

2.1.8 Node controller card

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

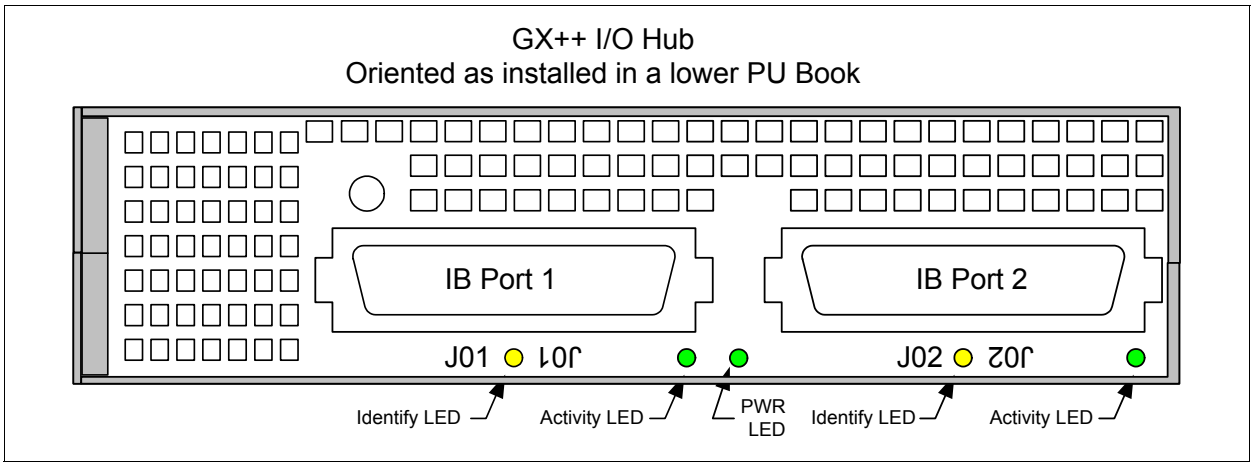

Figure 2-11 Node controller card

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 the 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

For the CEC, direct current (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 node planar is wired to support a unique core voltage/nest domain and a cache array voltage domain for each of the our SCMs. A common I/O voltage domain is shared among all CEC logic except the GX hub chip, which uses a separate 1.2V domain and the memory channel I/O which uses 1.0 V. A common memory domain powers all the memory buffers within a processor book. A common DRAM memory domain powers all the DRAMs within a PU book.

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

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

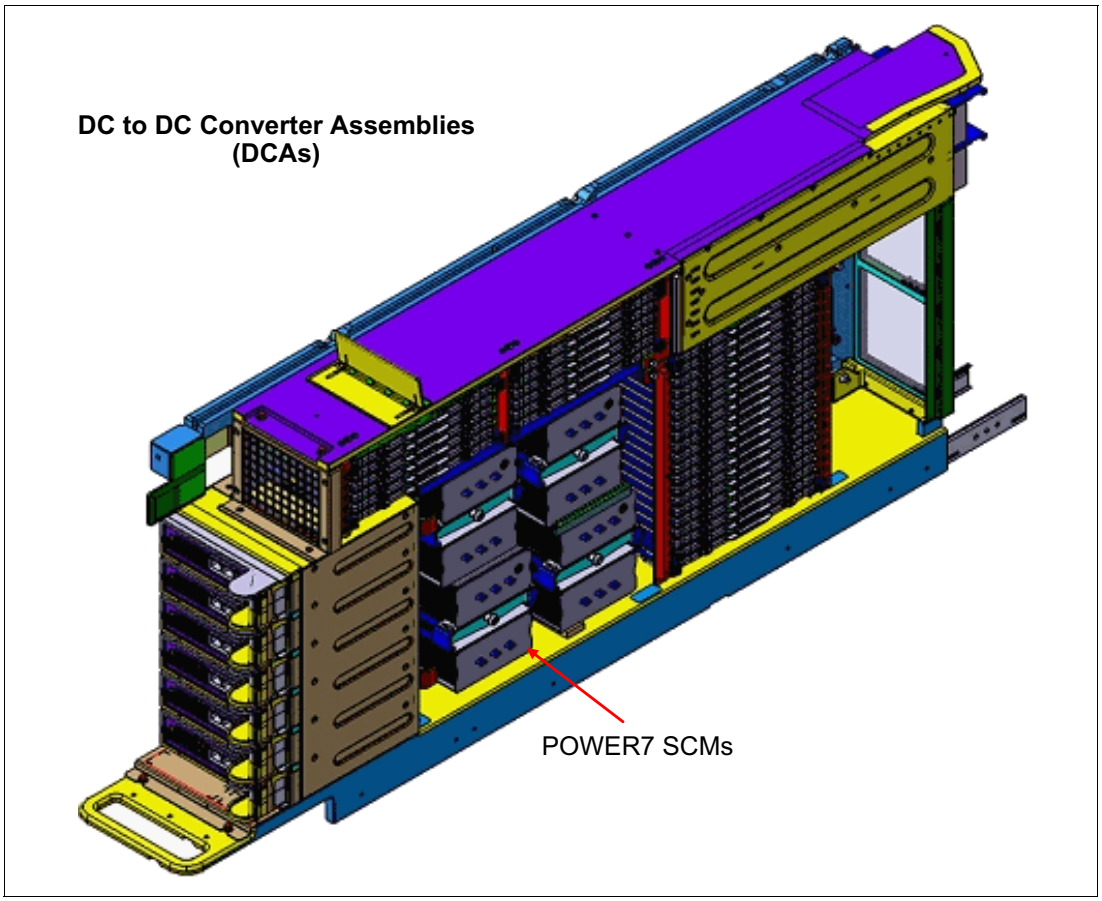

Figure 2-12 DC converter assembly (DCA)

When both DCAs are operational, adequate power is available for all operating conditions. For some configurations and workloads, the processor book might draw more current than a single DCA can supply. In the event of a DCA failure under heavy load or high ambient temperature, the remaining operational DCA can supply the necessary current for only a limited period before overheating. To prevent overheating when load is excessive, the remaining DCA can reduce the book load by communicating to the TPMD card over an link.

The TPMD card then places the book in “safe mode.” This mode is expected to set processor frequency to approximately 10% below nominal frequency. It can also include memory throttling. Partitions in power save mode must be left in power save mode. After a failing DCA is replaced, the system is returned to normal operation automatically.

2.2 System buses

The POWER7 processor interfaces can be divided into three categories:

- SMP interconnect: These interfaces connect the POWER7 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 POWER7 technology-based chip.
- External interconnect: Interfaces provide for communication with I/O devices outside the central system.

In the following section, we describe the SMP and external interconnects.

2.2.1 System interconnects

The Power 795 uses point-to-point SMP fabric interfaces between processor node books. Each processor book holds a processor node consisting of four POWER7 processors designated S, T, U, and V.

The bus topology is no longer ring-based as in POWER5, but rather a multitier, fully-connected topology to reduce latency, increase redundancy, and improve concurrent maintenance. Reliability is improved with error correction 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 book's read cache to address the requested data or control information. Inter-book communication takes place at the Level 2 (L2) cache.

The POWER7 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 POWER7 processor. The POWER7 processor has five fabric ports that can be used to connect to other POWER7 processors:

- Three for intranode bus interconnections: They are designated as X, Y, and Z, and are used to fully connect the POWER7 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 multinode systems.

Physically, the fabric bus is an 8-byte, 4-byte, or 2-byte wide, split-transaction, multiplexed address. For Power 795, the bus is 8 bytes and operates at 2.5 Ghz.

From a fabric perspective, a node (processor book node) consists of one to four processors fully connected with XYZ busses. AB busses are used to connect various fabric nodes together. The one to four processors on a node work together to broadcast address requests to other nodes in the system. Each node can have up to eight AB links (two for a processor, four processors per node). Figure 2-13 on page 55 illustrates the bus interconnections.

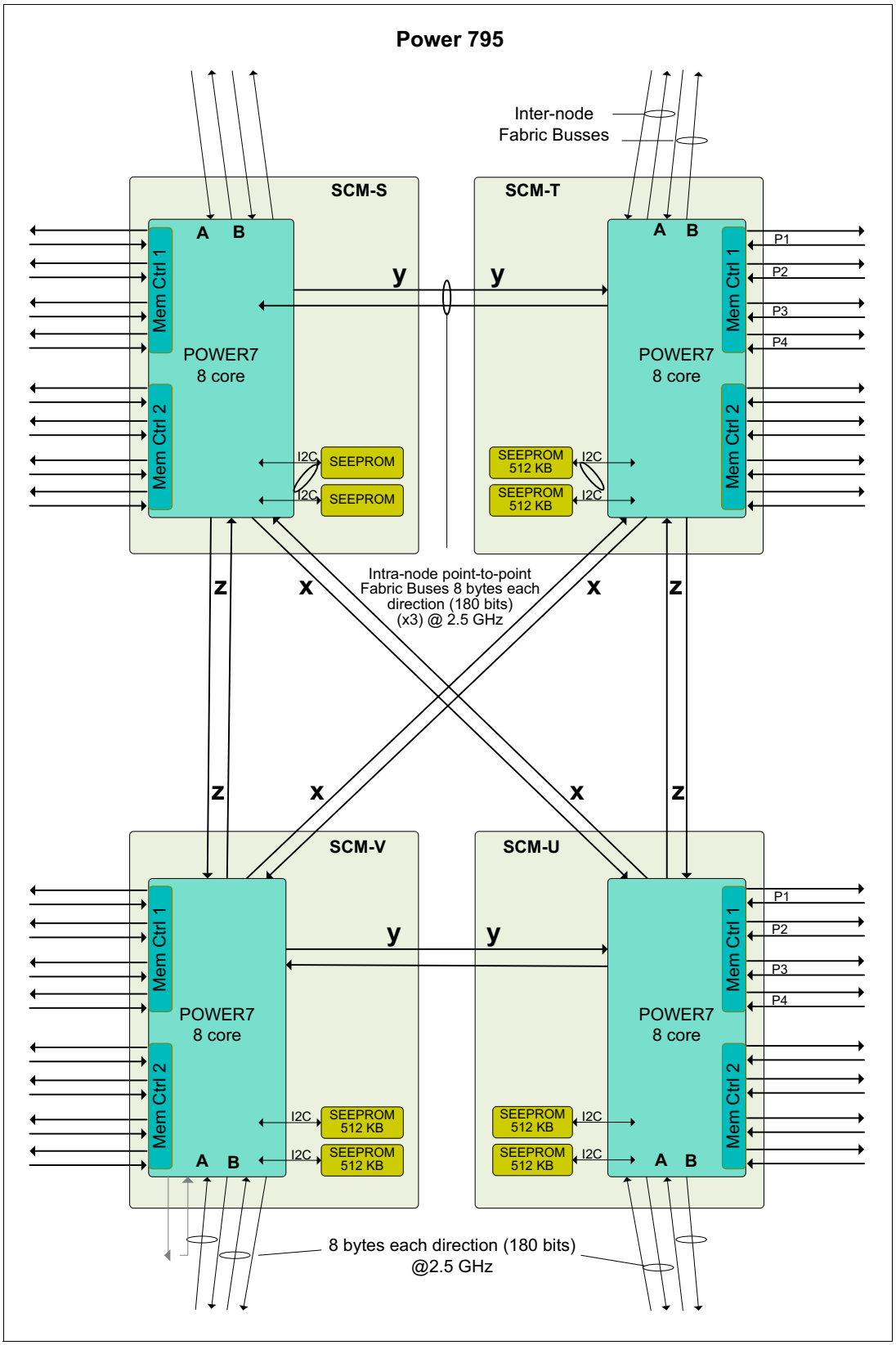

Figure 2-13 Fabric bus interconnections

With both the POWER5 and the POWER7 processor approaches, large systems are constructed by aggregating multiple nodes.

Figure 2-14 illustrates the potential for a large, robust, 256 core system that uses 8-byte SMP interconnect links, both L3 data ports to maximize L3 bandwidth, and all eight memory channels per chip.

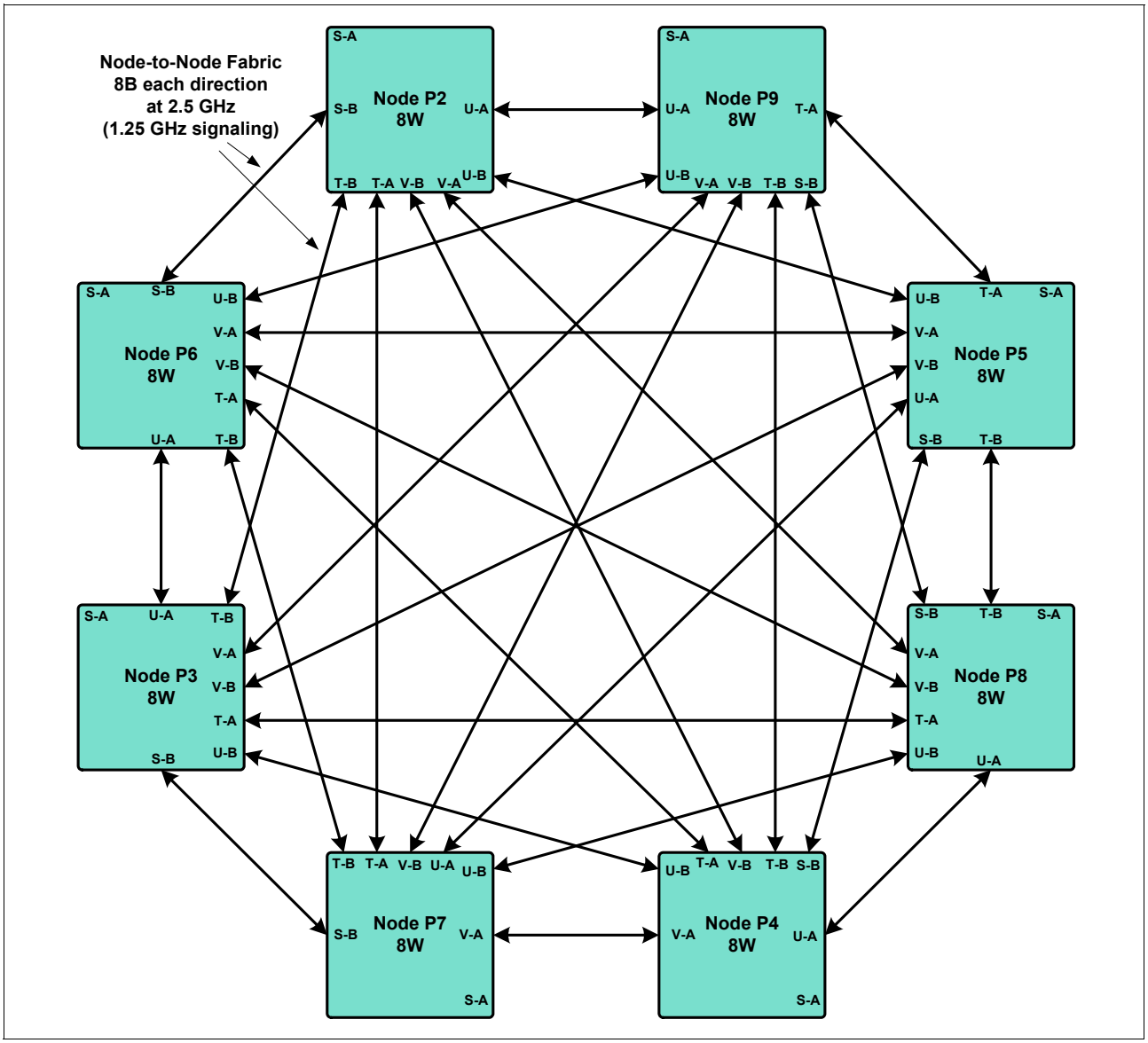

Figure 2-14 Power 795 256 core

2.2.2 I/O subsystem

The Power 795 uses remote I/O drawers for directly attached PCI or PCI-X adapters and disk capabilities. The Power 795 supports I/O DASD and media drawers through 12X host channel adapters (HCA), located in the front of the processor books. These are collectively referred to as *GX adapters*.

One type of GX adapter cards is supported in the Power 795 servers:

- GX Dual-port 12X HCA adapter (FC 1816)

Upgrade information:

There is no support for RIO-2 drawers on Power 795. Migration of Power 595 12X GX adapters to the Power 795 is supported. Migration of the remote I/O (RIO) GX adapter (FC 1814) is not supported.

Customers that plan to migrate adapters installed in RIO-based drawers or towers must ensure that there are spare slots available in 12X PCIe I/O Drawers (FC 5803) for these adapters, or must order such drawers to accommodate the adapters.

Movement of adapters must happen prior to the actual model upgrade.

Customers must avoid adding or removing I/O drawers during the upgrade because this can result in changes to I/O drawer bus IDs, which can cause problems with partition I/O assignments while partitions are re-created on the upgrade machine.

There is no I/O processor (IOP) support in Power 795. IOP-based adapters must be replaced before the upgrade.

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 might experience degraded performance and serviceability.

The 12X loop connections connect to the system CEC with 12X loop attachment adapters (FC 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.

Figure 2-15 shows the GX adapter layout for a two-processor book.

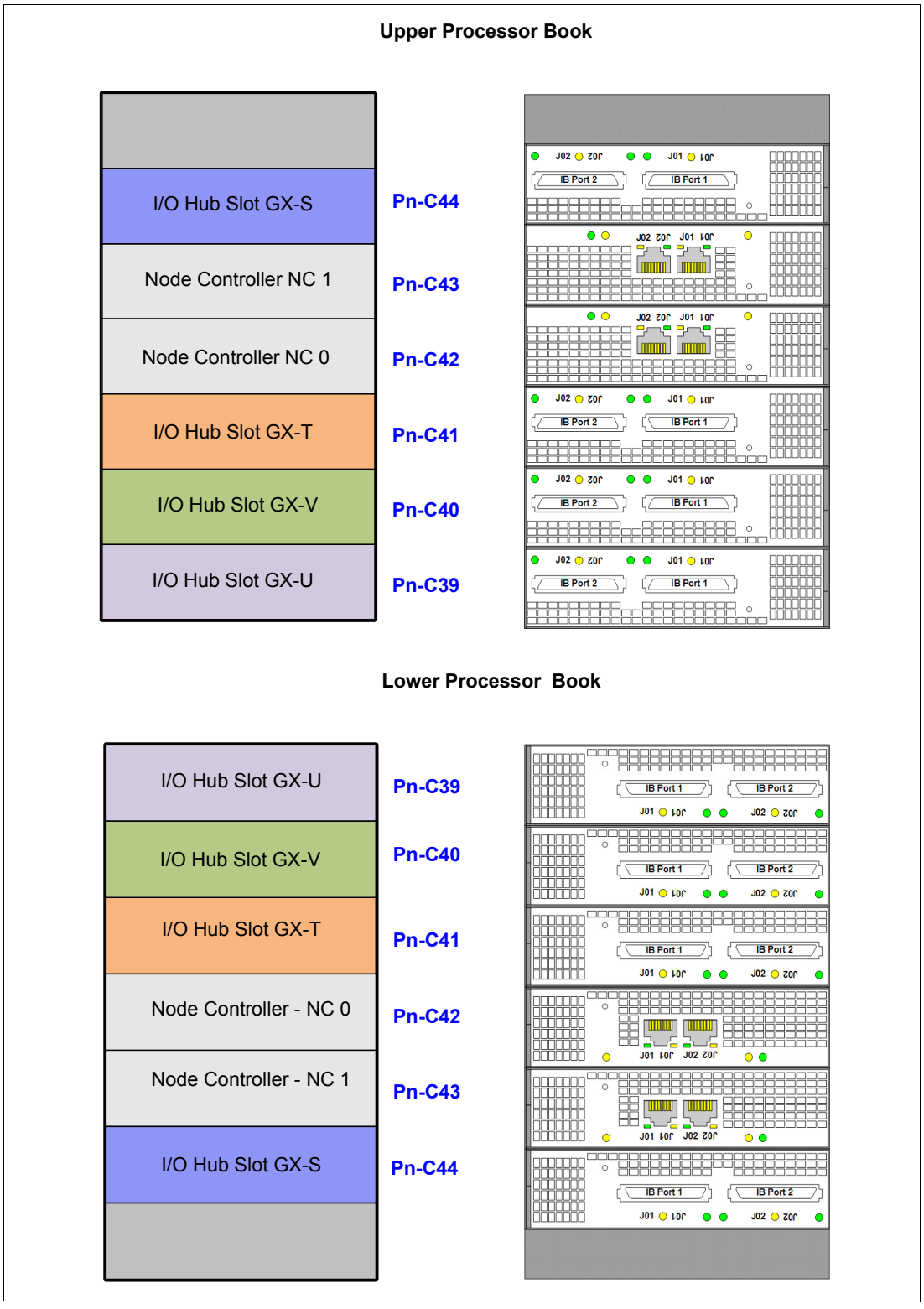

Figure 2-15 GX adapter

Rules for I/O hub plugging are in 2.8, “Internal I/O subsystem” on page 88.

Table 2-3 shows placement of the GX bus I/O hub 2-port adapter (FC 1816), which can be a maximum of four per node, depending on the processor books that are installed on a Power 795.

Table 2-3 I/O Hub adapter placement

Node 1 (P9)	Node 2 (P5)	Node 3 (P6)	Node 4 (P2)	Node 5 (P7)	Node 6 (P8)	Node 7 (P3)	Node 8 (P4)
1S, 1T, 1V, 1U	1S, 2S, 1T, 2T, 1V, 2V, 1U, 2U	1S, 2S, 3S, 1T, 2T, 3T, 1V, 2V, 3V, 1U, 2U, 3U	1S, 2S, 3S, 4S, 1T, 2T, 3T, 4T, 1V, 2V, 3V, 4V, 1U, 2U, 3U, 4U	1S, 2S, 3S, 4S, 5S, 1T, 2T, 3T, 4T, 5T, 1V, 2V, 3V, 4V, 5V, 1U, 2U, 3U, 4U, 5U	1S, 2S, 3S, 4S, 5S, 6S, 1T, 2T, 3T, 4T, 5T, 6T, 1V, 2V, 3V, 4V, 5V, 6V, 1U, 2U, 3U, 4U, 5U, 6U	1S, 2S, 3S, 4S, 5S, 6S, 7S, 1T, 2T, 3T, 4T, 5T, 6T, 7T, 1V, 2V, 3V, 4V, 5V, 6V, 7V, 1U, 2U, 3U, 4U, 5U, 6U, 7U	1S, 2S, 3S, 4S, 5S, 6S, 7S, 8S, 1T, 2T, 3T, 4T, 5T, 6T, 7T, 8T, 1V, 2V, 3V, 4V, 5V, 6V, 7V, 8V, 1U, 2U, 3U, 4U, 5U, 6U, 7U, 8U

Upgrade information:

- There is no support for RIO-2 drawers on Power 795. Migration of Power 595 12X GX adapters to the Power 795 is supported. Migration of the Remote I/O RIO GX adapter (FC 1814) is not supported.
- Customers that plan to migrate adapters installed in RIO based drawers or towers must ensure that there are spare slots available in 12X PCIe I/O Drawers (FC 5803) for these adapters or must order such drawers to accommodate the adapters.
- Movement of adapters must happen prior to the actual model upgrade.
- Customers must avoid adding or removing I/O drawers during the upgrade because this can result in changes to I/O drawer bus IDs, which can cause problems with partition I/O assignments as partitions are recreated on the upgrade machine.
- There is no IOP support in Power 795. IOP based adapters have to be replaced before the upgrade.

2.3 The IBM POWER7 processor

The IBM POWER7 processor represents a leap forward in technology achievement and associated computing capability. The multicore architecture of the POWER7 processor is matched with innovation across a wide range of related technologies to deliver leading throughput, efficiency, scalability, and RAS.

Although the processor is an important component in delivering outstanding servers, many elements and facilities have to be balanced on a server to deliver maximum throughput. As with previous generations of systems based on POWER processors, the design philosophy for POWER7 processor-based systems is one of system-wide balance in which the POWER7 processor plays an important role.

In many cases, IBM has been innovative to achieve required levels of throughput and bandwidth. Innovation for the POWER7 processor and POWER7 processor-based systems include (but are not limited to) the following areas:

- On-chip L3 cache implemented in embedded dynamic random access memory (eDRAM)
- Cache hierarchy and component innovation
- Advances in memory subsystem
- Advances in off-chip signalling
- Exploitation of long-term investment in coherence innovation

The superscalar POWER7 processor design also provides a variety of other capabilities:

- Binary compatibility with the prior generation of POWER processors
- Support for PowerVM virtualization capabilities, including PowerVM Live Partition Mobility to and from POWER6 and IBM POWER6+™ processor-based systems.

Figure 2-16 shows the POWER7 processor die layout with the major areas identified; processor cores, L2 cache, L3 cache and chip interconnection, symmetric multiprocessing (SMP) links, and memory controllers.

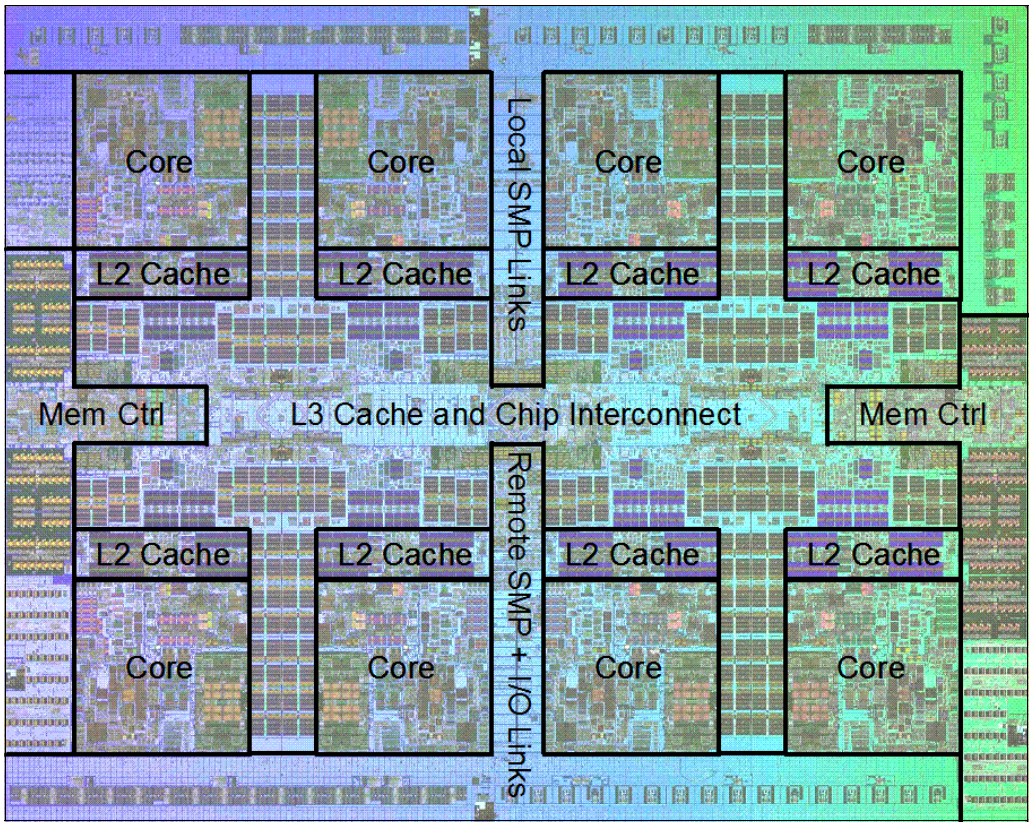

Figure 2-16 POWER7 processor die with key areas indicated

2.3.1 POWER7 processor overview

The POWER7 processor chip is fabricated by using the IBM 45 nm Silicon-On-Insulator (SOI) technology with copper interconnect, and implements an on-chip L3 cache using eDRAM.

The POWER7 processor chip has an area of 567 mm^2 and is built with 1.2 billion components (transistors). Eight processor cores are on the chip, each with 12 execution units, 256 KB of L2 cache, and access to 32 MB of shared on-chip L3 cache.

For memory access, the POWER7 processor includes two Double Data Rate 3 (DDR3) memory controllers, each with four memory channels. To be able to scale effectively, the POWER7 processor uses a combination of local and global SMP links with high coherency bandwidth and takes advantage of the IBM dual-scope broadcast coherence protocol.

Table 2-4 summarizes the technology characteristics of the POWER7 processor.

Table 2-4 Summary of POWER7 processor technology

Technology	Description
Die size	567 mm^2
Fabrication technology	► 45 nm lithography ► Copper interconnect ► Silicon-on-Insulator ► eDRAM
Components	1.2 billion components and transistors offering the equivalent function of 2.7 billion (for more details see 2.3.6, “On-chip L3 cache innovation and Intelligent Cache” on page 65)
Processor cores	8
Max execution threads per core/chip	4/32
L2 cache per core/chip	256 KB/2 MB
On-chip L3 cache per core/chip	4 MB/32 MB
DDR3 memory controllers	2
SMP design-point	32 sockets with IBM POWER7 processors
Compatibility	With prior generation of POWER processor

2.3.2 POWER7 processor core

Each POWER7 processor core implements aggressive out-of-order (OoO) instruction execution to drive high efficiency in the use of available execution paths. The POWER7 processor has an Instruction Sequence Unit that is capable of dispatching up to six instructions per cycle to a set of queues. Up to eight instructions per cycle can be issued to the Instruction Execution units.

The POWER7 processor has a set of twelve execution units:

- 2 fixed point units
- 2 load store units
- 4 double precision floating point units
- 1 vector unit
- 1 branch unit
- 1 condition register unit
- 1 decimal floating point unit

The following caches are tightly coupled to each POWER7 processor core:

- Instruction cache: 32 KB
- Data cache: 32 KB
- L2 cache: 256 KB, implemented in fast SRAM

2.3.3 Simultaneous multithreading

An enhancement in the POWER7 processor is the addition of the SMT4 mode to enable four instruction threads to execute simultaneously in each POWER7 processor core. Thus, the instruction thread execution modes of the POWER7 processor are as follows:

- SMT1: Single instruction execution thread per core
- SMT2: Two instruction execution threads per core
- SMT4: Four instruction execution threads per core

SMT4 mode

SMT4 mode enables the POWER7 processor to maximize the throughput of the processor core by offering an increase in core efficiency. SMT4 mode is the latest step in an evolution of multithreading technologies that IBM introduced. Figure 2-17 shows the evolution of simultaneous multithreading.

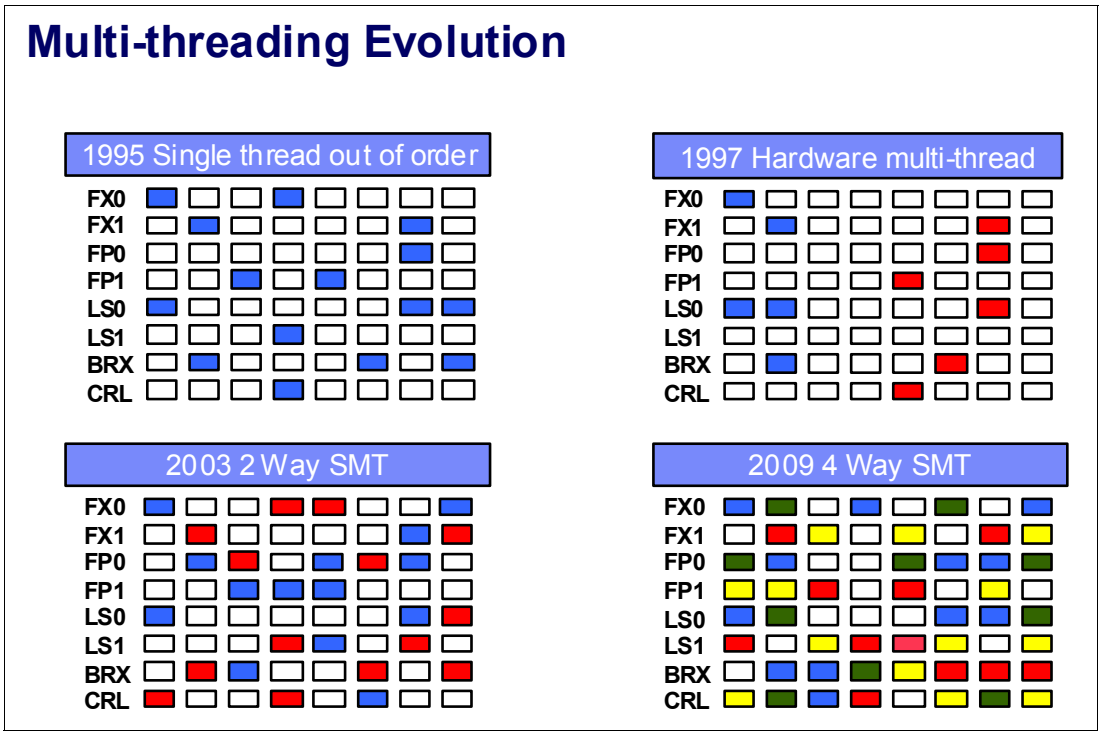

Figure 2-17 Evolution of simultaneous multithreading

The various SMT modes offered by the POWER7 processor allow flexibility, enabling users to select the threading mode that meets an aggregation of objectives such as performance, throughput, energy use, and workload enablement.

Intelligent Threads

The POWER7 processor features *Intelligent Threads* that can vary based on the workload demand. The system either automatically selects (or the system administrator can manually select) whether a workload benefits from dedicating as much capability as possible to a single thread of work, or if the workload benefits more from having capability spread across two or four threads of work. With more threads, the POWER7 processor can deliver more total capacity as more tasks are accomplished in parallel. With fewer threads, those workloads that need fast individual tasks can get the performance they need for maximum benefit.

2.3.4 Memory access

Each POWER7 processor chip has two DDR3 memory controllers, each with four memory channels (enabling eight memory channels per POWER7 processor chip). Each channel operates at 6.4 GHz and can address up to 32 GB of memory. Thus, each POWER7 processor chip is capable of addressing up to 256 GB of memory.

Note: In certain POWER7 processor-based systems, one memory controller is active with four memory channels being used.

Figure 2-18 provides a simple overview of the POWER7 processor memory access structure.

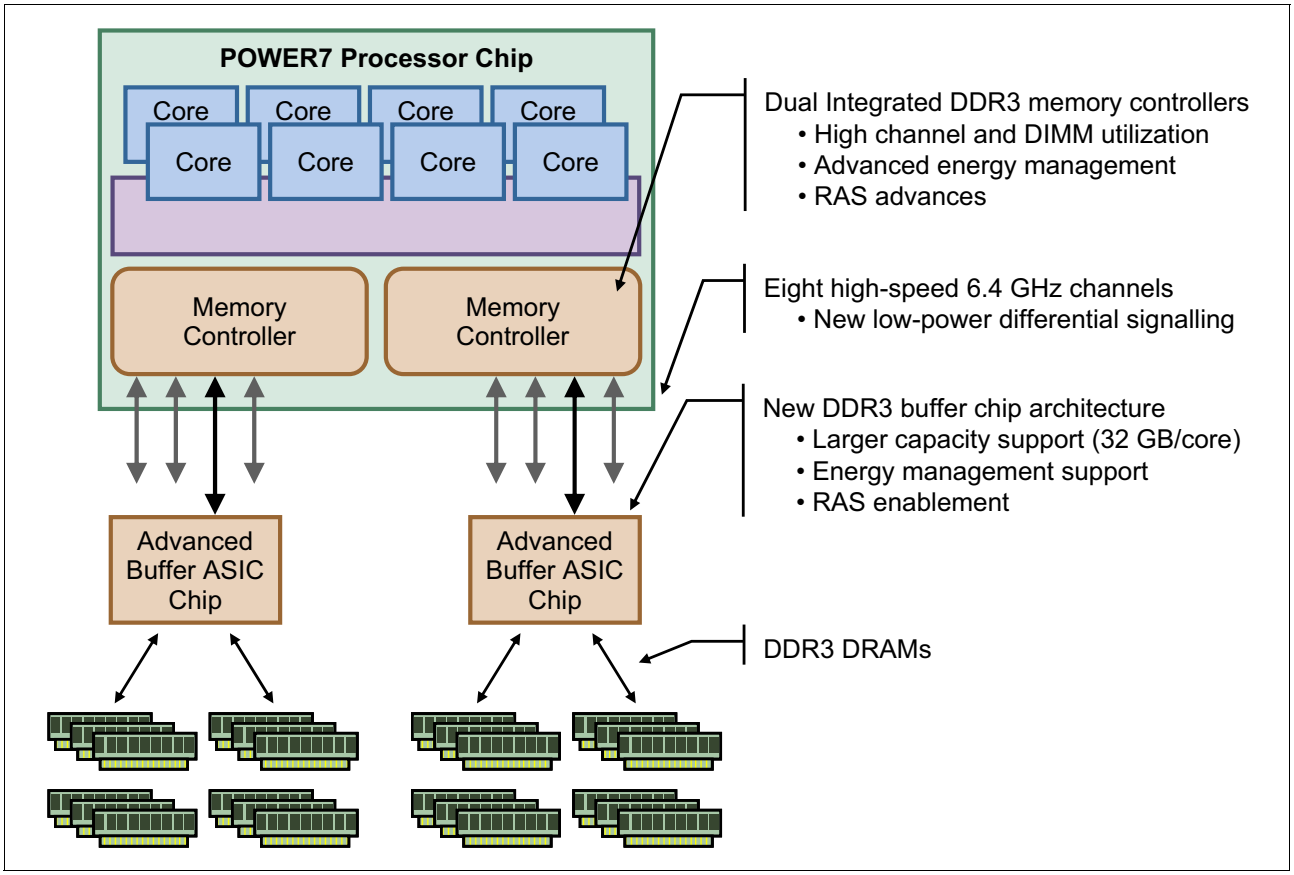

Figure 2-18 Overview of POWER7 memory access structure

2.3.5 Flexible POWER7 processor packaging and offerings

The POWER7 processor forms the basis of a flexible compute platform and can be offered in a number of guises to address differing system requirements.

The POWER7 processor can be offered with a single active memory controller with four channels for servers where higher degrees of memory parallelism are not required.

Similarly, the POWER7 processor can be offered with a variety of SMP bus capacities that are appropriate to the scaling-point of particular server models.

Figure 2-19 shows physical packaging options that are supported with POWER7 processors.

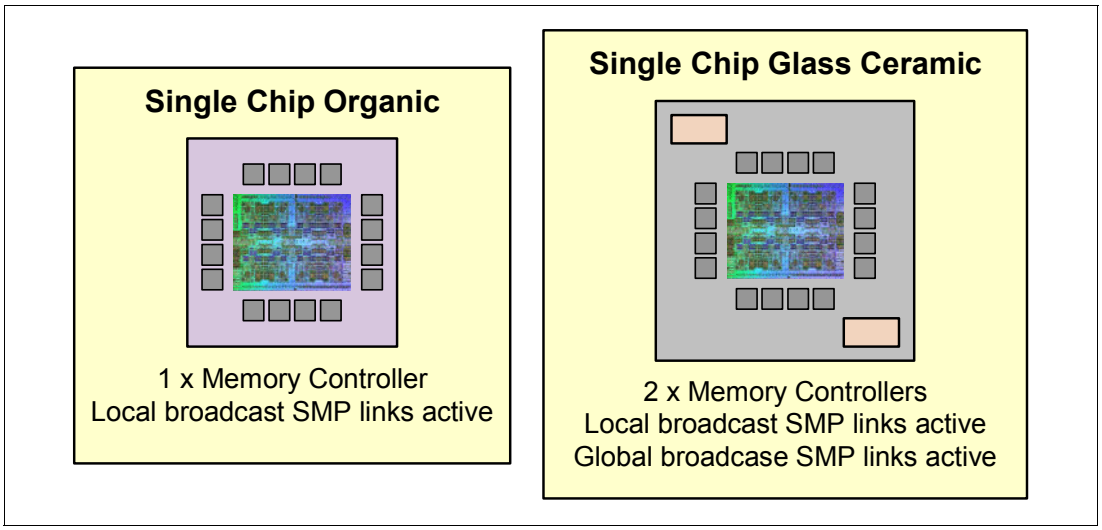

Figure 2-19 Outline of the POWER7 processor physical packaging

POWER7 processors have the unique ability to optimize to various workload types. For example, database workloads typically benefit from fast processors that handle high transaction rates at high speeds. Web workloads typically benefit more from processors with many threads that allow the breaking down of web requests into many parts and handle them in parallel. POWER7 processors uniquely have the ability to provide leadership performance in either case.

TurboCore mode

Users can opt to run selected servers in TurboCore mode. This mode uses four cores per POWER7 processor chip with access to the entire 32 MB of L3 cache (8 MB per core) and at a faster processor core frequency, which delivers higher performance per core, and can save on software costs for those applications that are licensed per core.

Availability: TurboCore is available on the Power 780 and Power 795.

MaxCore mode

MaxCore mode is for workloads that benefit from a higher number of cores and threads handling multiple tasks simultaneously, taking advantage of increased parallelism. MaxCore mode provides up to eight cores and up to 32 threads per POWER7 processor.

POWER7 processor 4-core and 6-core offerings

The base design for the POWER7 processor is an 8-core processor with 32 MB of on-chip L3 cache (4 MB per core). However, the architecture allows for differing numbers of processor cores to be active; 4-cores or 6-cores, and also the full 8-core version.

In most cases (MaxCore mode), the L3 cache that is associated with the implementation is dependent on the number of active cores. For a 6-core version, this behavior typically means that 6 x 4 MB (24 MB) of L3 cache is available. Similarly, for a 4-core version, the L3 cache available is 16 MB.

2.3.6 On-chip L3 cache innovation and Intelligent Cache

A breakthrough in material engineering and microprocessor fabrication has enabled IBM to implement the L3 cache in eDRAM and place it on the POWER7 processor die. L3 cache is critical to a balanced design, as is the ability to provide good signalling between the L3 cache and other elements of the hierarchy such as the L2 cache or SMP interconnect.

The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a Fast Local Region of L3 cache (FLR-L3) but also has access to other L3 cache regions as shared L3 cache. Additionally, each core can negotiate to use the FLR-L3 cache associated with another core, depending on reference patterns. Data can also be cloned to be stored in more than one core's FLR-L3 cache, again depending on reference patterns. This *Intelligent Cache* management enables the POWER7 processor to optimize the access to L3 cache lines and minimize overall cache latencies.

Figure 2-20 shows the FLR-L3 cache regions for each of the cores on the POWER7 processor die.

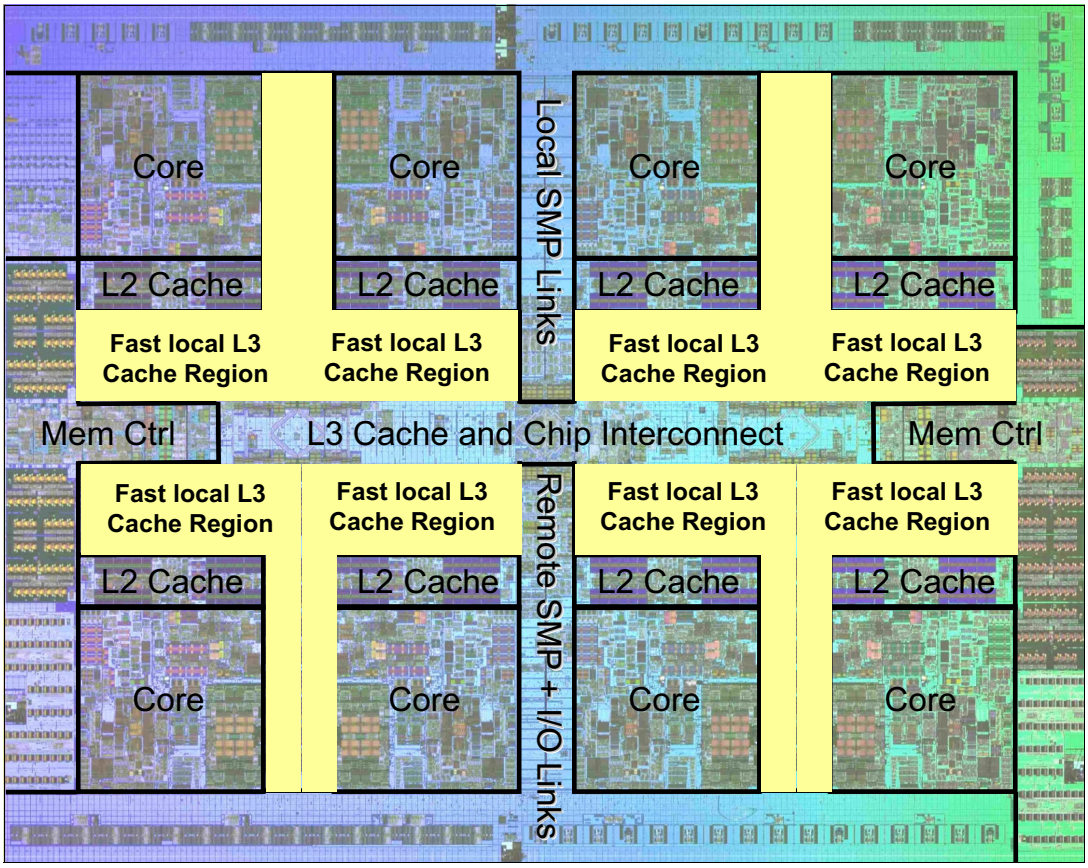

Figure 2-20 Fast local regions of L3 cache on the POWER7 processor

The innovation of using eDRAM on the POWER7 processor die is significant for several reasons:

- Latency improvement

  A six-to-one latency improvement occurs by moving the L3 cache on-chip as compared to L3 accesses on an external (on-ceramic) ASIC.

- Bandwidth improvement

  Improvement of twice the bandwidth occurs with on-chip interconnect. Frequency and bus sizes are increased to and from each core.

- No off-chip driver or receivers

  Removing drivers or receivers from the L3 access path lowers interface requirements, conserves energy, and lowers latency.

- Small physical footprint

  The eDRAM L3 cache requires far less physical space than an equivalent L3 cache implemented with conventional SRAM. IBM on-chip eDRAM uses only a third of the components used in conventional SRAM which has a minimum of 6 transistors to implement a 1-bit memory cell.

- Low energy consumption

  The on-chip eDRAM uses only 20% of the standby power of SRAM.

2.3.7 POWER7 processor and Intelligent Energy

Energy consumption is an important area of focus for the design of the POWER7 processor, which includes *Intelligent Energy* features that help to dynamically optimize energy usage and performance so that the best possible balance is maintained. Intelligent Energy features, such as EnergyScale, work with IBM Systems Director Active Energy Manager™ to dynamically optimize processor speed based on thermal conditions and system utilization.

2.3.8 Comparison of the POWER7 and POWER6 processors

Table 2-5 shows comparable characteristics between the generations of POWER7 and POWER6 processors.

Table 2-5 Comparison of technology for the POWER7 processor and the prior generation

Characteristic	POWER7	POWER6
Technology	45 nm	65 nm
Die size	567 mm^2	341 mm^2
Maximum cores	8	2
Maximum SMT threads per core	4 threads	2 threads
Maximum frequency	4.25 GHz	5 GHz
L2 Cache	256 KB per core	4 MB per core

Characteristic	POWER7	POWER6
L3 Cache	4 MB of FLR-L3 cache per core with each core having access to the full 32 MB of L3 cache, on-chip eDRAM	32 MB off-chip eDRAM ASIC
Memory support	DDR3	DDR2
I/O Bus	Two GX++	One GX++
Enhanced Cache Mode (TurboCore)	Yes	No
Sleep and Nap Mode[a]	Both	Nap only

a. For more information about Sleep and Nap modes, see 2.15.1, “IBM EnergyScale technology” on page 122.

2.4 POWER7 processor books

The IBM Power 795 server with POWER7 processors is a symmetric multiprocessing (SMP) rack-mounted server. Equipped with either eight 32-core or 24-core processor books, the Power 795 server can be deployed from 24-core to 256-core.

Each of the eight POWER7 processor books in a model Power 795 server contain four 6-core or 8-core single-chip modules (SCMs). The 6-core SCM operates at 3.7 GHz. The 8-core SCM operates at either 4.0 GHz or 4.25 GHz.

The model Power 795 server is available starting as low as six active cores, and you can activate one core at a time using optional CoD processor activations.

In standard mode, each 8-core SCM operates at 4.0 GHz with 32 MB of L3 cache and can scale up to 256-cores. When operating in TurboCore mode, each 8-core SCM operates with up to four active cores per SCM at 4.25 GHz and 32 MB of L3 cache, twice the L3 cache per core available than when in standard mode. In TurboCore mode, the Power 795 server can scale up to 128 cores.

TurboCore:

- The minimum number of processor books supported in TurboCore mode is three, with a total of 96 cores. TurboCore mode is specified by FC 9982.
- Switching in and out of TurboCore requires the system to be in a power standby state. The entire system must operate in either standard or TurboCore mode.

The available processor books are listed in Table 2-6.

Table 2-6 Available processor books

Feature code	Description
4700	0/32-core POWER7 4.0 GHz CoD 0-core Active Processor Book
4702	0/24-core POWER7 32 GHz CoD 0-core Active Processor Book

Initially, each processor book has no cores active, and they can be permanently activated using CoD activation features.

Permanent single-core activations for FC 4700 are done with FC 4713 (1-core activation) and 64-core activations with FC 4717. Permanent single-core activations for FC 4702 are done with FC 4714 (1-core activation) and 64-core activations with FC 4718.

Table 2-7 lists the CoD processor activation features and corresponding CoD modes. Additional information about the CoD modes is provided in 2.7, “Capacity on Demand (CoD)” on page 80.

Table 2-7 CoD processor activation features

Feature code	Description	CoD mode	Support		
			AIX	IBM i	Linux
4713	Single core activation for POWER7 CoD Processor Book FC 4700	Permanent	✓	✓	✓
4714	Single core activation for POWER7 CoD Processor Book FC 4702	Permanent	✓	✓	✓
4717	Sixty-four core activations for POWER7 CoD Processor Book FC 4700	Permanent	✓	✓	✓
4718	Sixty-four core activations for POWER7 CoD Processor Book FC 4702	Permanent	✓	✓	✓

The minimum number of permanently activated processors within a system is based on the number of processor books installed. A minimum of 25% of the available processor cores within a system must be activated, or a minimum of 24 activations, whichever is greater.

Table 2-8 details processor book activations for feature code 4700.

Table 2-8 Processor Books activations 4.0 GHz (feature code 4700)

Processor book		Minimum activation (4.0 GHz)		Maximum activation (4.0 GHz)		Maximum activation (4.25 GHz)	
Core	4700	Core	4713	Core	4713	Core	4713
32	1	24/32	24	32/32	32	N/A	N/A
64	2	24/64	24	64/64	64	N/A	N/A
96	3	24/96	24	96/96	96	48/96	48
128	4	32/128	32	128/128	128	64/128	64
160	5	40/160	40	160/160	160	80/160	80
192	6	48/192	48	192/192	192	96/192	96
224	7	56/224	56	224/224	224	112/224	112
256	8	64/256	64	256/256	256	128/256	128

Table 2-9 details processor book activations for FC 4702.

Table 2-9 Processor Books activations 3.7 GHz (FC 4702)

Processor book		Minimum activation		Maximum activation	
Core	**4702**	**Core**	**4714**	**Core**	**4714**
24	1	24/24	24	24/24	24
48	2	24/48	24	48/48	48
72	3	24/72	24	72/72	72
96	4	24/96	24	96/96	96
120	5	30/120	30	120/120	120
144	6	36/144	36	144/144	144
168	7	42/168	42	168/168	168
192	8	48/192	48	192/192	192

Minimum for activation: A minimum of 32 GB total or 2 GB of memory for each active processor core must be activated, whichever is greater. For a server with one FC 4702 processor book installed with six active processors, a minimum of 32 GB of memory must be activated. For a server with eight 4700 processor books installed with 64 activated processors, a minimum of 128 GB of memory must be activated.

The Power 795 can support two partitioning modes, referred to as *packed* and *scattered*:

- The packed mode has a maximum partition size of 32 cores. In packed mode, the hypervisor attempts to always pack partitions into a single processor book. Packed partitioning mode is supported for any number of books.
- The scattered partition mode allows for partition sizes up to the number of cores in the machine, but is restricted to systems with four or more processor books.

Parameter: The IBM POWER Hypervisor™, when assigning resources to the partition, tries to optimize the number of processor books, drawers, and SCMs owning such resources. The Power 795, because of its scalability, poses challenges of this process. The intent is to maximize the throughput needed to allocate partitions, avoiding overutilization of the links between processor books and SCMs. A configuration parameter, the System Partition Processor Limit (SPPL), gives directions to the IBM POWER Hypervisor to contain partitions to a processor book or spread partitions across multiple books.

Figure 2-21 shows the Power 795 processor book architecture.

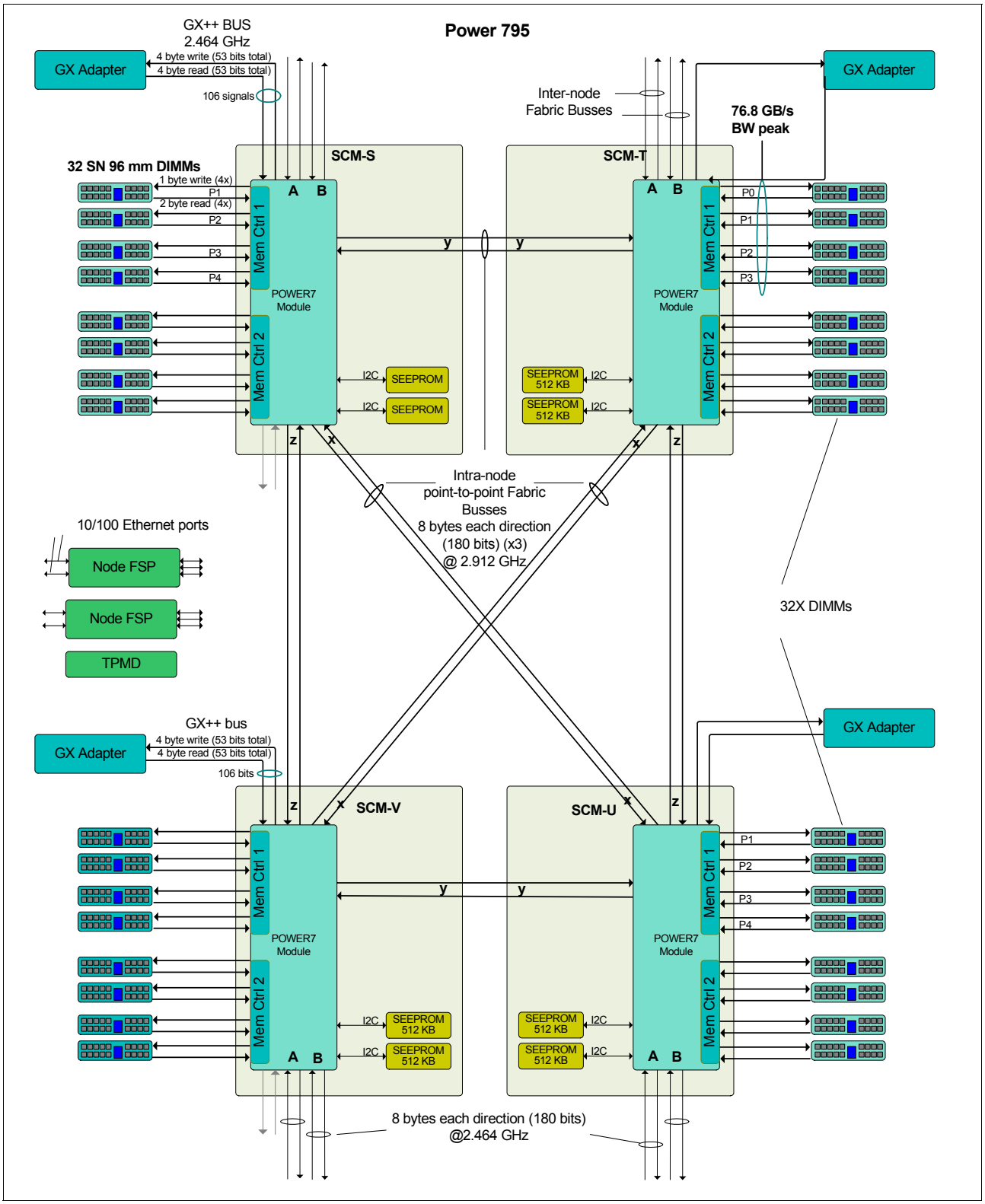

Figure 2-21 IBM Power 795 processor book architecture

Figure 2-22 details the processor book node layout. Each processor book provides the following items:

- Four SMP POWER7 processors chips packaged in four SCMs
- Thirty-two DIMM memory slots
- Four 12X GX++ adapter slots
- Two distributed converter assemblies (DCA)
- Two node controllers
- Redundant thermal power management device (TPMD) card
- Connection to the midplane

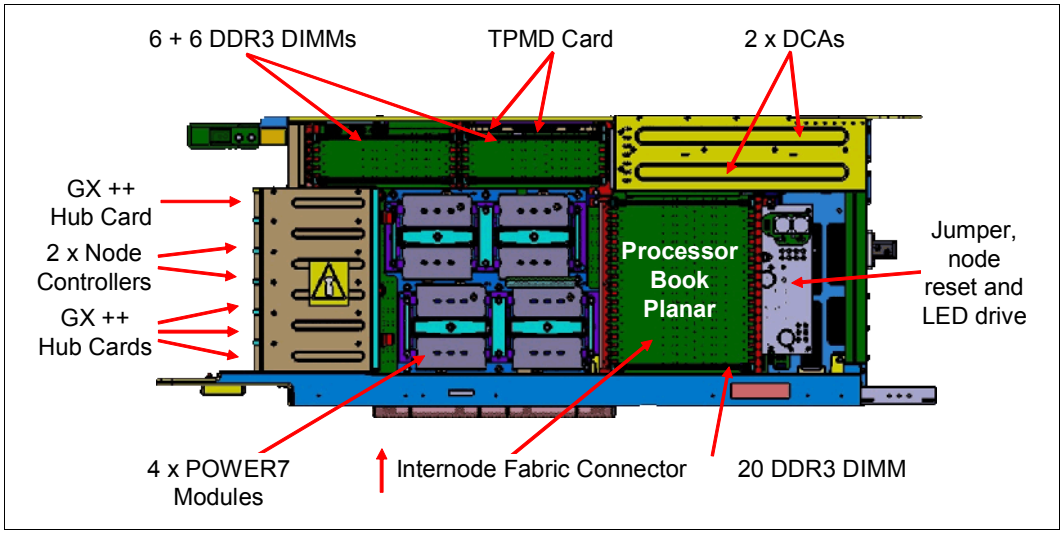

Figure 2-22 Power 795 processor book

Identical books: All eight processor books in a system are identical. They are simply inverted when plugged into the bottom side of the midplane.

2.5 Memory subsystem

The Power 795 server uses fully buffered, Double Data Rate (DDR3) memory DIMMs. It uses a new memory architecture to provide greater bandwidth and capacity, thus enabling operation at a higher data rate for large memory configurations. Each new POWER7 processor book can support up to 32 DDR3 DIMMs running at 1,066 MHz. A full system can contain up to 8 TB of memory.

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

- Error detection and correction code circuitry, designed to detect and correct faults that extend across multiple memory modules (DRAMs). This level includes tolerating a complete DRAM chip failure (Chipkill recovery).
- A spare memory (DRAM) per rank of memory that can be substituted for a failed DRAM module (DRAM sparing). The spares can be used when a DRAM fault is detected and provide additional protection beyond that provided by the error detection and correction circuitry.
- Scrubbing of memory is done to detect and correct intermittent errors.

- Active Memory Mirroring for Hypervisor is an option that mirrors the main memory used by the firmware.
- Special uncorrectable error (SUE) handling prevents an uncorrectable error in memory or cache from immediately causing the system to terminate.
- The L2 and L3 caches in the POWER7 processor are protected with double-bit detect, single-bit correct error correction code (ECC). In addition, the caches maintain a cache line delete capability.
- The bus transferring data between the processor and the memory uses CRC error detection with a failed operation retry mechanism and the ability to dynamically retune bus parameters when a fault occurs. The retry capability allows for correction of transient errors on the data bus and also allowing the memory buffer to correct certain transient errors that might occur internally in the buffer module.
- The memory bus has spare capacity to substitute a spare data bit-line for the one that is determined to be faulty. The bus transferring data between the memory buffer and the DRAMs on a DIMM uses ECC protection and retry of operations to tolerate transient uncorrectable bus errors.

Each POWER7 processor chip uses two memory controllers. Each memory controller has four ports and connects to four DIMMs by a differential channel as shown in Figure 2-23. Each processor book has 32 DIMM slots. Memory DIMMs are available in the following capacities: 8 GB, 16 GB, and 32 GB. Each memory feature (memory unit) provides four DIMMs.

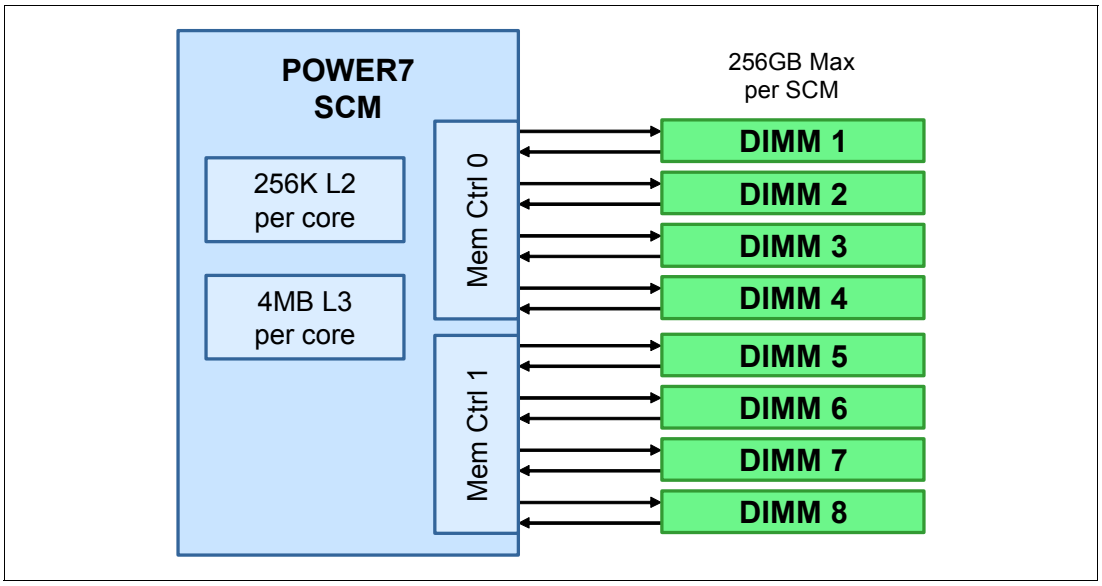

Figure 2-23 Memory system logical view

2.5.1 Memory bandwidth

The POWER7 processor has exceptional cache, memory, and interconnect bandwidths. Table 2-10 shows the bandwidth estimates for a processor book in the systems.

Table 2-10 Memory and IO bandwidth estimates per processor book

Memory		Processor frequency		
		3.72 GHz	**4.00 GHz**	**4.25 GHz**
L1 (data) cache		178.56 GBps	192.0 GBps	204.0 GBps
L2 cache		178.56 GBps	192.0 Gbps	204.0 GBps
L3 cache		119.04 GBps	128.0 GBps	113.6 GBps
System memory	Maximum read bandwidth	102.33 GBps	102.33 GBps	102.33 GBps
	Maximum write bandwidth	51.16 GBps	51.16 GBps	51.16 GBps
	Maximum combined bandwidth	136.44 GBps	136.44 GBps	136.44 GBps

2.5.2 Memory configuration and placement

Each processor book requires a minimum of two memory features installed (8 DIMMs).

All POWER7 memory features must be purchased with sufficient permanent memory activation features so that the system memory is at least 50% active. For two 0/32 GB (FC 5600) features, the minimum memory activated is 32 GB.

Minimums: A minimum of 50% of memory capacity installed must be activated using either memory FC 5600. There is also a recommended system minimum of 2 GB of active memory for each active processor core. Because there is a minimum of 24 cores activated in one processor book (FC 4700 or FC 4702), a minimum of 48 GB of memory is suggested to be activated when one processor book is installed.

Minimum activations that are ordered with all newly purchased FC 5602 or FC 8221 (0 - 1024 GB) is 75%.

Memory features FC 5600 (4x 8 GB), FC 5601(4x 16 GB), and FC 5602 (4x 32 GB) can be mixed within the same POWER7 processor book. For a reference of memory features, see Table 1-12 on page 18.

The best approach is for memory to be installed evenly across all processor books in the system. Balancing memory across the installed processor books allows memory access in a consistent manner and typically results in the best possible performance for your configuration.

Take plans for future memory upgrades into account when deciding which memory feature size to use at the time of initial system order.

Each book must have a minimum of 8 DIMMs. All DIMMs on a processor module must be the same size. All DIMMs on a memory controller must be the same speed. For best performance, ensure that all 32 DIMM are populated.

For an 8-core module, consider the following information if you plan to run an OLTP-type workload, a database, or a highly virtualized workload:

- If all 8 cores per module are licensed (that is, all cores in the system), all memory sites should be used.
- If three-quarters of the cores are licensed, three-quarters or more of the memory sites should be used.
- If half of the cores are licensed, half or more of the memory sites should be used.

For a 6-core module, consider the following information if you plan to run an OLTP type workload, a database, or a highly virtualized workload:

- If all 6 cores per module are licensed (that is, all cores in the system), three-quarters or more of the memory sites should be plugged.
- If three-quarters of the cores are licensed, half or more of the memory sites should be plugged.
- If half of the cores are licensed, half or more of the memory sites should be plugged.

In TurboCore mode, generally the best approach is that half or more of the memory sites be plugged.

Memory feature codes contain 4 DIMMs, so to plug all memory sites, there should be 8 memory features per book. To plug half of the memory sites, there should be 4 memory features per book.

2.6 Bulk power assembly

The Power 795 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 or DC inputs from 380 V to 520 V and converts this to a main isolated 350 V dc nominal bulk power. The bulk power assembly (BPA) holds the bulk power components.

The Power 795 has both primary and redundant BPAs. The BPAs provide the prime power conversion and dc distribution for devices located in the POWER7 795 CEC rack. They are composed of several individual components, all of which support concurrent maintenance and require no special tools.

The primary system rack and powered expansion rack always incorporate two bulk power assemblies for redundancy, which provide 350 V dc power for devices that are 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 must be powered from separate power sources with separate line cords.

Table 2-11 lists the individual components that are part of the BPA.

Table 2-11 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 components 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.
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.

Figure 2-24 shows the front view of a BPA.

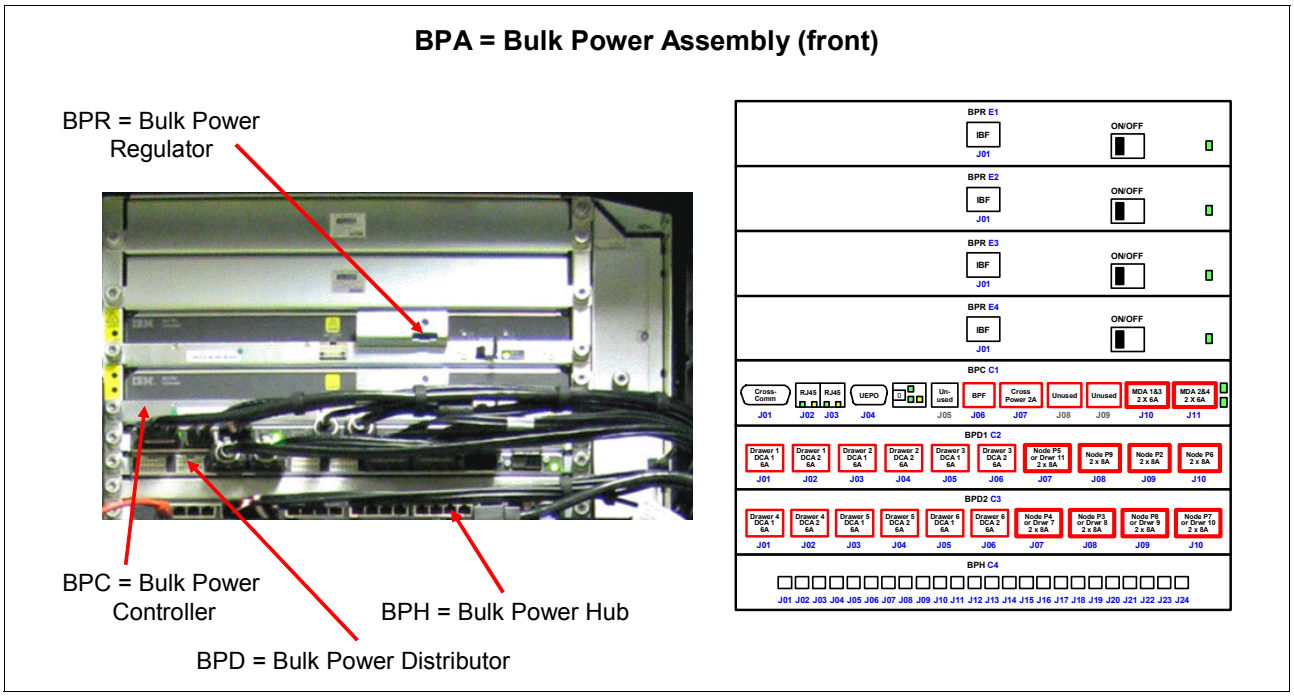

Figure 2-24 Bulk power assembly (BPA)

The power subsystem in the primary system rack is capable of supporting Power 795 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.

2.6.1 Bulk power hub

A 24-port 10/100 Ethernet switch serves as the Power 795 bulk power hub (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.

Figure 2-25 shows bulk power hubs.

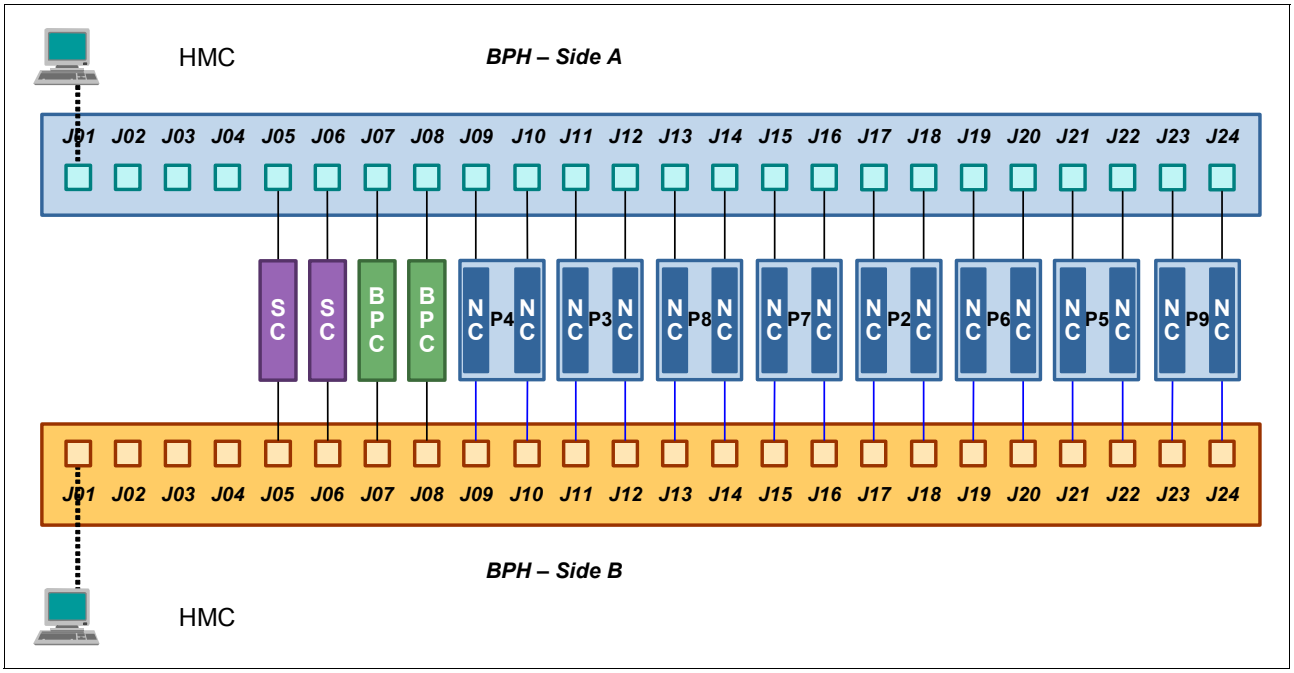

Figure 2-25 Bulk power hubs (BPH)

2.6.2 Bulk power controller

One bulk power controller (BPC), shown in Figure 2-26, 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.

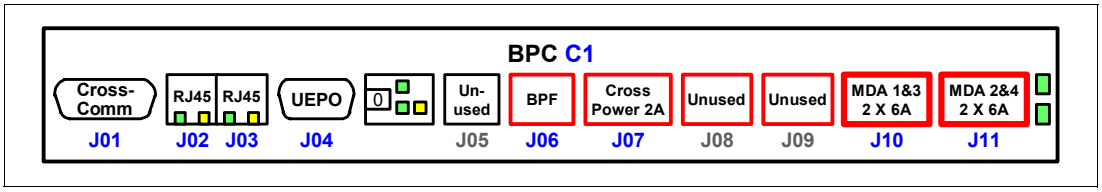

Figure 2-26 Bulk power controller (BPC)

The LED status in the BPC is highlighted in Figure 2-27 on page 77.

UPIC: Connectors J10 and J11 on the BPC contain two power domains and use a multi-universal power input cable (UPIC) cable to connect to the motor drive assemblies (MDAs) on the air movement devices (AMDs).

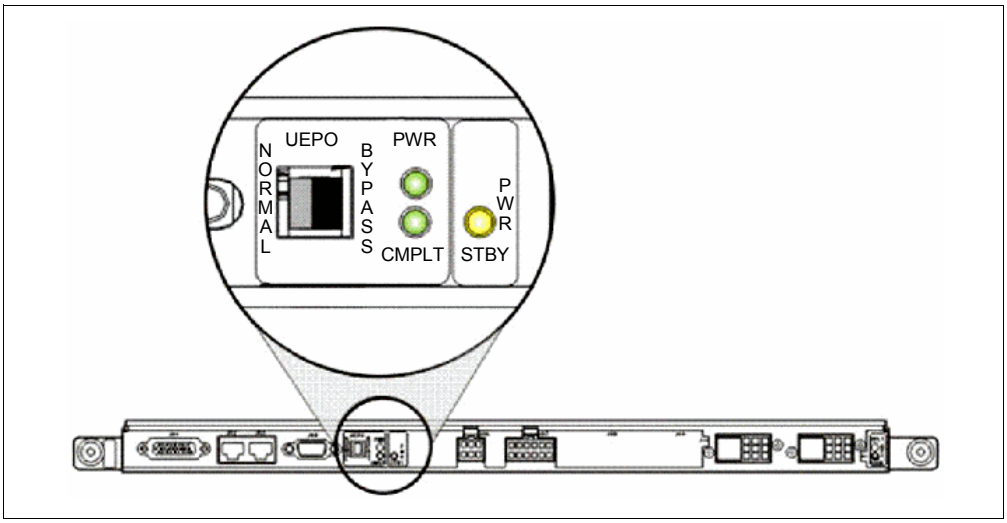

Figure 2-27 LED status in bulk power controller (BPC)

2.6.3 Bulk power distributor

Redundant bulk power distributor (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 2-28 details the BPD assembly.

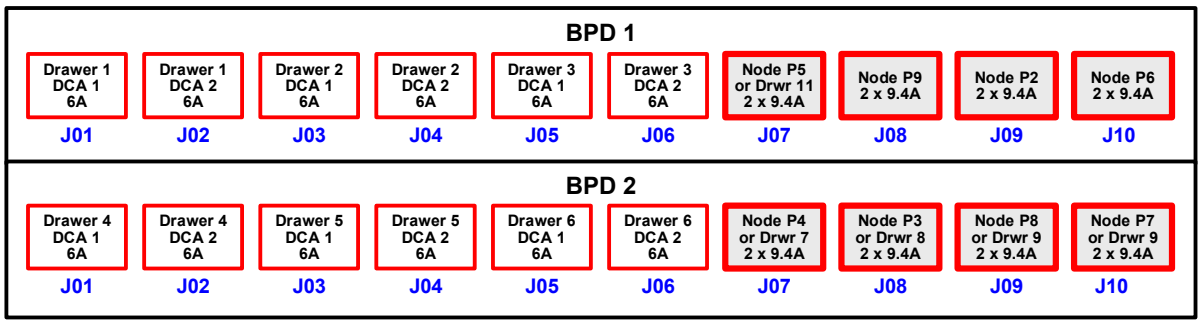

Figure 2-28 Bulk power distributor (BPD) assembly

2.6.4 Bulk power regulator

The redundant bulk power regulators (BPRs) interface to the bulk power assemblies to help ensure proper power is supplied to the system components. Figure 2-29 on page 78 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 that are required depends on configuration, based on the number of processor nodes and I/O drawers that are installed.

Figure 2-29 shows the BPR assembly.

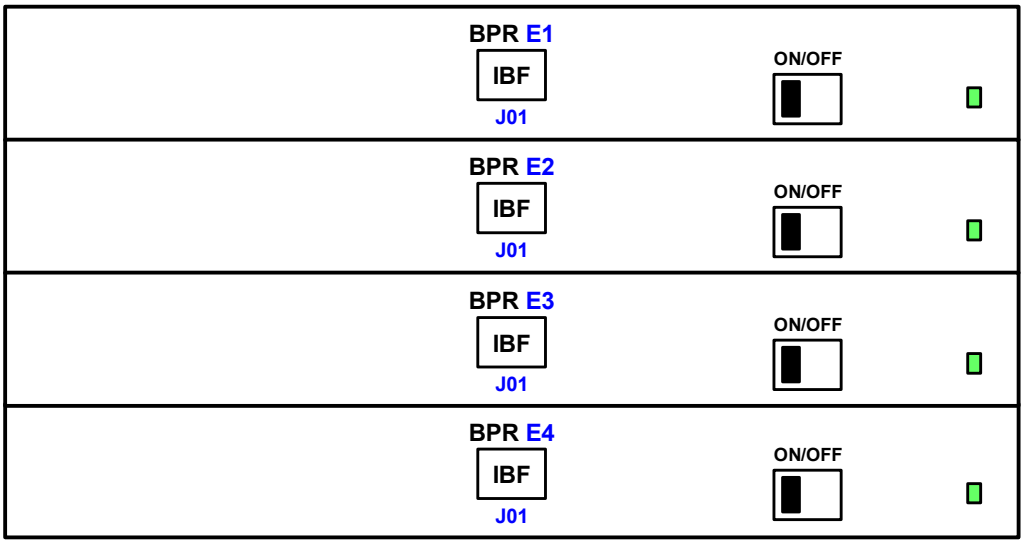

Figure 2-29 Bulk power regulator (BPR) assemblies

2.6.5 Bulk power fan

Each bulk power assembly has a bulk power fan (BPF) for cooling the components of the bulk power enclosure. The BPF is powered by the universal power input cable (UPIC) connected to connector J06 on the BPC. The BPF is shown in Figure 2-30.

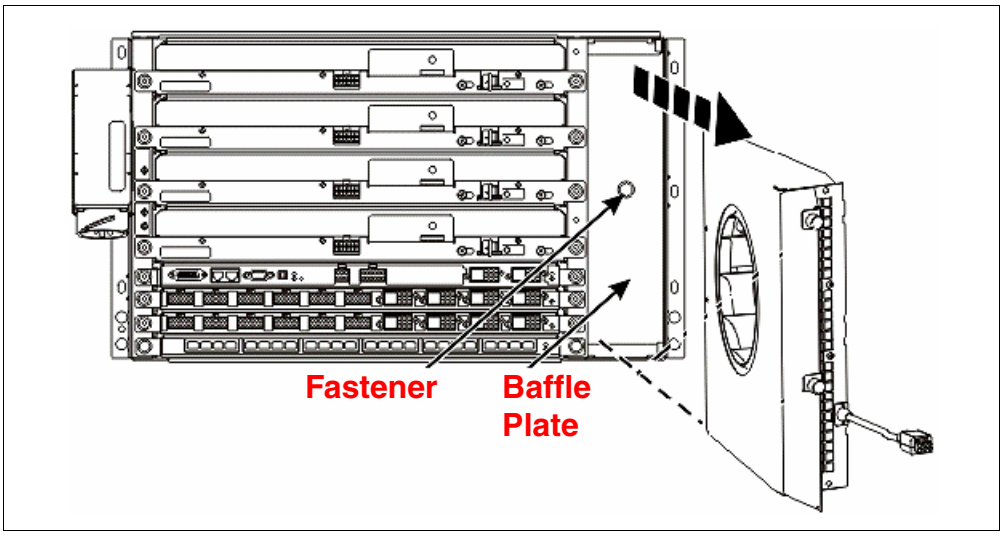

Figure 2-30 Bulk power fan (BPF)

2.6.6 Integrated battery feature

An optional integrated battery feature (IBF) is available for the Power 795 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 are located in each configured rack: CEC, powered expansion rack, and nonpowered bolt-on rack. When ordered, the IBFs displace the media drawer or an I/O drawer. In the CEC rack, two positions, U9 and U11 (located below the processor books) are each 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 occupies 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. In the powered expansion rack, 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.

2.7 Capacity on Demand (CoD)

Several types of CoD are available on the Power 795 server to help meet changing resource requirements in an on-demand environment, by using resources that are installed on the system but that are not activated.

As part of the IBM commitment to delivering the most flexible and resilient Power high-end systems, and as a standard, all new Power 780 (9179-MHD) and Power 795 servers include a specific number of Elastic On/Off CoD processor and memory days, depending on the configuration of the system at initial ship or upgrade time. For every new Power 780 or Power 795, 15 Elastic CoD On/Off processor days and 240 GB memory days are included at no charge for every processor core initially included with the system.These Elastic On/Off CoD processor and memory days must be used in accordance with the Temporary Capacity on Demand terms and conditions.

The Power Systems Pools offering is a multisystem IBM Power 780 and Power 795 infrastructure offering designed to enable a highly resilient and flexible IT environment in support of large-scale server consolidation and your most demanding business application requirements. Power Systems Pools allows for the aggregation of on/off and Elastic CoD compute resources, including processors and memory, across a number of Power 780 and Power 795 servers. It delivers greater flexibility to respond to critical application workload requirements and also to enhance the availability of your applications.

As part of the offering, those Power 780 and Power 795 servers that participate in an IBM Power Systems Pools environment are provided with regularly planned maintenance events (up to eight in a calendar year) that enable you to turn on the inactive processors and memory in another participating system in the pool if a system needs to be stopped for some type of maintenance occurrence. In addition, all on/off and Elastic CoD processor and memory days, including those that now come standard with all new Power 780 and Power 795 servers, can be accumulated and managed at a pool level to effectively and efficiently manage and balance application workload peak requirements. Power Systems Pools can consist of up to 10 Power 780 or Power 795 systems to support large scale application, database, and infrastructure requirements.

New to both Power 780 model MHD and Power 795 model FHB are 90-day temporary On/Off CoD processor and memory enablement features. These features enable a system to temporarily activate all inactive processor and memory CoD resources for a maximum of 90 days before you must order an other temporary on/off enablement feature number. Also new for Power 780 model MHD are high-density memory DIMMs that use 4 Gb technology. These memory DIMMs are for 64 GB, 128 GB, and 256 GB DDR3 memory features. This feature enables a new 256 GB DDR3 memory feature increase the mode FHB maximum memory capacity of 8 - 16 TB. IBM continues to offer the 32 GB, 2 Gb memory feature.

2.7.1 Capacity Upgrade on Demand (CUoD)

CUoD allows you to purchase an additional permanent processor or memory capacity and dynamically activate them when needed.

Processors can be activated in increments of one processor, while memory can be activated in increments of 1 GB. As your workload demands require more processing power, activating the inactive processors or memory is possible simply by placing an order for an activation feature. You can retrieve, over the Internet, an electronically encrypted activation code that unlocks the desired amount of capacity. There is no hardware to ship and install, and no additional contract is required

2.7.2 On/Off Capacity on Demand (On/Off CoD)

On/Off CoD enables processors or memory to be temporarily activated in full-day increments as needed.

Charges are based on usage reporting that is collected monthly. Processors and memory can be activated and turned off an unlimited number of times, when additional processing resources are needed.

This offering provides to a system administrator an interface at the HMC to manage the activation and deactivation of resources. A monitor that resides on the server records the usage activity. This usage data must be sent to IBM on a monthly basis. A bill is then generated based on the total amount of processor and memory resources utilized, in increments of processor and memory (1 GB) days.

Before using temporary capacity on your server, you must enable your server. To do this step, an enablement feature (MES only) must be ordered and the required contracts must be in place.

If a Power 795 server uses the IBM i operating system in addition to any other supported operating system on the same server, the client must inform IBM which operating system caused the temporary On/Off CoD processor usage so that the correct feature can be used for billing.

The features that are used to order enablement codes and support billing charges are detailed in Table 2-12 and in Table 2-13 on page 81.

Table 2-12 On/Off CoD processor order enablement codes and support billing charges

Processor	Processor enablement feature	AIX, Linux processor billing feature	IBM i processor billing feature
4700	7971	4704	4712
4702	7971	4711	4724
7560	7971	4704	4712

Table 2-13 On/Off COD memory order enablement codes and support billing charges

Memory feature	Memory enablement feature	AIX, Linux, IBM i memory billing feature
5600	7973	7377
5601	7973	7377
5602	7973	7377

The On/Off CoD process consists of three steps: enablement, activation, and billing.

- Enablement

  Before requesting temporary capacity on a server, you must enable it for On/Off CoD. To do this step, order an enablement feature and sign the required contracts. IBM will generate an enablement code, mail it to you, and post it on the web for you to retrieve and enter on the target server.

  A *processor enablement* code allows you to request up to 360 processor days of temporary capacity. If the 360 processor-day limit is reached, place an order for another processor enablement code to reset the number of days that you can request back to 360.

  A *memory enablement* code allows you request up to 999 memory days of temporary capacity. If you reach the limit of 999 memory days, place an order for another memory enablement code to reset the number of allowable days you can request back to 999.

- Activation

  When On/Off CoD temporary capacity is needed, request it on the HMC menu for On/Off CoD. Specify how many of the inactive processors or GB of memory are required to be temporarily activated for some number of days. You will be billed for the days requested, whether the capacity is assigned to partitions or left in the shared processor pool.

  At the end of the temporary period (days that were requested), you must ensure that the temporarily activated capacity is available to be reclaimed by the server (not assigned to partitions), or you are billed for any unreturned processor days.

- Billing

  The contract, signed by the client before receiving the enablement code, requires the On/Off CoD user to report billing data at least once a month (whether or not activity occurs). This data is used to determine the proper amount to bill at the end of each billing period (calendar quarter). Failure to report billing data for use of temporary processor or memory capacity during a billing quarter can result in default billing equivalent to 90 processor days of temporary capacity.

For more information regarding registration, enablement, and usage of On/Off CoD, see the following location:

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

2.7.3 Utility Capacity on Demand (Utility CoD)

Utility CoD autonomically provides additional processor performance on a temporary basis within the shared processor pool.

Utility CoD enables you to place a quantity of inactive processors into the server's shared processor pool, which then becomes available to the pool's resource manager. When the server recognizes that the combined processor utilization within the shared processor pool exceeds 100% of the level of base (purchased and active) processors that are assigned across uncapped partitions, 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 automatically allows the additional Utility CoD processors to be used. The system automatically and continuously monitors and charges for the performance needed above the base (permanent) level.

Registration and usage reporting for Utility CoD is made using a public website and payment is based on reported usage.

If a Power 795 server uses the IBM i operating system in addition to any other supported operating system on the same server, the client must inform IBM which operating system caused the temporary Utility CoD processor usage so that the correct feature can be used for billing.

Table 2-14 lists utility COD billing features.

Table 2-14 Utility COD Billing features

Utility COD feature	AIX, Linux billing	IBM i billing
100 Processor minutes for FC 4700	FC 4606	FC 4719
100 Processor minutes for FC 4702	FC 4607	FC 4722

Registration and usage reporting for Utility CoD is made using a public website and payment is based on reported usage. Utility CoD requires PowerVM Standard Edition (FC 7943) or PowerVM Enterprise Edition (FC 8002) to be active on the 9119-FHB. For more information about registration, enablement, and use of Utility CoD, visit the following web location:

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

2.7.4 Trial Capacity On Demand (Trial CoD)

Trial CoD is a function delivered with all Power Systems servers supporting CUoD resources. Those servers with standby CUoD processors or memory will be capable of using a one-time, no-cost activation for a maximum period of 30 consecutive days. This enhancement allows for benchmarking of CUoD resources or can be used to provide immediate access to standby resources when the purchase of a permanent activation is pending.

A *standard request* for Trial CoD requires you to complete a form including contact information and vital product data (VPD) from your Power 795 system with inactive CoD resources.

A standard request activates two processors or 4 GB of memory (or both two processors and 4 GB of memory) for 30 days. Subsequent standard requests can be made after each purchase of a permanent processor activation. An HMC is required to manage Trial CoD activations.

An *exception request* for Trial CoD requires you to complete a form including contact information and VPD from your Power 795 system with inactive CoD resources. An exception request will activate all inactive processors or all inactive memory (or all inactive processor and memory) for 30 days. An exception request can be made only one time over the life of the machine. An HMC is required to manage Trial CoD activations.

To request either a standard or an exception trial, see the following web location:

https://www-912.ibm.com/tcod_reg.nsf/TrialCod?OpenForm

2.7.5 Capacity backup

The Power Systems capacity backup (CBU) offerings are designed to support our client's disaster recovery and high availability needs. The CU offerings recognize that true high availability or disaster recovery solutions require at least two systems. If one system is not available the other one “takes over.” The CBU offering is designed to give clients flexible and economic options for deploying business continuity operations.

The IBM Power 795 CBU solution for disaster recovery (DR) offers an offsite DR machine at an affordable price with a choice of two CBU features, FC 7560 or FC 7562. Power 795 CBU offerings allow flexibility in selecting active and standby processors in four configurations using equivalents of the two standard Power 795 processor books, FC 4700 or FC 4702.

CBU feature FC 7560 offers four separate maximum standby configurations using a 0/32-core, equivalent FC 4700 processor book. Each FC 7560 configuration provides multiple increments of 900 On/Off CoD billing credits. Each of the CBU offerings provides a fixed number of permanently activated processors. CBU FC 7560 allows either 4/64, 8/128, 12/192, or 16/256 permanently activated cores.

CBU feature FC 7562 offers standby configurations that use a 0/24-core, equivalent FC 4702 processor book. Each FC 7562 configuration provides multiple increments of 680 On/Off CoD billing credits. CBU FC 7562 allows either 3/48, 6/96, 9/144, or 12/192 permanently activated cores.

Difference: This activation differs from the 25% minimum activation that is required with orders for the standard FC 4700 or FC 4702 processor book.

CoD activations of Power 795 CBU servers require the On/Off CoD Enablement function discussed under “Enablement” in 2.7.2, “On/Off Capacity on Demand (On/Off CoD)” on page 81 (FC 7971, FC 7973). The On/Off CoD enablement code does not ship with the CBU server because customer contracts and Sales Channel registration are required prior to processing an MES order for On/Off CoD enablement feature FC 7971. The On/Off CoD usage days that are provided with the CBU server will be reflected as credits in the quarterly On/Off CoD Usage and Billing report, which is distributed to the Sales Channel that registers the CBU system for On/Off CoD.

The no-charge processor days provided with the FC 7560 and FC 7562 are intended to be available for use in a disaster, other failover situations, or for failover testing; nevertheless, they can be used as you determine are appropriate. However, when the no-charge processor days provided with the CBU system are exhausted, additional temporary CoD processor-days may be purchased under the same terms and conditions that are available when ordering the Processor Billing feature for processor feature FC 4700 or FC 4702.

Temporary on/off memory (gigabyte days) are not provided with CBU features FC 7560 and FC 7562. On/off memory can be enabled for a CBU server by ordering On/Off Memory Enablement FC 7973.

The standby processors of FC 7560 and FC 7562 cannot be permanently activated. After exhausting the On/Off CoD billing credits provided with the CBU offering, additional On/Off CoD usage days may be purchased at regular On/Off CoD billing feature prices. Regular use of standby processors outside a disaster might be costly.

Table 2-15 shows the available processor options for a CBU Power 795 system.

Table 2-15 Capacity Backup options for Power 795

Active/Installed cores	Processor speed	Feature codes required	Minimum activation feature codes	On/Off CoD processing-days credit included
4/64	4.0 GHz	2 x 7560	4713 x4	1800[a]
8/128	4.0 GHz	4 x 7560	4713 x8	3600
12/192	4.0 GHz	6 x 7560	4713 x12	5400
16/256	4.0 GHz	8 x 7560	4713 x16	7200
3/48	3.7 GHz	2 x 7562	4714 x3	1360[b]
6/96	3.7 GHz	4 x 7562	4714 x6	2720
9/144	3.7 GHz	6 x 7562	4714 x9	4080
12/192	3.7 GHz	8 x 7562	4714 x12	5440

a. Each 7560 CBU node includes 900 on/off processor days.
b. Each 7562 CBU node includes 680 on/off processor days.

Feature codes: FC 4700 and FC 4702 in standard Power 795 configurations are replaced by FC 7560 and FC 7562 in CBU configurations. FC 4700 and FC 4702 cannot be installed in the same server as FC 7560 and FC 7562.

For more information about CoD usage, see the following website:

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

2.7.6 Capacity backup for IBM i

The Power 795 server's capacity backup (CBU) designation can help meet your requirements for a second system to use for backup, high availability, and disaster recovery. You can temporarily transfer IBM i processor license entitlements and 5250 Enterprise Enablement entitlements purchased for a primary machine to a secondary CBU-designated system. Temporarily transferring these resources instead of purchasing them for your secondary system can result in significant savings. Processor activations cannot be transferred.

The CBU specify feature FC 4896 is available only as part of a new server purchase or during an MES upgrade from an existing system to a 9119-FHB. Certain system prerequisites must be met and system registration and approval are required before the CBU specify feature can be applied on a new server.

The CBU for IBM i specify feature FC 4896 can be used with any of the Power 795 processor book features, including the CBU for DR. However, temporary transfers of IBM i licensing and 5250 entitlements are to permanently activated processors only. The CBU for DR processor books FC 7560 and FC 7562 have relatively few permanently activated processors lessening the usefulness of CBU of IBM i in this situation. IBM i licenses cannot be transferred to on/off processors.

For more information about CBU and to register a CBU machine, go to the following location:

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

The Power 795 server's CBU designation can help meet your requirements for a second system to use for backup, high availability, and disaster recovery. You can temporarily transfer IBM i processor license entitlements and 5250 Enterprise Enablement entitlements, purchased for a primary machine, to a secondary CBU-designated system. Temporarily transferring these resources instead of purchasing them for your secondary system might result in significant savings. Processor activations cannot be transferred.

The CBU specify feature FC 4896 is available only as part of a new server purchase or during an MES upgrade from an existing system to a 9119-FHB. Certain system prerequisites must be met and system registration and approval are required before the CBU specify feature can be applied on a new server.

The CBU for IBM i specify FC 4896 can be used with any of the Power 795 processor book features, including the CBU for DR. However, temporary transfers of IBM i licensing and 5250 entitlements are to permanently activated processors. The CBU for DR processor books 7560 and 7562 have relatively few permanently activated processors, lessening the usefulness of CBU of IBM i in this situation.

Standard IBM i terms and conditions do not allow either IBM i processor license entitlements or 5250 OLTP (Enterprise Enablement) entitlements to be transferred permanently or temporarily. These entitlements remain with the machine they were ordered for. When you register the association between your primary and on-order CBU system, you must agree to certain terms and conditions regarding the temporary transfer.

After a CBU system designation is approved and the system is installed, you can temporarily move your optional IBM i processor license entitlement and 5250 Enterprise Enablement entitlements from the primary system to the CBU system when the primary system is down or while the primary system processors are inactive. The CBU system can then better support failover and role swapping for a full range of test, disaster recovery, and high availability scenarios. Temporary entitlement transfer means that the entitlement is a property transferred from the primary system to the CBU system and may remain in use on the CBU system as long as the registered primary and CBU system are in deployment for the high availability or disaster recovery operation.

The primary system for a Power 795 server can be one of the following systems:

- 795 (9119-FHB)
- 780 (9179 MHB)
- 595 (9119-FHA)
- 595 (9406-595)

These systems have IBM i software licenses with an IBM i P50 software tier or higher. The primary machine must be in the same enterprise as the CBU system.

Before you can temporarily transfer IBM i processor license entitlements from the registered primary system, you must have more than one IBM i processor license on the primary machine and at least one IBM i processor license on the CBU server. An activated processor must be available on the CBU server to use the transferred entitlement. You may then transfer any IBM i processor entitlements above the minimum one, assuming the total IBM i workload on the primary system does not require the IBM i entitlement you would like to transfer during the time of the transfer.

During this temporary transfer, the CBU system's internal records of its total number of IBM i processor license entitlements are not updated, and you might see IBM i license noncompliance warning messages from the CBU system. Such messages that arise in this situation do not mean you are not in compliance.

Before being able to temporarily transfer 5250 entitlements, you must have more than one 5250 Enterprise Enablement entitlement on the primary server and at least one 5250 Enterprise Enablement entitlement on the CBU system. You may then transfer the entitlements that are not required on the primary server during the time of transfer and that are above the minimum of one entitlement.

2.7.7 Software licensing and CoD

For software licensing considerations with the various CoD offerings, see the most recent revision of the *Capacity on Demand User's Guide*:

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

2.7.8 MaxCore, TurboCore, and CoD

The Power 795 has the capability to run in two modes:

- MaxCore mode allows for all processor cores in the system to be activated.
- TurboCore mode allows for half of the processor cores in the system to be activated, but the cores run at a higher speed and have access to the entire L3 cache on the chip.

Choosing which mode the Power 795 starts in when powering on the system is possible.

In MaxCore mode, the number of processor cores available for use by partitions remains equal to the number or processor activations purchased. In TurboCore mode, the same is true, but if the number of processor core activations purchased is greater than 50% of the number of processor cores physically installed, then the number of processor cores available for partitions is limited to 50% of installed processor cores. The excess core activations are not available until the Power 795 is restarted in MaxCore mode.

Table 2-16 shows examples of processor activations.

Table 2-16 Processor activations with MaxCore and TurboCore

Physically installed cores	Processor activations purchased	Cores available for use by partitions	
		MaxCore	TurboCore
128	32	32	32
128	80	80	64
128	128	128	64

2.8 Internal I/O subsystem

Each processor book on the Power 795 server provides four GX++ busses for the attachment of GX++ bus adapters (also known as GX++ hub adapters, or I/O hubs). A fully configured Power 795 server with eight processor books supports up to 32 GX++ bus adapters.

The GX++ bus adapter locations on upper and lower processor book are shown in Figure 2-31 on page 88.

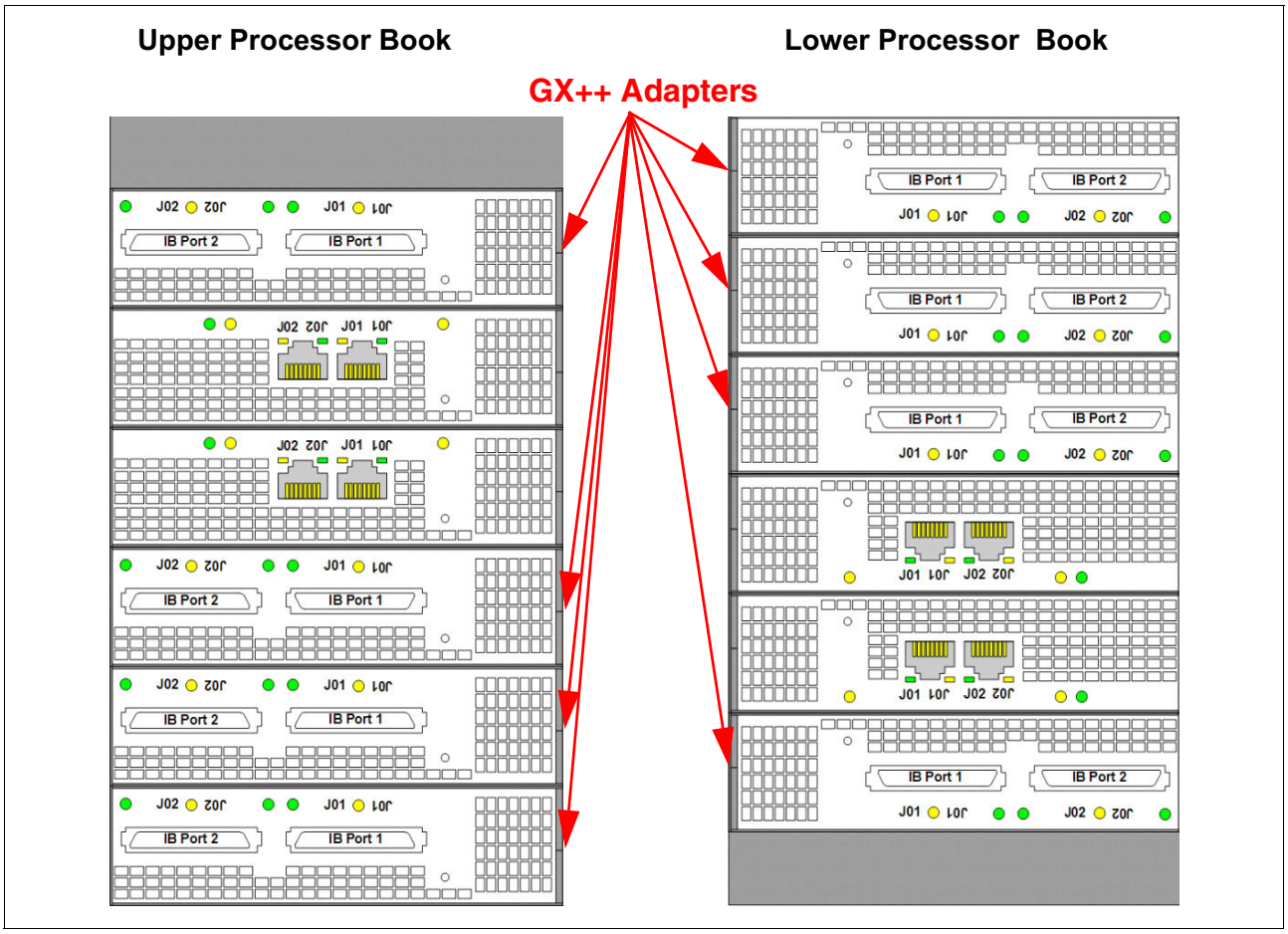

Figure 2-31 GX++ bus adapters

2.8.1 Connection technology

The GX++ adapters that are installed on the processor books provide connectivity to the I/O drawers, and also to InfiniBand switches through a 10 Meter 12X to 4X Enhanced Channel Conversion (FC 1854). Each GX++ adapter provides two ports. Table 2-17 shows the feature code and description for the GX++ adapter.

Table 2-17 GX++ adapter card

Feature	Description	Form factor	Attach to drawers	Support		
				AIX	IBM i	Linux
1816	GX Dual-Port 12X HCA	narrow	5797, 5798, 5803, 5873	✓	✓	✓

GX++ adapter plugging rules

The GX++ adapters are evenly distributed across the installed processor books. When adding GX++ adapter cards to newly added nodes, use the rules in the following steps:

1. Populate slots farthest from the midplane first.
2. Distribute cards among all the processor books, starting with the first book plugged, and proceeding in book plug order up to a maximum of 32 cards.

Important: When your Power 795 server is manufactured, the GX++ adapters are evenly distributed across the installed processor books. I/O connections will then be distributed across these installed GX++ adapters. If you add more GX++ adapters during an upgrade, install them so that the result is an even balance across all new and existing processor books. Therefore, the cabling relationship between the GX++ adapters and drawers can vary with each Power 795 server. We suggest that you document these connections to assist with system layout and maintenance.

2.9 Internal I/O subsystem

Each processor book on the Power 795 server provides four GX++ busses for the attachment of GX++ bus adapters (also known as GX++ hub adapters, or I/O hubs). A fully configured 795 server with eight processor books supports up to 32 GX++ bus adapters. The GX++ bus adapter locations on the upper and lower processor book are shown in Figure 2-32 on page 90.

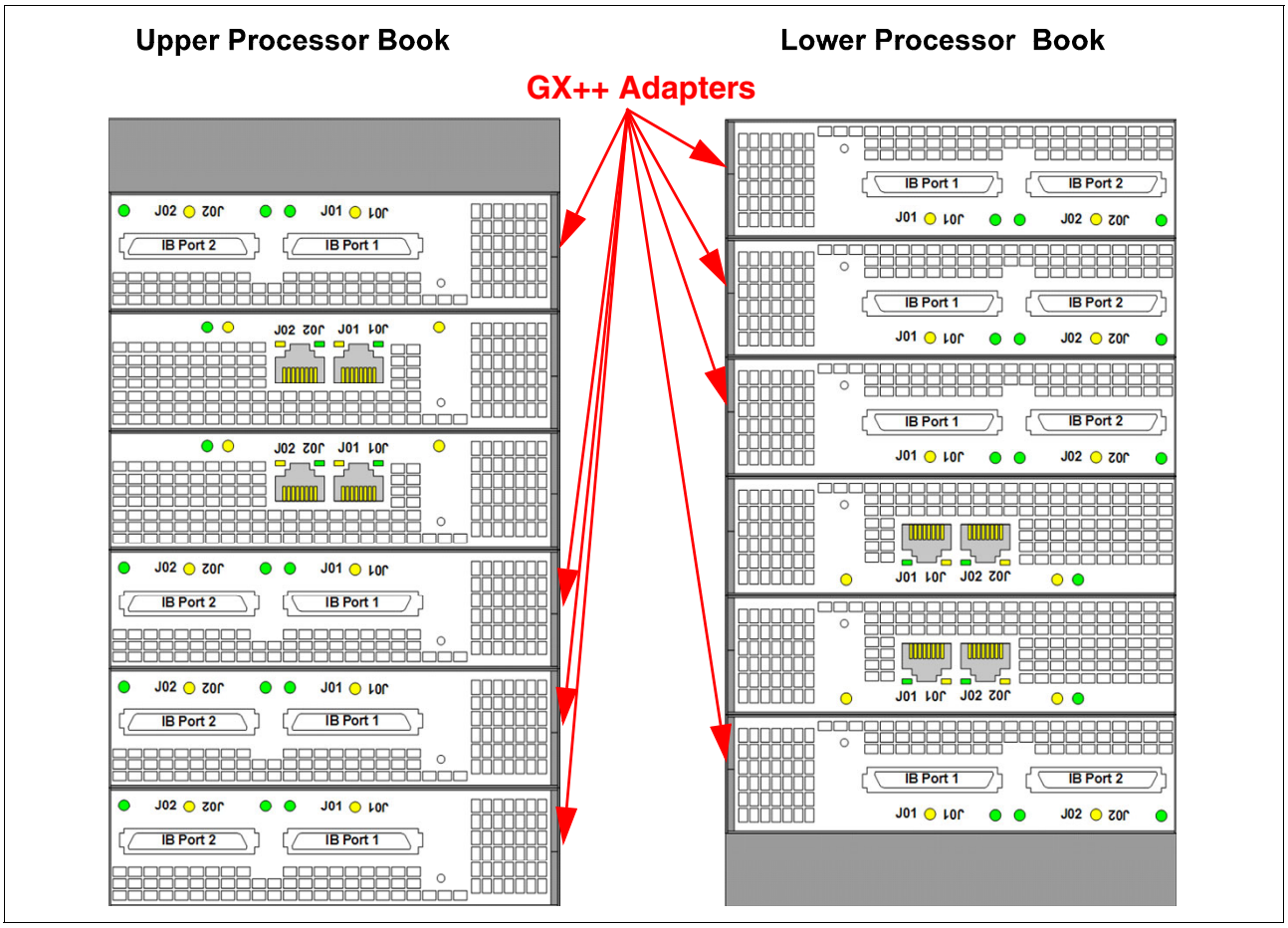

Figure 2-32 GX++ bus adapters

2.9.1 Connection technology

The GX++ adapters that are installed on the processor books provide connectivity to the I/O drawers and to InfiniBand switches through a 10 Meter 12X to 4X Enhanced Channel Conversion (FC 1854). Each GX++ adapter provides two ports. Table 2-18 on page 90 lists the feature code and description for the GX++ adapter.

Table 2-18 GX++ adapter card

Feature code	Description	Form factor	Attach to drawer	Support		
				AIX	IBM i	Linux
1816	GX Dual-Port 12X HCA	Narrow	FC 5797, FC 5798, FC 5803, FC 5873	✓	✓	✓

GX++ adapter plugging rules

The GX++ adapters are evenly distributed across the installed processor books. When adding GX++ adapter cards to newly added nodes, use the rules in the following steps:

1. Populate slots farthest from the midplane first.
2. Distribute cards among all the processor books, starting with the first book plugged, and proceeding in book plug order up to a maximum of 32 cards.

Important: When your Power 795 server is manufactured, the GX++ adapters are evenly distributed across the installed processor books. I/O connections are then distributed across these installed GX++ adapters. If you add more GX++ adapters during an upgrade, install them so that the result is an even balance across all new and existing processor books. Therefore, the cabling relationship between the GX++ adapters and drawers can vary with each Power 795 server. We suggest that you document these connections to assist with system layout and maintenance.

2.9.2 Internal I/O drawers

The internal I/O drawers (24 inches) provide storage and I/O connectivity for the Power 795 server. The available internal I/O drawers are listed in Table 2-19.

Table 2-19 Internal I/O drawers

Feature code	Description	Support		
		AIX	IBM i	Linux
5803	12X I/O Drawer PCIe, SFF disk,20 PCIe slots, 26 SFF disk bays	✓	✓	✓
5873	12X I/O Drawer PCIe, SFF no disks, 20 PCIe slots	✓	✓	✓
5797[a]	12X I/O drawer, 20 PCI-X slots, 16 disk bays, with repeater	✓	✓	✓
5798[a]	12X I/O drawer, 20 PCI-X slots, 16 disk bays, no repeater	✓	✓	✓

a. Supported, but no longer orderable

The primary I/O drawer used in Power 795 is the 12X I/O Drawer PCIe, SFF disk,20 PCIe slots, 26 SFF disk bays (FC 5803). I/O drawers FC 5797/FC 5798 are allowed as a primary drawer when the system is an upgrade from Power 595.

12X I/O Drawer PCIe

The 12X I/O Drawer with SFF disks (FC 5803) is a 4U high I/O drawer containing 20 PCIe slots and 26 SAS hot-swap SFF disk bays. This drawer attaches to the central electronics complex by 12X DDR cables (FC 1862, FC 1863 or FC 864). When SFF disks are installed on the FC 5803, they are driven by at least one SAS PCIe adapter and SAS AT cable (FC 3688). Using a mode switch, the 26 SFF slots can be configured as one group of 26 slots (AIX and Linux) or two groups of 13 slots (AIX, IBM i, and Linux) or four groups of 6 or 7 slots (AIX and Linux).

The FC 5873 drawer has the same PCIe slots as the FC 5803, but does not have the SFF disk bays.

Figure 2-33 shows a diagram of the FC 5803 internal I/O drawer.

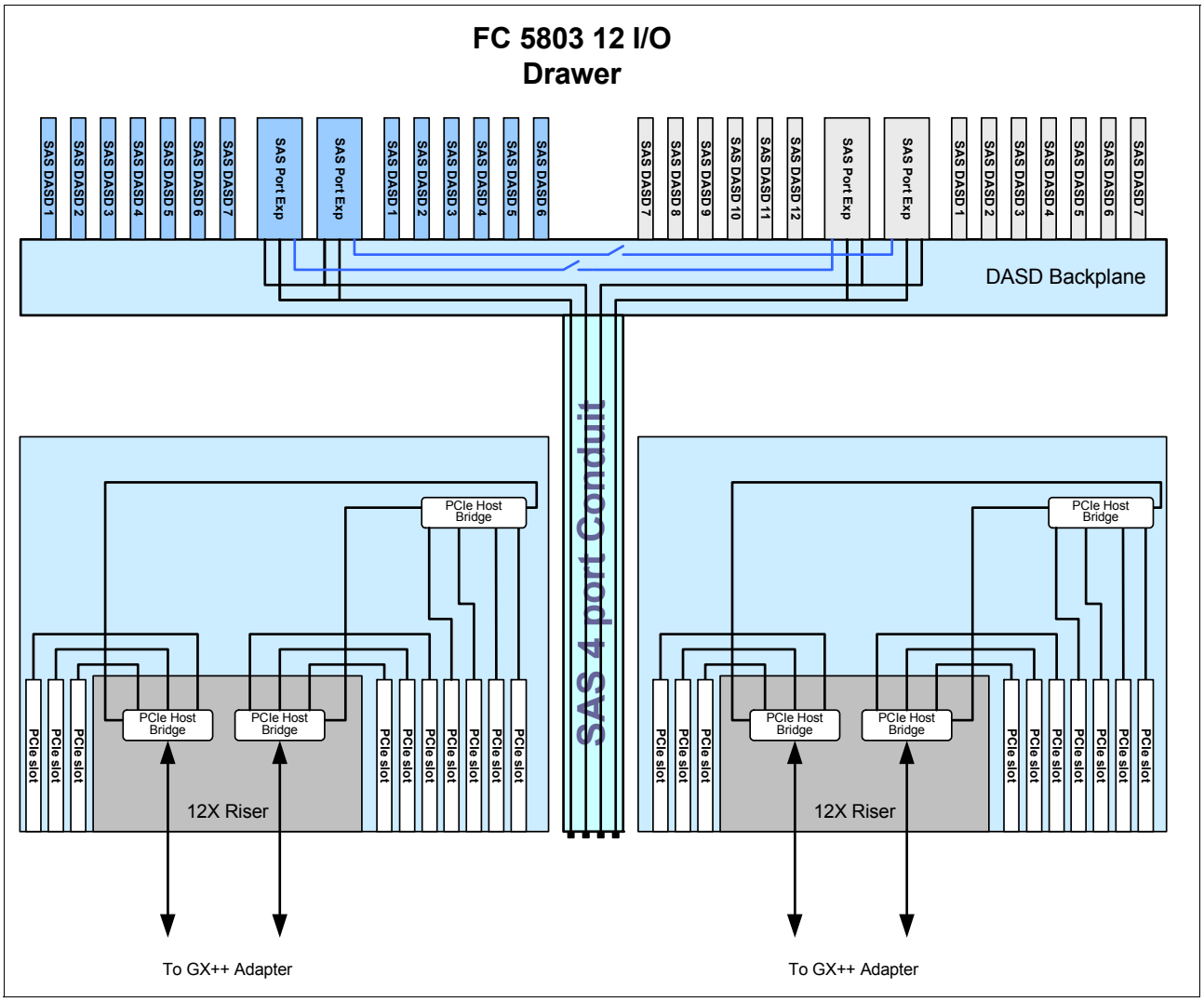

Figure 2-33 FC *5803 internal I/O drawer*

2.9.3 12X I/O Drawer PCI-X

The 12X I/O Drawer PCI-X with repeater (FC 5797) provides a 4U high I/O drawer containing 20 PCI-X slots and 16 hot-swap SCSI disk bays. This drawer attaches to the CEC by 12X attachment cables and includes a repeater card installed. The repeater card is designed to strengthen signal strength over the longer 12X cables used with the powered expansion rack (FC 6954) and nonpowered expansion rack (FC 6983).

The 12X I/O Drawer PCI-X (FC 5798) is equivalent to the FC 5797 drawer, but configured without a repeater card and can only be used in the primary system rack.

Unsupported: IBM i does not support SCSI drives in the FC 5797 drawer.

Figure 2-34 shows an internal diagram of the FC 5797 and FC 5798 internal I/O drawers.

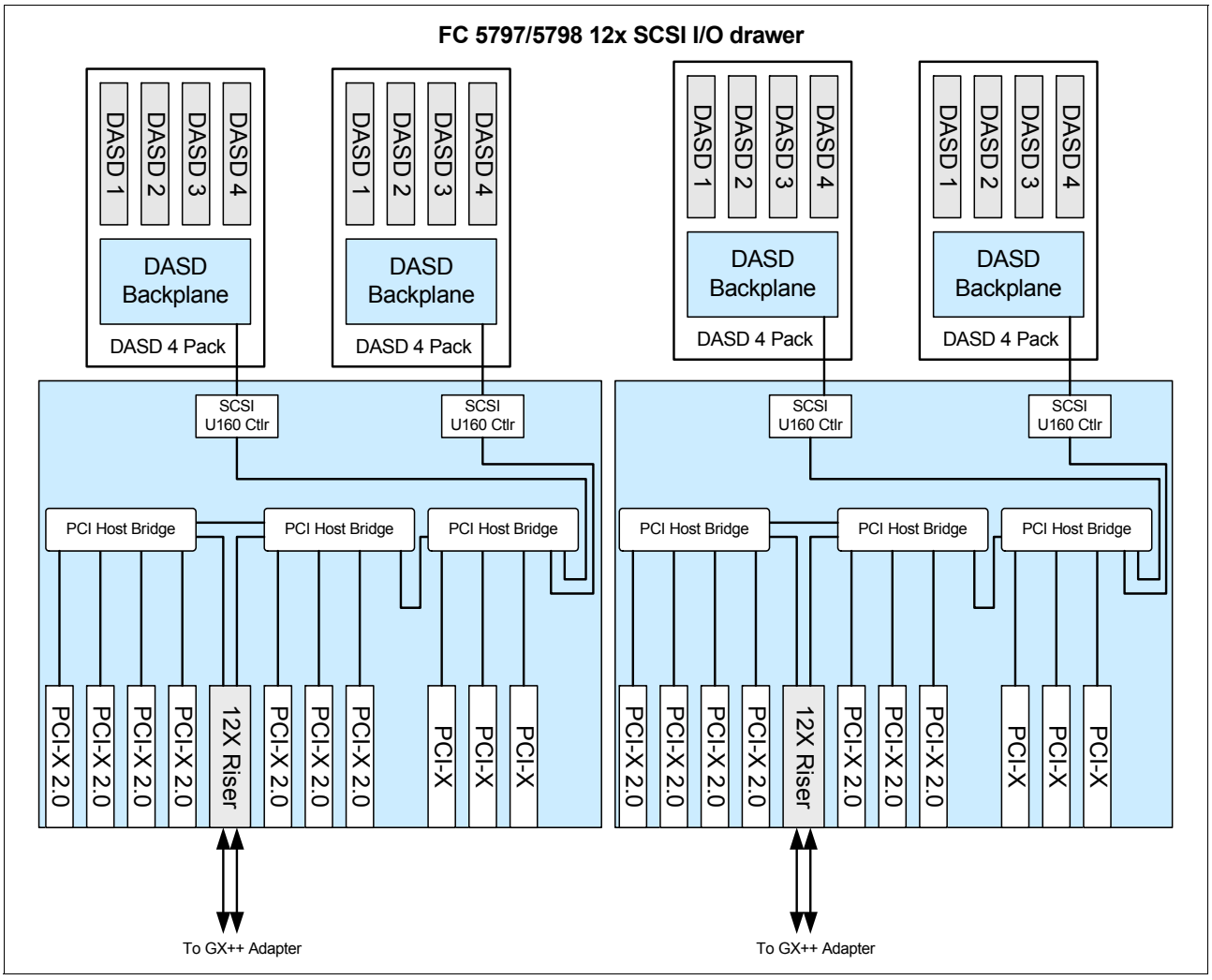

Figure 2-34 FC 5797 and FC 5798 internal I/O drawers

Consider additional I/O drawer configurations:

- Each half of the 12X I/O drawers are powered separately for redundancy.
- One single-wide, blind-swap cassette (equivalent to those in FC 4599 is provided in each PCIe 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, order FC 4599.
- For maximum throughout, use two GX++ adapters per adapter drawer (one GX++ adapter 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 2-20 details features of 12X-based internal I/O drawers.

Table 2-20 Internal I/O drawer feature comparison

Feature or function	FC 5803, FC 5873 drawers	FC 5797, FC 5798 drawers
Connection technology	12X	12X
Bandwidth per connection port (4 ports per drawer)	6.40 GBps 9.50 GBps	5 GBps sustained 6 GBps peak
PCI-X 2.0 (266 MHz) PCI-X (133 MHz) slots	N/A	14/6
PCIe slots	20	N/A
Ultra3 SCSI busses	N/A	N/A
SCSI disk bays	N/A	16
SAS SFF disk bays	26 (FC 5803), 0 (FC 5873)	N/A
Maximum drawers per system	32 (FC 5803) or 31 (FC 5873)	30 (FC 5797) 2 (FC 5798)

A maximum of 32 FC 5803 drawers or 31 FC 5873 I/O drawers can be connected to a Power 795 server. A maximum total of 30 FC 5797 plus FC 5798 I/O drawers can be connected to a 795 server, two of which can be FC 5798. The maximum sum of FC 5803 + FC 5797 + FC 5798 + FC 5873 is 32 that are configured in a Power 795 server.

Figure 2-35 shows an example of a FC 5803 I/O drawer installation sequence when the integrated battery feature (IBF) is not installed. If the IBF is installed, the battery backup units are located where an I/O drawer might have been located. Subsequent drawer numbering must be shifted by one.

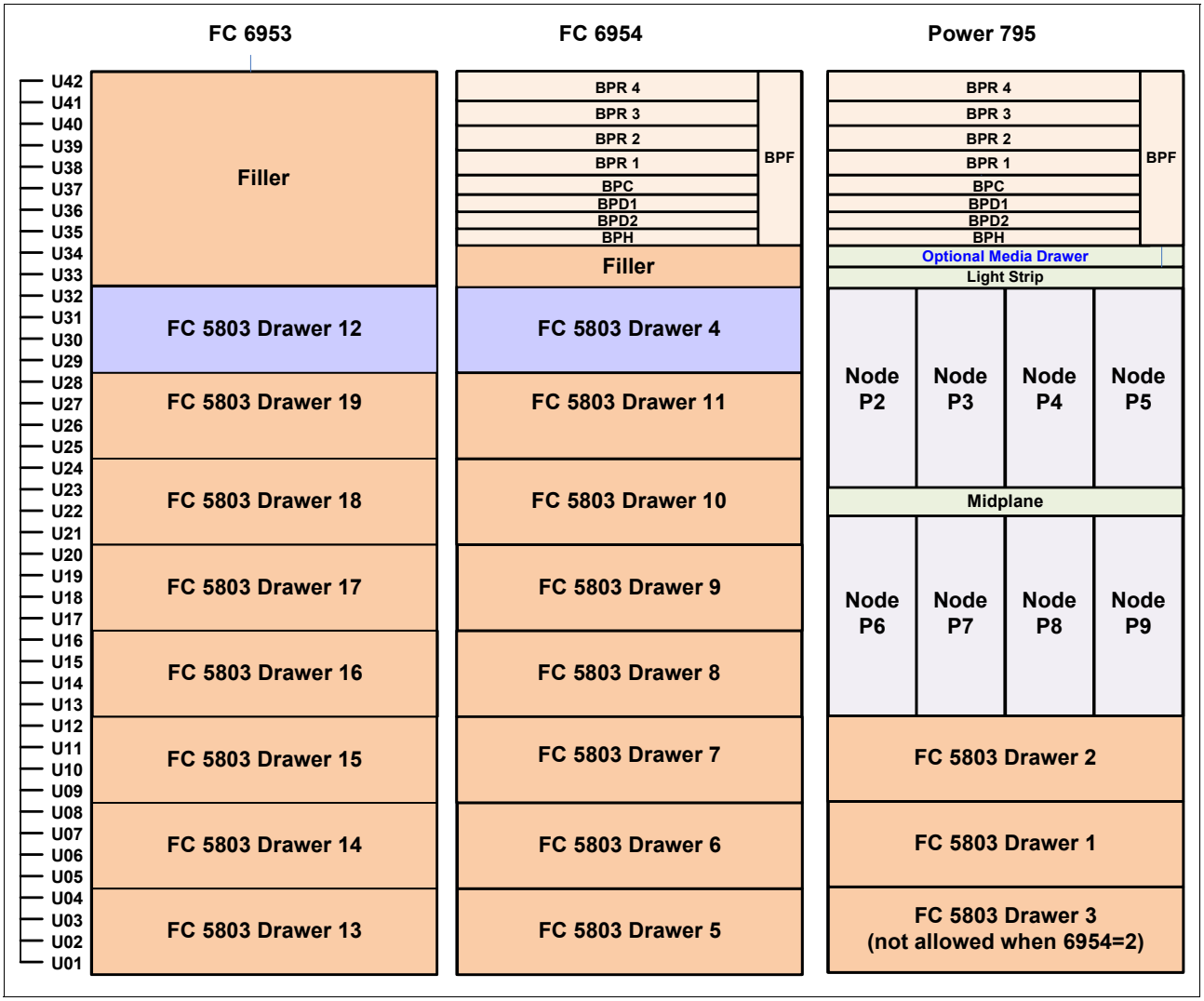

Figure 2-35 Power 795 I/O Expansion Drawer locations

2.9.4 Internal I/O drawer attachment

The internal I/O drawers are connected to the Power 795 server CEC using 12X technology. Drawer connections are made in loops to help protect against errors resulting from an open, missing, or disconnected cables. 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 12X I/O attachment adapter (GX++ adapter) 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 GX++ adapter attachment adapters can be installed in each processor book. Up to 32 12X I/O drawers are supported per Power 795 server.

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

- Single-loop (Figure 2-36 on page 97) mode connects an entire I/O drawer to the CEC using 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 GX Dual-Port 12X (FC 1816) per I/O drawer.
- Dual-loop (Figure 2-37 on page 98) 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 GX++ adapters (FC 1816) per connected I/O drawer. With a dual-loop configurations, the overall I/O bandwidth per drawer is higher.

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

Table 2-21 lists the number of single-looped and double-looped I/O drawers that can be connected to a 795 server based on the number of processor books installed.

Table 2-21 Number of I/O drawers that can be connected per processor book

Number of installed processor books	12X (FC 5803,FC 5873)		12X (FC 5797, FC 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	16	8	16	8
5	20	10	20	10
6	24	12	24	12
7	28	14	28	14
8	32	16	30	16

2.9.5 Single loop (full-drawer) cabling

Single loop I/O drawer connections are shown in Figure 2-36.

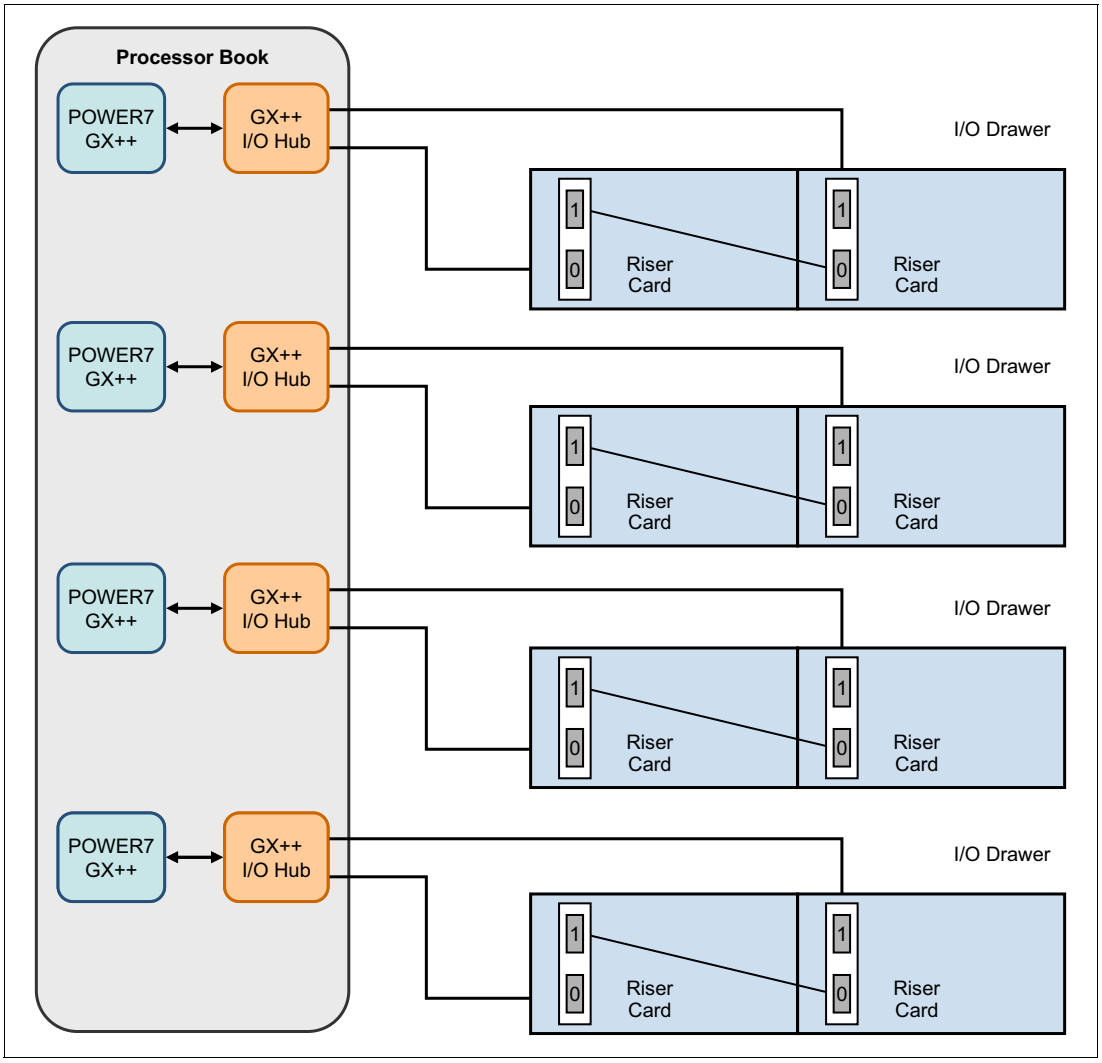

Figure 2-36 Single loop I/O drawer cabling

A 0.6m 12X DDR cable (FC 1861) must be ordered for each I/O drawer to enable single loop configuration.

2.9.6 Dual looped (half-drawer) cabling

Each of the two internal I/O drawer planars can be cabled and addressed by a GX++ adapter individually by using the preferred, dual loop (half drawer) cabling, shown in Figure 2-37.

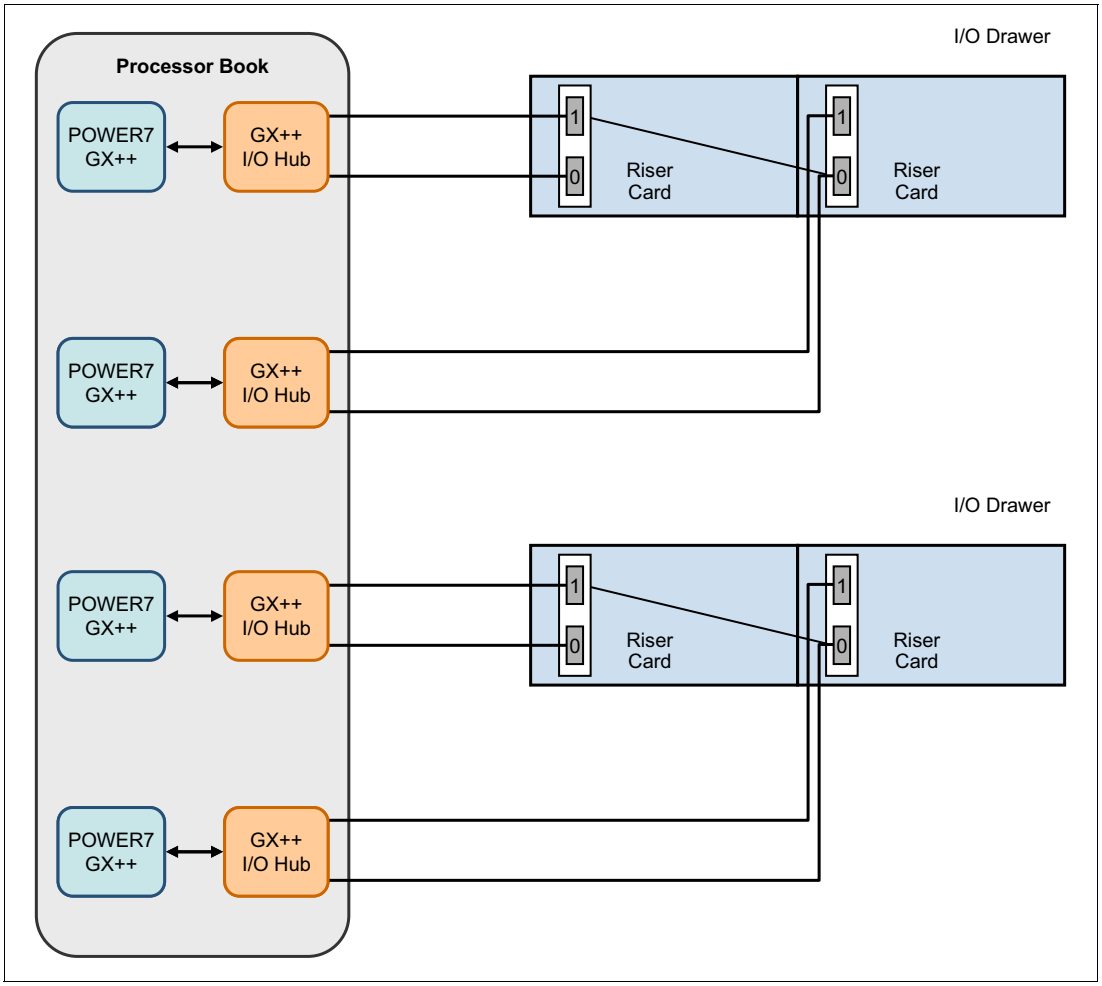

Figure 2-37 Dual loop I/O drawer cabling

2.10 External disk subsystems

The following external disk subsystems can be attached to the Power 795 server:

- EXP 12S SAS Expansion Drawer (FC 5886; supported, but no longer orderable).
- EXP24S SFF Gen2-bay Drawer for high-density storage (FC 5887)
- TotalStorage EXP24 Disk Drawer (FC 5786; no longer orderable)
- IBM 7031 TotalStorage EXP24 Ultra320 SCSI Expandable Storage Disk Enclosure; no longer orderable
- IBM System Storage

The next sections describe the supported external disk subsystems in more detail.

2.10.1 EXP 12S SAS Expansion Drawer

The EXP 12S (FC 5886) is an expansion drawer with twelve 3.5-inch form factor SAS bays. The FC 5886 supports up to 12 hot-swap SAS Hard Disk Drives (HDD) or up to 8 hot-swap solid-state drives (SSD). The EXP 12S includes redundant ac power supplies and two power cords. The drawer is one set of 12 drives which is run by one SAS controller, or one pair of SAS controllers. It has two SAS attachment ports and two Service Managers for redundancy. The EXP 12S takes up a 2 EIA space in a 19-inch rack. The SAS controller can be a SAS PCI-X or PCIe adapter or pair of adapters.

A maximum of 185 EXP 12S SAS Expansion Drawers can be used in the Power 795.

Attachment to the controller is through the appropriate external SAS cables. A pair of FC 5886 EXP12S drawers can be attached together with SAS cables allowing the SAS controllers to access 24 SAS bays.

For detailed information about the SAS cabling, see the serial-attached SCSI cable planning documentation:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7had/p7hadsascabling.htm

The drawer can be attached using one of the following adapters:

- PCIe 380 MB Cache Dual -x4 3 GB SAS RAID adapter (FC 5805)
- PCI-X DDR Dual -x4 SAS adapter (FC 5900)
- PCIe Dual -x4 SAS adapter (FC 5901 CCIN 57B3)
- PCIe 380 MB Cache Dual -x4 3 GB SAS RAID adapter (FC 5903)
- PCI-X DDR 1.5 GB Cache SAS RAID adapter (FC 5904)
- PCI-X DDR Dual -x4 SAS adapter (FC 5912 CCIN 572A)

With proper cabling and configuration, multiple wide ports are used to provide redundant paths to each dual port SAS disk. The adapter manages SAS path redundancy and path switching in case a SAS drive failure occurs. The SAS cables (Y) attach to an EXP 12S disk drawer.

Use SAS cable (YO) system to SAS enclosure, single controller/dual path 1.5M (FC 3450) or SAS cable (YO) system to SAS enclosure, single controller/dual path 3M (FC 3451) to attach SFF SAS drives in an EXP12S drawer.

Table 2-22 details the available SAS cable types.

Table 2-22 SAS Cables type

Cable type	Function
AA	This cable is used to connect between the top ports on two tri-port SAS adapters in a RAID configuration
AI	This cable is used to connect from a SAS adapter to internal SAS disk slots using an FC3650 or FC3651 cable card or FC3669 to the System External SAS port on your system.
AE	These cables are used to connect a SAS adapter to a media expansion drawer. These cables can also be used to connect two SAS adapters to a disk expansion drawer in a unique JBOD configuration.
AT	This cable is used with a PCIe 12X I/O drawer to connect from a PCIe SAS adapter to the internal SAS disk slots.

Cable type	Function
EE	This cable is used to connect one disk expansion drawer to another in a cascaded configuration. Disk expansion drawers can only be cascaded one level deep, and only in certain configurations.
YO	This cable is used to connect a SAS adapter to a disk expansion drawer. The cable must be routed along the right side of the rack frame (as viewed from the rear) when connecting to a disk expansion drawer.
YI	This cable is used to connect a system external SAS port to a disk expansion drawer. The cable must be routed along the right side of the rack frame (as viewed from the rear) when connecting to a disk expansion drawer.
X	This cable is used to connect two SAS adapters to a disk expansion drawer in a RAID configuration. The cable must be routed along the right side of the rack frame (as viewed from the rear) when connecting to a disk expansion drawer.

Various disk options are available to be installed in the EXP 12S drawer. Table 2-23 shows the available disk drive feature codes.

Table 2-23 Disk options for the EXP 12S drawer

Feature code	Description	Support		
		AIX	IBM i	Linux
3586	69 GB 3.5" SAS Solid State Drive	✓	—	✓
3587	69 GB 3.5” SAS Solid State Drive	—	✓	—
3646	73.4 GB 15K RPM SAS Disk Drive	✓	—	✓
3647 (CCIN 433C)	146.8 GB 15K RPM SAS Disk Drive	✓	—	✓
3648 (CCIN 433D)	300 GB 15K RPM SAS Disk Drive	✓	—	✓
1751	900 GB 10K RPM SAS SFF HD	✓	—	✓
1752	900 GB 10K RPM SAS HDD in Gen2-S Carrier	✓	—	✓
3649	450 GB 15K RPM SAS Disk Drive	✓	—	✓
3658 (CCIN 198E)	428.4 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	—	✓	—
1737 (CCIN 19A4)	856 GB 10K RPM SFF SAS Disk Drive	—	✓	—
1738 (CCIN 19B4)	856 GB 10K RPM HDD in Gen2-S Carrier	—	✓	—

A second EXP12S drawer can be attached to another drawer using two SAS EE cables, providing 24 SAS bays instead of 12 bays for the same SAS controller port. This technique is

called *cascading*. In this configuration, all 24 SAS bays are controlled by a single controller or a single pair of controllers.

For detailed information about the SAS cabling, see the serial-attached SCSI cable planning documentation:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7had/p7hadsascabling.htm

2.10.2 EXP24S SAS Expansion Drawer

The EXP24S (FC 5887) SFF Gen2-bay Drawer is an expansion drawer with twenty-four 2.5-inch small form factor (SFF) SAS bays. It supports up to 24 hot-swap SFF SAS Hard Disk drives (HDD) on POWER6 or POWER7 servers in 2U of 19-inch rack space.

Unsupported: SSD not currently supported

The SFF bays of the EXP24S are different from the SFF bays of the POWER7 system units or 12X PCIe I/O Drawers (FC 5802, FC 5803). The EXP24S uses Gen-2 or SFF-2 SAS drives that physically do not fit in the Gen-1 or SFF-1 bays of the POWER7 system unit or 12X PCIe I/O Drawers or vice versa.

The EXP24S SAS ports are attached to SAS controllers which can be a SAS PCI-X or PCIe adapter or pair of adapters. The EXP24S can also be attached to an imbedded SAS controller in a server with an imbedded SAS port. Attachment between the SAS controller and the EXP24S SAS ports is with the appropriate SAS Y or X cables.

The drawer can be attached with the following SAS adapters/controllers:

- PCI-X 1.5 GB Cache SAS RAID Adapter 3 Gb (FC 5904, FC 5906, FC 5908)
- PCIe 380 MB Cache SAS RAID Adapter 3 Gb (FC 5805, FC 5903)
- PCIe Dual-x4 SAS Adapter 3 Gb (FC 5901 CCIN 57B3)

A maximum of 168 EXP24S SAS Expansion Drawers can be used in the Power 795.

The EXP24S can be ordered in one of three possible manufacturing-configured mode settings (not customer set-up):

- With IBM AIX, Linux, Virtual I/O Server (VIOS):
  - Mode 4: four sets of 6 bays
  - Mode 2: two sets of 12 bays
  - Mode 1: one set of 24 bays
- With IBM i:
  - Mode 1: one set of 24 bays

Notes:

- FC 5887 Modes are set by IBM Manufacturing. No option exists to reset after IBM shipment is announced.
- If ordering multiple EXP24S, avoid mixing modes within that order. There is no externally visible indicator regarding the drawer's mode.
- SSDs are not currently supported in EXP24S.
- No cascading of FC 5887s; maximum of one FC 5887 per SAS connector.

Six SAS connectors are on the rear of the EXP24S to which SAS adapters/controllers are attached. They are labeled T1, T2, and T3, and there are two T1, two T2, and two T3, as shown in Figure 2-38.

- In mode 1, two or four of the six ports are used. Two T2 are used for a single SAS adapter, and two T2 and two T3 are used with a paired set of two adapters or dual adapters configuration.
- In mode 2 or mode 4, four ports are used, two T2 and two T3 to access all SAS bays.

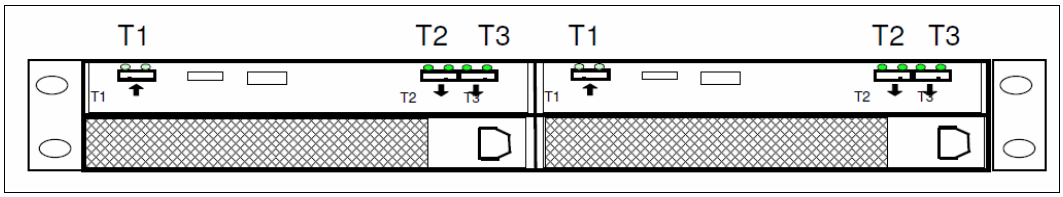

Figure 2-38 FC 5887 rear connectors

An EXP24S in mode 4 can be attached to two or four SAS controllers and provide a great deal of configuration flexibility. An EXP24S in mode 2 has similar flexibility. Up to 24 HDDs can be supported with any of the supported SAS adapters/controllers.

EXP24S no-charge specify codes should be included with EXP24S orders to indicate IBM Manufacturing the mode to which the drawer should be set and the adapter/controller/cable configuration which will be used. Table 2-24 lists the no-charge specify codes, the physical adapters/controllers, and cables with their own chargeable feature numbers.

Table 2-24 EXP24S Cabling

Feature code	Mode	Adapter/controller	Cable to drawer	Environment
9359	1	One FC 5901	1 YO cable	AIX, Linux, VIOS
9360	1	Pair FC 5901	2 YO cables	AIX, Linux, VIOS
9361	2	Two FC 5901	2 YO cables	AIX, Linux, VIOS
9365	4	Four FC 5901	2 X cables	AIX, Linux, VIOS
9366	2	Two pair FC 5901	2 X cables	AIX, Linux, VIOS
9367	1	Pair FC 5903/FC 5805	2 YO cables	AIX, IBM i, Linux, VIOS
9368	2	Two FC 5903/FC 5805	2 X cables	AIX, Linux, VIOS
9382	1	One FC 5904/06/08	1 YO cable	AIX, IBM i, Linux, VIOS
9383	1	Pair FC 5904/06/08	2 YO cables	AIX, IBM i, Linux, VIOS

Cable to EXP24S Drawer options:

- X cables for FC 5901, FC 5903, FC 5805, FC 5904, FC 5906, and FC 5908 (3 Gb): 3 m (FC 3661), 6 m (FC 3662), 15 m (FC 3663)
- YO cables for FC 5901, FC 5903, FC 5805, FC 5904, FC 5906, FC 5908 (3 Gb): 1.5 m (FC 3691), 3 m (FC 3692), 6 m (FC 3693), 15 m (FC 3694)

Rails: EXP24S drawer rails are fixed length and designed to fit Power Systems provided racks of 28 inches (8xx mm) deep. EXP24S uses two EIA of space in a 19-inch wide rack. Other racks might have different depths and these rails do not adjust. No adjustable depth rails are orderable at this time

For details about SAS cabling, see the serial-attached SCSI cable planning documentation:

http://ibm.co/UQEGdh

2.10.3 TotalStorage EXP24 Disk Drawer

The TotalStorage EXP24 Disk Drawer (FC 5786) provides 24 disk bays. The FC 5876 requires 4U of mounting space in a 19-inch rack and features redundant power, redundant cooling. The DASD drawer 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 (FC 5741/FC 5742) that are connected to an Ultra 320 SCSI adapters. Up to four repeater cards are supported per FC 5786 unit.

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

Up to 110 TotalStorage EXP24 Disk Drawers can be attached to the Power 795 server.

Ordering notes: New TotalStorage EXP24 Disk Drawers cannot be ordered for the Power 795, and thus only existing FC 5786 drawers can be moved to the Power 795.

The TotalStorage EXP24 Disk Drawer is supported in initial orders only with IBM i as the primary operating system.

2.10.4 IBM TotalStorage EXP24

The IBM 7031 TotalStorage EXP24 Ultra320 SCSI Expandable Storage Disk Enclosure supports up to 24 Ultra320 SCSI Disk Drives arranged in four independent SCSI groups of up to six drives or in two groups of up to twelve drives. Each SCSI drive group can be connected by either a Single Bus Ultra320 SCSI Repeater Card or a Dual Bus Ultra320 SCSI Repeater Card allowing a maximum of eight SCSI connections per TotalStorage EXP24.

The IBM 7031 Model D24 (7031-D24) is an expandable disk storage enclosure that is a horizontal 4 EIA by 19-inch rack drawer for mounting in equipment racks.

The IBM 7031 Model T24 (7031-T24) is an expandable disk storage enclosure that is a vertical tower for floor-standing applications.

Ordering notes: A new IBM 7031 TotalStorage EXP24 Ultra320 SCSI Expandable Storage Disk Enclosure cannot be ordered for the Power 795, and thus only existing 7031-D24 drawers or 7031-T24 towers can be moved to the Power 795.

AIX and Linux partitions are supported along with the usage of a IBM 7031 TotalStorage EXP24 Ultra320 SCSI Expandable Storage Disk Enclosure.

2.10.5 IBM System Storage

The IBM System Storage Disk Systems products and offerings provide compelling storage solutions with superior value for all levels of business, from entry-level up to high-end storage systems.

IBM System Storage N series

The IBM System Storage N series is a network-attached storage (NAS) solution and provides the latest technology to customers to help them improve performance, virtualization manageability, and system efficiency at a reduced total cost of ownership. For more information about the IBM System Storage N series hardware and software, go to the following location:

http://www.ibm.com/systems/storage/network

IBM System Storage DS3500

The IBM System Storage DS3500 is an entry-level storage system designed to meet the availability and consolidation needs for a wide range of users. New features, including larger capacity 1 TB SAS drives, increased data protection features such as RAID 6, encryption, and more FlashCopies per volume, provide a reliable virtualization platform. For more information about DS3000, go to the following location:

http://www.ibm.com/systems/storage/disk/ds3500/index.html

IBM System Storage DS5000

DS5000 enhancements help reduce cost by introducing SSD drives. Also, with the EXP5060 expansion unit supporting up to 1.34 PB in a 4U package, customers can see up to a 1/3 reduction in floor space over standard enclosures. With the addition of 1 Gbps iSCSI host attach, customers can reduce cost for their less demanding applications while continuing to provide high performance where necessary utilizing the 8 Gbps Fibre Channel host ports. With the DS5000 family, you get consistent performance from a smarter design that simplifies your infrastructure, improves your total cost of ownership (TCO), and reduces your cost. For more information about DS5000 series, see the following location:

http://www.ibm.com/systems/storage/disk/ds5000/index.html

IBM Storwize V7000 Midrange Disk System

IBM Storwize® V7000 is a virtualized storage system to complement virtualized server environments that provides unmatched performance, availability, advanced functions, and highly scalable capacity never seen before in midrange disk systems. Storwize V7000 is a powerful midrange disk system that was designed to be easy to use and enable rapid deployment without additional resources. Storwize V7000 is virtual storage that offers greater efficiency and flexibility through built-in solid-state drive (SSD) optimization and thin provisioning technologies. Storwize V7000 advanced functions also enable nondisruptive migration of data from existing storage, simplifying implementation and minimizing disruption to users. Storwize V7000 also enables you to virtualize and reuse existing disk systems, supporting a greater potential return on investment (ROI). For more information about Storwize V7000, go to the following location:

http://www.ibm.com/systems/storage/disk/storwize_v7000/index.html

IBM XIV Storage System

IBM offers a high-speed configuration of its self-optimizing, self-healing, resilient disk solution, the IBM XIV® Storage System; storage reinvented for a new era. Now, organizations with mid-size or high-size capacity requirements can take advantage of the latest IBM technology

for their most demanding applications, up to 243 TB of usable capacity and incremental upgrades. For more information about XIV Storage System, go to the following location:

http://www.ibm.com/systems/storage/disk/xiv/index.html

IBM System Storage DS8000

The IBM System Storage DS8000® series is designed to offer high availability, multiplatform support, and simplified management tools. With its high capacity, scalability, broad server support, and virtualization features, the DS8000 family is well suited for simplifying the storage environment by consolidating data from multiple storage systems on a single system.

The high-end model DS8800 is the most advanced model in the IBM DS8000 lineup and introduces new dual IBM POWER6 processor-based controllers that usher in a new level of performance for the company's flagship enterprise disk platform. The DS8800 offers a huge maximum physical storage capacity. For more information about the DS8000 series, go to the following location:

http://www.ibm.com/systems/storage/disk/ds8000/index.html

2.11 PCI adapter support

Most PCI, PCI-X, and PCIe adapters for the system are capable of being hot-plugged. Any PCI adapter supporting a boot device or system console must not be hot-plugged. The following adapters are not hot-plug capable:

- POWER GXT145 PCI Express Graphics Accelerator (FC 5748)
- POWER GXT135P Graphics Accelerator with Digital Support (FC 2849, supported)
- The 2-Port Multiprotocol PCI Adapter (FC 2962, supported)

The maximum number of a specific PCI, PCI-X, or PCIe adapters allowed per Power 795 server can be less than the number allowed per I/O drawer multiplied by the maximum number of I/O drawers.

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 POWER7 configuration, including the system unit and I/O enclosures attached to loops, see the *Power Systems PCI Adapter Placement Guide*:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/areab/areabkickoff.htm

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

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

If you are installing a new feature, ensure that you have the software that is required to support the new feature, and determine whether there are any existing program temporary fix (PTF) prerequisites to install. To do this, use the IBM Prerequisite website at:

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

In the feature summary tables of this section, note that over time, additional features might be withdrawn from marketing. This information is typically included in announcement letters. To

ensure that you have the latest information about ordering a feature, go to the Support for IBM Systems website:

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 Planning tab, as in the following sample steps:

1. Select **Power**.
2. Select **9119-FHB** from the **Hardware** list.
3. Click **Continue**.
4. Select the **Plan and Install** tab.

Alternatively, for example, for IBM System i® models, go to the following website:

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

2.11.1 LAN adapters

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

Table 2-25 Available LAN adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
5700[a]	Gigabit Ethernet-SX PCI-X Adapter	Short	512	✓	✓	✓
5706	2-Port 10/100/1000 Base-TX Ethernet PCI-X Adapter	Short	512	✓	✓	✓
5717	4-Port 10/100/1000 Base-TX PCI Express Adapter	Short	256	✓	—	✓
5732	10 Gigabit Ethernet-CX4 PCI Express Adapter	Short	256	✓	—	✓
5767	2-Port 10/100/1000 Base-TX Ethernet PCI Express Adapter	Short	512	✓	✓	✓
5768	2-Port Gigabit Ethernet-SX PCI Express Adapter	Short	512	✓	✓	✓
5769	10 Gigabit Ethernet-SR PCI Express Adapter	Short	256	✓	—	✓
5772	10 Gigabit Ethernet-LR PCI Express Adapter	Short	256	✓	✓	✓
5899	4-Port 1 Gb Ethernet PCIe 4x Adapter	Short	24	✓	✓	✓
5701[a]	10/100/1000 Base-TX Ethernet PCI-X Adapter	Short	512	✓	✓	✓
5721[a]	10 Gigabit Ethernet-SR PCI-X (Fiber)	Short	256	✓	✓	✓
5722[a]	10 Gigabit Ethernet-LR PCI-X (Fiber)	Short	256	✓	✓	✓
5740	4-port 10/100/1000 Gigabit Ethernet PCI-X	Short	256	✓	✓	✓
EC28	2-Port 10Gb Ethernet/RoCE SFP+ Adapter	Long	5	✓	—	✓
EC30	2-Port 10Gb Ethernet/RoCE SFP+ Adapter with Optics	Short	5	✓	—	✓
EN22	GX++ 2-Port 10Gb Ethernet/FCoE (SR Fibre) Adapter	GX++	24	✓	—	—
EN23	GX++ 2-Port 16Gb Ethernet/FCoE (LC Fibre) Adapter	GX++	24	✓	—	—

a. Supported, but no longer orderable

Support: AIX Network Installation Manager (NIM) boot capability is supported with adapter FC 5700, FC 5701, FC 5706, FC 5707, FC 5717, FC 5767, and FC 5768.

2.11.2 SCSI adapters

Table 2-26 lists the SCSI adapters that are available for the server.

Table 2-26 Available SCSI adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
1912[a]	PCI-X DDR Dual Channel Ultra320 SCSI Adapter	Short	180	✓	✓	✓
5736	PCI-X Dual Channel Ultra320 SCSI Adapter	Short	180	✓	✓	✓
5780[a]	PCI-X EXP24 Ctl-1.5 GB No IOP	Long	180	—	✓	—

a. Supported, but no longer orderable

2.11.3 iSCSI

Internet SCSI (iSCSI) is an open, standards-based approach by which SCSI information is encapsulated with 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, go to the following location:

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

Gigabit Ethernet

Although iSCSI can, by design, be 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 Power 795 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 2-27 on page 108 lists the iSCSI adapters that are available for the Power 795 server.

Table 2-27 Available iSCSI adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
FC 5713	Gigabit iSCSI TOE PCI-X on copper media adapter	Short	160	✓	✓	✓
FC 5714[a]	Gigabit iSCSI TOE PCI-X on optical media adapter	Short	160	✓	✓	✓

a. Supported, but no longer orderable

IBM iSCSI software Host Support Kit

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

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

The iSCSI Host Support Kit on IBM 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.

2.11.4 SAS adapters

Serial Attached SCSI (SAS) is an 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 2-28 lists the SAS adapters that are available for the Power 795 server.

Table 2-28 Available SAS adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
5805	PCIe 380MB Cache Dual - x4 3 Gb SAS RAID Adapter	Short	240	✓	✓	✓
5900[a]	PCI-X DDR Dual - x4 SAS Adapter	Short	180	✓	—	✓
5901	PCIe Dual - x4 SAS Adapter	Short	240	✓	✓	✓
5902[ab]	PCI-X DDR Dual - x4 3 Gb SAS RAID Adapter	Long	180	✓	—	✓
5903[ab]	PCIe 380 MB Cache Dual - x4 3 Gb SAS RAID Adapter	Short	240	✓	✓	✓
5906	PCI-X DDR 1.5 GB Cache SAS RAID Adapter (BSC)	Long	180	—	✓	—
5912	PCI-X DDR Dual - x4 SAS Adapter	Short	192	✓	✓	✓
5913	PCIe 6 Gbps SAS Controller Adapter	Short	18	✓	✓	✓
ESA1	PCIe 6 Gbps Cacheless SAS Controller Adapter	Short	20	✓	✓	✓

a. Supported, but is no longer orderable.
b. The SAS RAID adapter must be installed in pairs; it cannot be used to drive. tape, and DVD media devices.

2.11.5 Fibre Channel adapters

The Power 795 server supports direct attachment, or SAN connection, to devices using Fibre Channel adapters. 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.

Table 2-29 summarizes the Fibre Channel adapters that are available for the Power 795 server.

Table 2-29 Available Fibre Channel adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
5716[a]	2 Gigabit Fibre Channel PCI-X Adapter	Short	192	✓	—	✓
5735[b]	8 Gigabit PCI Express Dual Port Fibre Channel Adapter	Short	256	✓	✓	✓
5749	4 Gbps Fibre Channel (2-port)	Short	480	—	✓	—
5758[a]	4 Gigabit single-port Fibre Channel PCI-X 2.0 Adapter (LC)	Short	480	✓	—	✓
5759	4 Gigabit dual-port Fibre Channel PCI-X 2.0 Adapter (LC)	Short	480	✓	—	✓

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
5773[a]	4 Gigabit PCI Express Single Port Fibre Channel Adapter	Short	512	✓	—	✓
5774	4 Gigabit PCI Express Dual Port Fibre Channel Adapter	Short	512	✓	✓	✓
EN23	GX++ 2-Port 16Gb FC Adapter	GX++	24	✓	—	—

a. Supported, but no longer orderable
b. N_Port ID Virtualization (NPIV) capability is supported through the Virtual I/O Server.

2.11.6 Fibre Channel over Ethernet (FCoE)

FCoE allows for the convergence of Fibre Channel and Ethernet traffic onto a single adapter and converged fabric.

Figure 2-39 compares the existing FC and network connection and the FCoE connection.

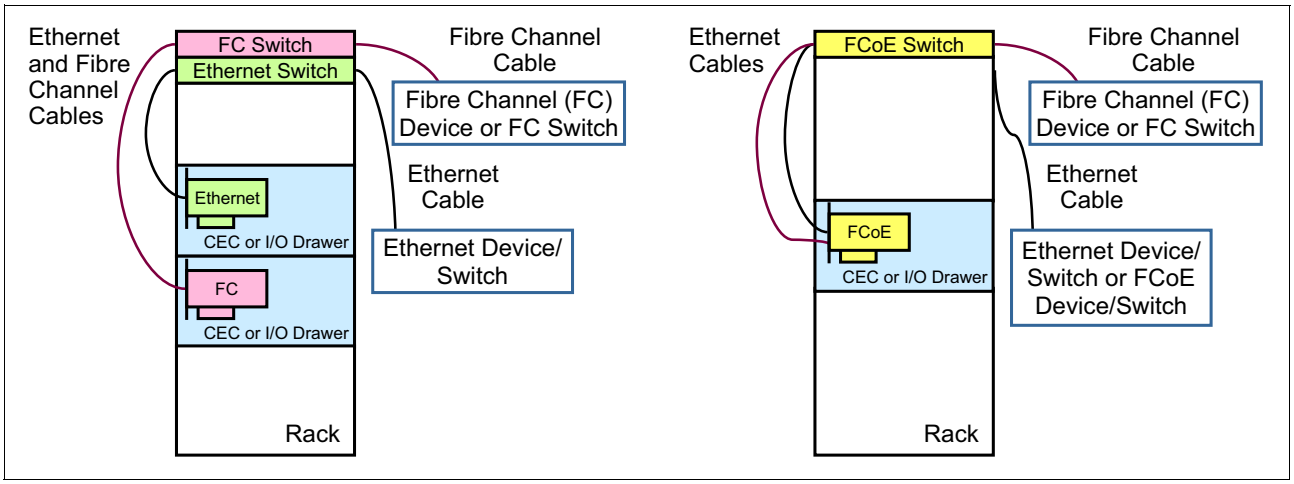

Figure 2-39 Comparison between existing FC and network connection and FCoE connection

Table 2-30 lists the available Fibre Channel over Ethernet Adapter. It is a high performance, converged network adapter (CNA) using SR optics. Each port can provide network interface card (NIC) traffic and Fibre Channel functions simultaneously.

Table 2-30 Available FCoE adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
5708	10Gb FCoE PCIe Dual Port Adapter	Short	256	✓	—	✓

2.11.7 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 with 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 2-31 lists the asynchronous, WAN, and modem adapters that are available for the server.

Table 2-31 Available Asynchronous, WAN, and modem adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
2943[a]	8-Port Asynchronous Adapter EIA-232/RS-422	Short	18	✓	—	—
5289	2-Port Asynchronous Adapter EIA-232 and 2-Port RJ45	Short	12	✓	✓	✓
5723[a]	2-Port Asynchronous EIA-232 PCI Adapter	Short	18	✓	—	✓
5785	4 Port Async EIA-232 PCIe Adapter	Short	18	✓	—	✓
6805[a]	PCI 2-Line WAN IOA No IOP	Short	199	—	✓	✓
6808[a]	PCI 4-Modem WAN IOA No IOP	Long	155	—	✓	—
6833[a]	PCI 2-Line WAN with Modem IOA No IOP	Short	239	—	✓	—
2893	PCI 2-Line WAN with Modem	Long	24	✓	✓	✓
2894	PCI 2-Line WAN with Modem (CIM)	Long	24	✓	✓	✓

a. Supported, but no longer orderable

2.11.8 Cryptographic coprocessor

The cryptographic coprocessor cards provide both cryptographic coprocessor and cryptographic accelerator functions in a single card.

Table 2-32 lists the cryptographic adapters.

Table 2-32 Available cryptographic adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
4764[a]	PCI-X Cryptographic Coprocessor (FIPS 4)	Short	32	✓	✓	—
4808	PCIe Crypto Coprocessor Gen3 BSC 4765-001	Short	10	✓	✓	—

a. Supported, but no longer orderable

2.11.9 GX adapter

The FC 816, GX Dual-port 12X Host Channel Adapter, provides two 12X connections for dual channel applications. Connection to supported InfiniBand switches is accomplished by using the 12X to 4X Channel Conversion Cable FC 1842 for model conversions, or the FC 1854 cable available at the time of order. The connection using the FC 1842 cable is limited to the base ports of the InfiniBand switches.

Additional information about this adapter is provided in Table 2-33 on page 112.

Table 2-33 Available GX adapter

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
1816	GX Dual-port 12X HCA	Short	32	✓	✓	✓

2.11.10 USB and graphics adapters

The PCI-Express 4 port USB adapter (FC 2728) provides support for USB devices. The 2-Port USB PCI adapter (FC 2738) is available for the connection of a keyboard and a mouse. The POWER GXT135P (FC 2849) and the POWER GXT145 (FC 5748) are 2-D graphics adapters that provide support for analog and digital monitors.

Table 2-34 lists the available USB and graphics adapters.

Table 2-34 USB and graphics adapters

Feature code	Adapter description	Size	Maximum	Support		
				AIX	IBM i	Linux
2728	4 port USB PCIe Adapter	Short	8	✓	—	✓
2738[a]	2-Port USB PCI Adapter	Short	16	✓	—	✓
2849[a]	POWER GXT135P Graphics Accelerator with Digital Support	Short	8	✓	—	✓
5748	POWER GXT145 PCI Express Graphics Accelerator	Short	8	✓	—	✓

a. Supported, but no longer orderable

2.12 Hardware Management Console (HMC)

Each Power 795 server must be connected to an HMC for system control, partitioning, Capacity on Demand, and service functions. Preferably, ensure that each Power 795 server is connected to two HMCs for redundancy. For more details, see the following sections.

The HMC is a dedicated workstation that provides a graphical user interface for configuring, operating, and performing basic system tasks for the POWER7, and also the previous POWER5, POWER5+, POWER6 and POWER6+ processor-based systems. Either non-partitioned, partitioned, or clustered environments are supported. In addition, the HMC is used to configure and manage logical partitions (LPARs). One HMC is capable of controlling

multiple POWER5, POWER5+, POWER6, POWER6+, and POWER7 processor-based systems.

At the time of writing, one HMC supports up to 1024 LPARs using HMC machine code Version 7 Release 720 or later, and a maximum of 32 Power 795 servers are supported. An HMC also supports up to 48 Power 710, Power 720, Power 730, Power 740, Power 750, Power 755, Power 770, or Power 780.

For updates of the machine code and also HMC functions and hardware prerequisites, see the following website:

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

2.12.1 HMC functional overview

The HMC provides three groups of tasks:

- Server management
- Virtualization management
- HMC management

Server management

The first group contains all tasks related to the management of the physical servers under the control of the HMC.

- System password
- Power on/off
- Capacity on Demand
- Energy Management Policies
- Power Trending and Capping
- Error management
- System indicators
- Error/event collection reporting
- Dump collection reporting
- Call home
- Customer notification
- Hardware replacement (guided repair)
- SNMP events
- Concurrent add/repair
- Redundant service processor
- Firmware updates

Virtualization management

The second group contains all tasks related to virtualization features such as the Logical Partition configuration or dynamic reconfiguration of resources.

- System plans
- System profiles
- Logical partitions (create, activate, shutdown)
- Profiles
- Partition Mobility
- Dynamic LPAR operations (add/remove/mode processors, memory, I/O, and so on)
- Custom groups

HMC management

The last group relates to the management of the HMC itself, its maintenance, security, or configuration for example:

- Set-up wizard
- User management
  - User IDs
  - Authorization levels
  - Customizable authorization
- Disconnect/reconnect
- Network security
  - Remote operation enable/disable
  - User definable SSL certificates
- Console logging
- HMC redundancy
- Scheduled operations
- Back up and restore
- Updates, upgrades
- Customizable message of the day

The HMC version V7R720 adds support for Power 710, Power 720, Power 730, Power 740, and Power 795 servers.

The HMC provides both graphical and command-line interfaces for all management tasks. Running HMC Version 7, a remote connection to the HMC through a web browser (previous versions required a special client program, called WebSM) or SSH are possible. The command-line interface is also available by using the SSH secure shell connection to the HMC. It can be used by an external management system or a partition to perform HMC operations remotely.

2.12.2 HMC connectivity to the Power 795

The Power 795 server requires one primary HMC that can communicate to all bulk power hubs (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.

For improved system availability, a redundant HMC is highly preferable. 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 establish the connections between the the HMC and the Power 795 server.

Figure 2-40 shows the connection between two HMCs to the Power 795 server.

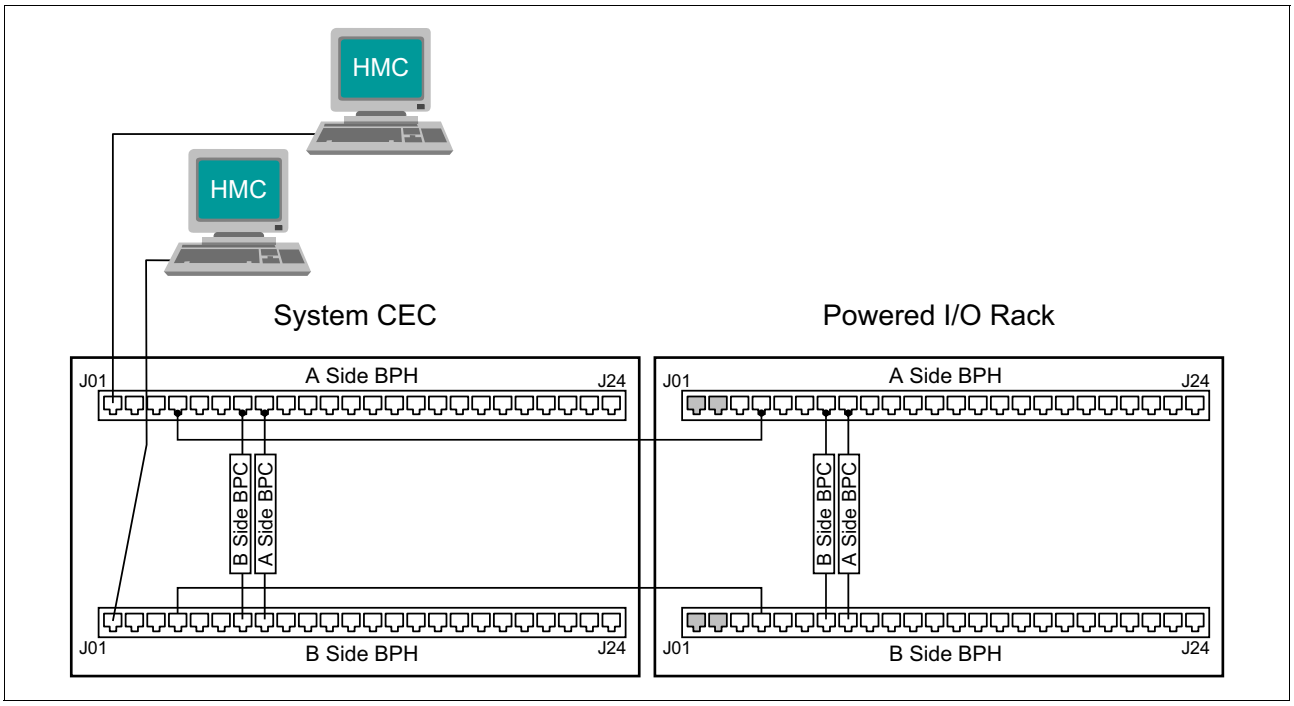

Figure 2-40 HMC connection to the bulk power hub of the Power 795

A minimum of two Ethernet ports are needed on the HMC to provide connectivity to the Power 795. The rack-mounted 7042-CR5 and 7042-CR6 HMC default configuration provides four Ethernet ports. The deskside 7042-C07 and 7042-C08 HMC standard configuration offers one Ethernet port. Preferably, order an optional PCIe adapter to provide additional Ethernet ports.

The default mechanism for allocation of the IP addresses for the Power 795 server HMC ports is dynamic. The HMC can be configured as a DHCP server, providing the dynamic IP address at the time the managed server is powered on. In this case, the flexible service processor (FSP) are allocated IP address from a set of address ranges predefined in the HMC software.

Service processor: The service processor is used to monitor and manage the system hardware resources and devices. The service processor offers two Ethernet 10/100 Mbps ports. Both Ethernet ports are visible only to the service processor and can be used to attach the server to an HMC or to access the Advanced System Management Interface (ASMI) options from a client web browser, using the HTTP server integrated into the service processor internal operating system.

For more information, see “Service processor” on page 192.

2.12.3 HMC high availability

The HMC is an important hardware component. After being put into operation, POWER7 processor-based servers and their hosted partitions can continue to operate when no HMC is available. However, in such conditions, some operations cannot be performed, such as a dynamic LPAR reconfiguration, a partition migration using PowerVM Live Partition Mobility, or the creation of a new partition. Therefore, you might decide to install two HMCs in a

redundant configuration so that one HMC is always operational, even when performing maintenance of the other one, for example.

If you want a redundant HMC function, the servers can be attached to two separate HMCs to address availability requirements. Both HMCs must have the same level of the Hardware Management Console Licensed Machine Code Version 7 (FC 0962) to manage POWER7 processor-based servers or an environment with a mixture of POWER5, POWER5+, POWER6, POWER6+, and POWER7 processor-based servers. The HMCs provide a locking mechanism so that only one HMC at a time has write access to the service processor. Depending on your environment, you have multiple options to configure the network.

Figure 2-41 shows a high available HMC configuration managing two servers. Each HMC is connected to one FSP port of all managed servers.

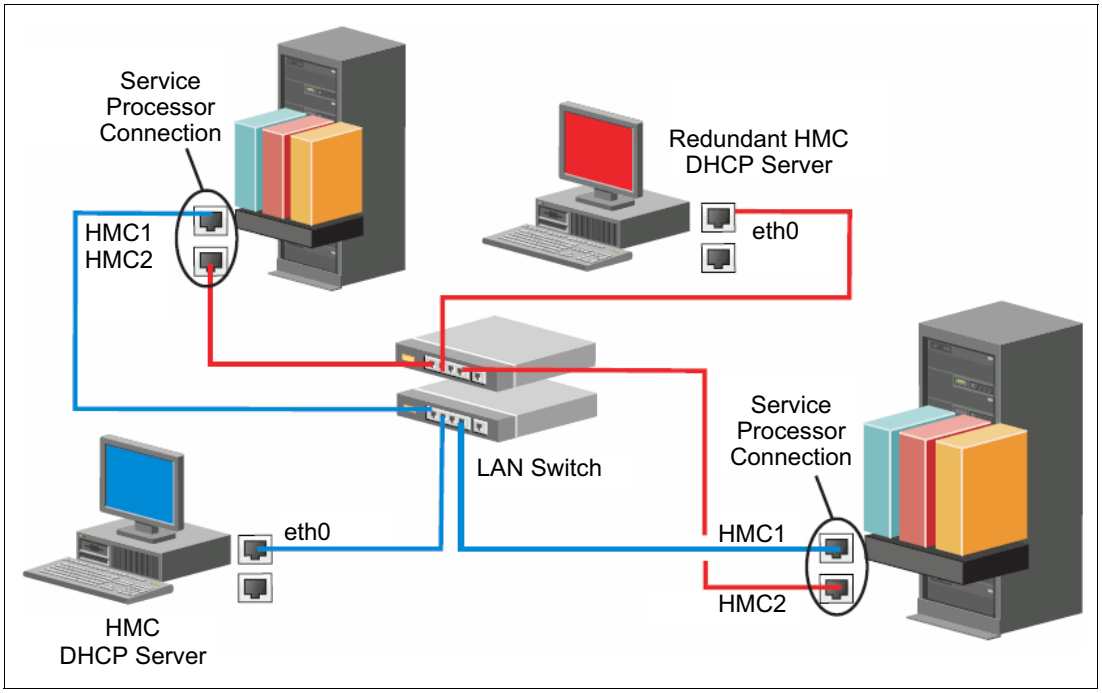

Figure 2-41 High availability HMC architecture

Both HMCs must be on a separate VLAN, to protect from a network failure. Each HMC can be DHCP server for its VLAN.

In a configuration with multiple systems or HMCs, the customer is required to provide switches or hubs to connect each HMC to the Power 795 FSP Ethernet ports. This approach provides redundancy both for the HMCs and the service processors.

For more details about redundant HMCs, see the *Hardware Management Console V7 Handbook*, SG24-7491.

2.12.4 HMC code level

The HMC code must be at level V7R720 to support the Power 795 systems.

In a dual HMC configuration, both HMCs must be at the same HMC version and release as each other.

Tips:

- When upgrading the code of a HMC in a dual HMC configuration, a good practice is to disconnect one of the HMCs and avoid having both HMCs connected to the same server with different code levels. If no profile or partition changes take place during the upgrade, both HMCs can stay connected. If the HMCs are at different code levels and a profile change is performed from the HMC at the higher code level, the format of the data stored in the server can change, and the HMC at the lower code level might go into recovery state if it does not understand the new data format.
- There are compatibility rules between the various software products that are executing within a POWER7 processor-based server environment: HMC, Virtual I/O Server, system firmware, or partition operating systems. To check what combinations are supported, and to identify required upgrades, you can use the Fix Level Recommendation Tool web page:

  http://www14.software.ibm.com/webapp/set2/flrt/home

The following two rules that are related to HMC code level are important when using PowerVM Live Partition Mobility:

- To use PowerVM Live Partition Mobility between a POWER6 processor-based server and a POWER7 processor-based server, if the source server is managed by one HMC and the destination server is managed by a different HMC, ensure that the HMC managing the POWER6 processor-based server is at version 7, release 3.5 or later, and the HMC managing the POWER7 processor-based server is at version 7, release 7.1 or later.
- To use PowerVM Live Partition Mobility for a partition configured for Active Memory Expansion, ensure that the HMC that manages the destination server is at version 7, release 7.1 or later.

2.13 Operating system support

The IBM POWER7 processor-based systems support three families of operating systems:

- AIX
- IBM i
- Linux

In addition, the Virtual I/O Server can be installed in special partitions that provide support to the other operating systems for using features such as virtualized I/O devices, PowerVM Live Partition Mobility, or PowerVM Active Memory Sharing.

For details about the software that is available on IBM POWER servers, visit the IBM Power Systems Software™ site:

http://www.ibm.com/systems/power/software/index.html

2.13.1 Virtual I/O Server (VIOS)

The minimum required level of Virtual I/O Server software depends on the server model:

Power 710 and 730	Virtual I/O Server Version 2.2, or later
Power 720 and 740	Virtual I/O Server Version 2.2, or later
Power 750	Virtual I/O Server Version 2.1.2.11 with Fix Pack 22.1 and Service Pack 1
Power 755	The Virtual I/O Server feature is not available on this model.
Power 770 and 780	Virtual I/O Server Version 2.1.2.12 with Fix Pack 22.1 and Service Pack 2
Power 795	Virtual I/O Server Version 2.2, or later

IBM regularly updates the Virtual I/O Server code. To find information about the latest updates, visit the Virtual I/O Server site:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/home.html

2.13.2 IBM AIX operating system

The following sections discuss the support for the various levels of AIX operating systems.

IBM periodically releases maintenance packages (service packs or technology levels) for the AIX operating system. Information about these packages, downloading, and obtaining the CD-ROM is on the Fix Central website. The Fix Central website also provides information about ho to obtain the fixes included on CD-ROM.

http://www-933.ibm.com/support/fixcentral/

The Service Update Management Assistant, which can help you 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, go to following website:

http://www14.software.ibm.com/webapp/set2/sas/f/genunix/suma.html

IBM AIX Version 5.3

IBM AIX Version 5.3 is supported on all models of POWER7 processor-based servers delivered in 2010.

The following minimum levels of AIX Version 5.3 support Power 710, 720, 730, and 740:

- AIX 5.3 with the 5300-10 Technology Level and Service Pack 5, or later
- AIX 5.3 with the 5300-11 Technology Level and Service Pack 5, or later
- AIX 5.3 with the 5300-12 Technology Level and Service Pack 2, or later

The following minimum levels of AIX Version 5.3 support Power 750, 755, 770, and 780:

- AIX 5.3 with the 5300-09 Technology Level and Service Pack 7, or later
- AIX 5.3 with the 5300-10 Technology Level and Service Pack 4, or later
- AIX 5.3 with the 5300-11 Technology Level and Service Pack 2, or later

The following minimum levels of AIX Version 5.3 to support Power 795:

- AIX 5.3 with the 5300-10 Technology Level and Service Pack 5, or later
- AIX 5.3 with the 5300-11 Technology Level and Service Pack 5, or later
- AIX 5.3 with the 5300-12 Technology Level and Service Pack 1, or later

A partition that uses AIX Version 5.3 will execute in POWER6 or POWER6+ compatibility mode. Therefore, although the POWER7 processor has the ability to run four hardware threads per core simultaneously, using AIX 5.3 limits the number of hardware threads per core to two.

A partition with AIX 5.3 is limited to a maximum of 64 cores.

IBM AIX Version 6.1

If you install AIX 6.1 on a POWER7 processor-based server, the minimum level requirements depend on the target server model:

The following minimum levels of AIX Version 6.1 support Power 710, 720, 730, 740, and 795:

- AIX 6.1 with the 6100-04 Technology Level and Service Pack 7, or later
- AIX 6.1 with the 6100-05 Technology Level and Service Pack 3, or later
- AIX 6.1 with the 6100-06 Technology Level

The following minimum levels of AIX Version 6.1 support Power 750 and 755:

- AIX 6.1 with the 6100-02 Technology Level and Service Pack 8, or later
- AIX 6.1 with the 6100-03 Technology Level and Service Pack 5, or later
- AIX 6.1 with the 6100-04 Technology Level and Service Pack 2, or later

The following minimum levels of AIX Version 6.1 support Power 770 and 780:

- AIX 6.1 with the 6100-02 Technology Level and Service Pack 8, or later
- AIX 6.1 with the 6100-03 Technology Level and Service Pack 5, or later
- AIX 6.1 with the 6100-04 Technology Level and Service Pack 3, or later

A partition that uses AIX 6.1 with TL6 can run in POWER6, POWER6+ or POWER7 mode. The best approach is to run the partition in POWER7 mode to allow exploitation of new hardware capabilities such as SMT4 and Active Memory Expansion.

A partition with AIX 6.1 is limited to a maximum of 64 cores.

IBM AIX Version 7.1

AIX Version 7.1 comes will full support for the Power 710, 720, 730, 740, 750, 755, 770, 780, and 795, exploiting all the hardware features from the POWER7 processor, and also from the server architecture. A partition with AIX 7.1 can run in POWER6, POWER6+ or POWER7 mode, to enable Live Partition Mobility to different POWER6 and POWER7 systems. When running in POWER7 mode, a partition with AIX 7.1 can scale up to 256 cores and 8 TB of RAM.

Software key: Partition sizes greater than 128-cores (up to 256-cores) require a software key to enable. The purchase requires lab services pre-analysis as a prerequisite to shipment. The software key requires FC 1256 to be installed.

2.13.3 IBM i operating system

The IBM i operating system is supported on Power 710, 720, 730, 740, 750, 770, 780, and 795 at the following minimum levels:

- IBM i Version 6.1 with i 6.1.1 machine code, or later
- IBM i Version 7.1, or later

IBM i Standard Edition, and Application Server Edition options are available the Power 740, 750, 770, 780, and 795:

- IBM i Standard Edition offers an integrated operating environment for business processing
- IBM i Application Server Edition offers IBM i without DB2® for application and infrastructure serving.

IBM i is not supported on Power 755.

IBM periodically releases maintenance packages (service packs or technology levels) for the IBM i operating system. Information about these packages, downloading, and obtaining the CD-ROM is on the Fix Central website:

http://www-933.ibm.com/support/fixcentral/

2.13.4 Linux operating system

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 following versions of Linux on POWER7 processor-based servers are supported:

- SUSE Linux Enterprise Server 10 with SP3, enabled to run in POWER6 Compatibility mode
- SUSE Linux Enterprise Server 11, supporting POWER6 or POWER7 mode
- Red Hat Enterprise Linux AP 5 Update 5 for POWER, or later

Clients who want to configure Linux partitions in virtualized Power Systems should be aware of the following conditions:

- Not all devices and features that are 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. You may acquire Linux operating system licenses from IBM, to be included with the POWER7 processor-based servers, or from other Linux distributors.

See the following resources for more information:

- Features and external devices supported by Linux:

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

- SUSE Linux Enterprise Server 10,:

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

- Red Hat Enterprise Linux Advanced Server:

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

- Supported virtualization features are listed in 3.4.9, “Operating system support for PowerVM” on page 157.

2.14 Compiler technology

You can boost performance and productivity with IBM compilers on IBM Power Systems.

IBM XL C, XL C/C++ and XL FORTRAN compilers for AIX and for Linux exploit the latest POWER7 processor architecture. Release after release, these compilers continue to help improve application performance and capability, exploiting architectural enhancements made available through the advancement of the POWER technology.

IBM compilers are designed to optimize and tune your applications for execution on IBM POWER platforms, to help you unleash the full power of your IT investment, to create and maintain critical business and scientific applications, to maximize application performance, and to improve developer productivity. The performance gain from years of compiler optimization experience is seen in the continuous release-to-release compiler improvements that support the POWER4 processors, through to the POWER4+, POWER5, POWER5+ and POWER6 processors, and now including the new POWER7 processors. With the support of the latest POWER7 processor chip, IBM advances a more than 20-year investment in the XL compilers for POWER series and IBM PowerPC® series architectures.

XL C, XL C/C++, and XL FORTRAN features that were introduced to use the latest POWER7 processor include the following items:

- Vector unit and vector scalar extension (VSX) instruction set to efficiently manipulate vector operations in your application
- Vector functions within the Mathematical Acceleration Subsystem (MASS) libraries for improved application performance
- Built-in functions or intrinsics and directives for direct control of POWER instructions at the application level
- Architecture and tune compiler options to optimize and tune your applications

COBOL for AIX enables you to selectively target code generation of your programs to either exploit POWER7 systems architecture or to be balanced among all supported POWER systems. The performance of COBOL for AIX applications is improved by means of an enhanced back-end optimizer. With the back-end optimizer, a component that is also common to the IBM XL compilers, your applications can use the latest industry-leading optimization technology.

PL/I for AIX supports application development on the latest POWER7 processor.

IBM Rational® Development Studio for IBM i 7.1 provides programming languages for creating modern business applications. These languages include the ILE RPG, ILE COBOL, C, and C++ compilers and also the heritage RPG and COBOL compilers. The latest release includes performance improvements and XML processing enhancements for ILE RPG and ILE COBOL, improved COBOL portability with a new COMP-5 data type, and easier Unicode migration with relaxed USC2 rules in ILE RPG. Rational also released a product named Rational Open Access: RPG Edition. This product opens the ILE RPG file I/O processing, enabling partners, tool providers, and users to write custom I/O handlers that can access other devices such as databases, services, and web user interfaces.

IBM Rational Power Appliance is a preconfigured application development environment solution for Power Systems that includes a Power Express server preinstalled with a comprehensive set of Rational development software along with the AIX operating system. The Rational development software includes support for Collaborative Application Lifecycle Management (C/ALM) through IBM Rational Team Concert™ for Power Systems Software, a

set of software development tools from Rational Developer for Power Systems Software, and a choice between the XL C/C++ for AIX or COBOL for AIX compilers.

2.15 Energy management

The Power 795 offers support for power and thermal management through hardware monitoring and control. The IBM Systems Director Active Energy Manager exploits EnergyScale technology, enabling advanced energy management features to dynamically save power and further improve energy efficiency. *Intelligent Energy* optimization capabilities enable the POWER7 processor to operate at a higher frequency for increased performance, or significantly reduce frequency to save energy.

2.15.1 IBM EnergyScale technology

IBM EnergyScale technology provides functions to help the user understand and dynamically optimize the processor performance versus processor energy consumption, and system workload, to control IBM Power Systems power and cooling usage.

On POWER7 processor-based systems, the thermal power management device (TPMD) card is responsible for collecting the data from all system components, changing operational parameters in components, and interacting with the IBM Systems Director Active Energy Manager (an IBM Systems Directors plug-in) for energy management and control.

IBM EnergyScale makes use of power and thermal information that is collected from the system to implement policies that can lead to better performance, or better energy utilization. IBM EnergyScale includes the functions described in this section.

Power trending

EnergyScale provides continuous collection of real-time server energy consumption. This feature enables administrators to predict power consumption across their infrastructure and to react to business and processing needs. For example, administrators can use such information to predict data center energy consumption at various times of the day, week, or month.

Thermal reporting

IBM Director Active Energy Manager can display measured ambient temperature and calculated exhaust heat index temperature. This information can help identify data center hot spots that need attention.

Power Saver Mode

Power Saver Mode lowers the processor frequency and voltage on a fixed amount, reducing the energy consumption of the system while still delivering predictable performance. This percentage is predetermined to be within a safe operating limit and is not user configurable. The server is designed for a fixed frequency drop of up to 30% down from nominal frequency (the actual value depends on the server type and configuration). 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.

Dynamic Power Saver Mode

Dynamic Power Saver Mode varies processor frequency and voltage based on the utilization of the POWER7 processors. Processor frequency and utilization are inversely proportional for most workloads, implying that as the frequency of a processor increases, its utilization decreases, given a constant workload.Dynamic Power Saver Mode takes advantage of this relationship to detect opportunities to save power, based on measured real-time system utilization.

When a system is idle, the system firmware will lower the frequency and voltage to Power Energy Saver Mode values. When fully utilized, the maximum frequency will vary, depending on whether the user favors power savings or system performance. If an administrator prefers energy savings and a system is fully utilized, the system is designed to reduce the maximum frequency to 95% of nominal values. If performance is favored over energy consumption, the maximum frequency can be increased to up to 109% of nominal frequency for extra performance.

Dynamic Power Saver Mode is mutually exclusive with Power Saver mode. Only one of these modes can be enabled at a given time.

Power capping and soft power capping

There are two power ranges into which the power cap can be set: power capping and soft power capping.

- Power capping enforces a user-specified limit on power usage. Power Capping is not a power saving mechanism. It enforces power caps by actually throttling the processors in the system, degrading performance significantly. The idea of a power cap is to set a limit that must never be reached but frees up extra power never used in the data center. The *margined* power is this 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 because server specifications are always based on maximum configurations and worst case scenarios. The user must set and enable an energy cap from the IBM Director Active Energy Manager user interface.
- Soft power capping extends the allowed energy capping range further, beyond a region that can be guaranteed in all configurations and conditions. If the energy management goal is to meet a particular consumption limit, then Soft Power Capping is the mechanism to use.

Processor core Nap mode

The IBM POWER7 processor uses a low-power mode called Nap that stops processor execution when there is no work to do on that processor core. The latency of exiting Nap is small, typically not generating any impact on applications running. For that reason, the POWER Hypervisor can use the Nap mode as a general purpose idle state. When the operating system detects that a processor thread is idle, it yields control of a hardware thread to the POWER Hypervisor. The POWER Hypervisor immediately puts the thread into Nap mode. Nap mode allows the hardware to turn the clock off on most circuits inside the processor core. Reducing active energy consumption by turning off the clocks allows the temperature to fall, which further reduces leakage (static) power of the circuits causing a cumulative effect. Nap mode saves from 10 -15% of power consumption in the processor core.

Processor core Sleep mode

To be able to save even more energy, the POWER7 processor has an even lower power mode called Sleep. Before a core and its associated L2 and L3 caches enter Sleep mode, caches are flushed and transition lookaside buffers (TLB) are invalidated, and hardware clock is

turned off in the core and in the caches. Voltage is reduced to minimize leakage current. Processor cores inactive in the system (such as CoD processor cores) are kept in Sleep mode. Sleep mode saves about 35% power consumption in the processor core and associated L2 and L3 caches.

Fan control and altitude input

System firmware dynamically adjusts fan speed based on energy consumption, altitude, ambient temperature, and energy savings modes. Power Systems are designed to operate in worst-case environments, in hot ambient temperatures, at high altitudes, and with high power components. In a typical case, one or more of these constraints are not valid. When no power savings setting is enabled, fan speed is based on ambient temperature, and assumes a high-altitude environment. When a power savings setting is enforced (either Power Energy Saver Mode or Dynamic Power Saver Mode), fan speed will vary, based on power consumption, ambient temperature, and altitude available. System altitude can be set in IBM Director Active Energy Manager. If no altitude is set, the system will assume a default value of 350 meters above sea level.

Processor folding

Processor folding is a consolidation technique that dynamically adjusts, over the short-term, the number of processors available for dispatch to match the number of processors demanded by the workload. As the workload increases, the number of processors made available increases; as the workload decreases, the number of processors made available decreases. Processor folding increases energy savings during periods of low to moderate workload because unavailable processors remain in low-power idle states (Nap or Sleep) longer.

EnergyScale for I/O

IBM POWER7 processor-based systems automatically power off hot-pluggable, PCI adapter slots that are empty or not being used. System firmware automatically scans all pluggable PCI slots at regular intervals, looking for those that meet the criteria for being not in use and powering them off. This support is available for all POWER7 processor-based servers, and the expansion units that they support.

Server Power Down

If overall data center processor utilization is low, workloads can be consolidated on fewer numbers of servers so that some servers can be turned off completely. This step is sensible when long periods of low utilization, such as weekends, will occur. Active Energy Manager provides information, such as the power that will be saved and the time it will take to bring a server back online, that can be used to help make the decision to consolidate and power off. As with many of the available features in IBM Systems Director and Active Energy Manager, this function can be scripted and can be automated.

Partition Power Management

Available with Active Energy Manager 4.3.1 or newer, and POWER7 systems with EM730 firmware or newer, is the capability to set a power savings mode for partitions or the system processor pool. As with the system-level power savings modes, the per-partition power savings modes can be used to achieve a balance between the power consumption and the performance of a partition. Only partitions that have dedicated processing units can have a unique power savings setting. Partitions that run in shared processing mode have a common power savings setting, which is that of the system processor pool. The reason is because processing unit fractions cannot be power-managed.

As in the case of system-level power savings, two Dynamic Power Saver mode options are offered:

- Favor partition performance
- Favor partition power savings

The user must configure this setting from Active Energy Manager. When Dynamic Power Saver is enabled in either mode, system firmware continuously monitors the performance and utilization of each of the computer's POWER7 processor cores that belong to the partition. Based on this utilization and performance data, the firmware dynamically adjusts the processor frequency and voltage, reacting within milliseconds to adjust workload performance and also deliver power savings when the partition is underutilized.

In addition to the two Dynamic Power Saver options, the customer can select to have no power savings on a given partition. This option keeps the processor cores that are assigned to the partition running at their nominal frequencies and voltages.

A new power savings mode, referred to as Inherit Host Setting, is available and is applicable only to partitions. When configured to use this setting, a partition adopts the power savings mode of its hosting server. By default, all partitions with dedicated processing units, and the system processor pool, are set to Inherit Host Setting.

On POWER7 processor-based systems, several EnergyScales are imbedded in the hardware and do not require an operating system or external management component. More advanced functionality requires Active Energy Manager (AEM) and IBM Systems Director.

Table 2-35 lists all supported features, showing all cases where AEM is or is not required. The table also lists features that can be activated by traditional user interfaces (ASMI or HMC).

Table 2-35 AEM support

Feature	Active Energy Manager required	ASMI	HMC
Power trending	Yes	No	No
Thermal reporting	Yes	No	No
Static Power Saver	No	Yes	Yes
Dynamic Power Saver	Yes	No	No
Power capping	Yes	No	No
Energy-optimized Fans	No	—	—
Processor core Nap	No	—	—
Processor core Sleep	No	—	—
Processor folding	No	—	—
EnergyScale for I/O	No	—	—
Server power down	Yes	—	—
Partition Power Management	Yes	—	—

The Power 795 system implements all the Energy Scale capabilities listed in 2.15.1, “IBM EnergyScale technology” on page 122.

2.15.2 Thermal power management device card

The thermal power management device (TPMD) card is a separate microcontroller installed on some POWER6 processor-based systems, and in all POWER7 processor-based systems. It runs real-time firmware whose sole purpose is to manage system energy.

The TPMD card monitors the processor modules, memory, environmental temperature, and fan speed; and based on this information, it can act upon the system to maintain optimal power and energy conditions (for example, increase the fan speed to react to a temperature change). It also interacts with the IBM Systems Director Active Energy Manager to report power and thermal information, and to receive input from AEM on policies to be set. The TPMD is part of the EnergyScale infrastructure.

On the Power 795, each book has its own redundant TPMD card.

3

Virtualization

As you look for ways to maximize the return on your IT infrastructure investments, consolidating workloads becomes an attractive proposition.

IBM Power Systems combined with PowerVM technology are designed to help you consolidate and simplify your IT environment with the following key capabilities:

- Improve server utilization and sharing I/O resources to reduce total cost of ownership and make better use of IT assets.
- Improve business responsiveness and operational speed by dynamically re-allocating resources to applications as needed, to better match changing business needs or handle unexpected changes in demand.
- Simplify IT infrastructure management by making workloads independent of hardware resources, thereby enabling you to make business-driven policies to deliver resources based on time, cost, and service-level requirements.

This chapter discusses the virtualization technologies and features on IBM Power Systems:

- POWER Hypervisor
- POWER modes
- Active Memory Expansion
- PowerVM
- System Planning Tool

3.1 POWER Hypervisor

Combined with features designed into the POWER7 processors, the POWER Hypervisor delivers functions that enable other system technologies, including logical partitioning technology, virtualized processors, IEEE VLAN compatible virtual switch, virtual SCSI adapters, virtual Fibre Channel adapters, 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 "Processing mode" on page 139.)
- 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 performs a reset or 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 and also when not connected to the managed console. It requires memory to support the resource assignment to the logical partitions on the server. The amount of memory that is 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 specified in the logical partition profiles

The minimum amount of physical memory that is required to create a partition will be the size of the system's logical memory block (LMB). The default LMB size varies according to the amount of memory that is configured in the CEC (Table 3-1).

Table 3-1 Configured CEC memory-to-default logical memory block size

Configurable CEC memory	Default logical memory block
Up to and including 32 GB	128 MB
Greater than 32 GB	256 MB

In most cases, however, the actual minimum requirements and recommendations of the supported operating systems are above 256 MB. Physical memory is assigned to partitions in increments of LMB.

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

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

Virtual SCSI

The POWER Hypervisor provides a virtual SCSI mechanism for virtualization of storage devices. Storage virtualization is accomplished by using two, paired adapters:

- A virtual SCSI server adapter
- A virtual SCSI client adapter

A Virtual I/O Server partition or a IBM i partition can define virtual SCSI server adapters. Other partitions are *client* partitions. The Virtual I/O Server partition is a special logical partition, as described in 3.4.4, "Virtual I/O Server (VIOS)" on page 145. The Virtual I/O Server software is included on all PowerVM Editions and when using the PowerVM Standard Edition and PowerVM Enterprise Edition, dual Virtual I/O Servers can be deployed to provide maximum availability for client partitions when performing Virtual I/O Server maintenance.

Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions on the same server to use fast and secure communication without any need for physical interconnection. The virtual Ethernet allows a transmission speed up to 20 Gbps, depending on the maximum transmission unit (MTU) size, type of communication and CPU entitlement. Virtual Ethernet support began with IBM AIX Version 5.3, Red Hat Enterprise Linux 4 and SUSE SLES 9, and it is supported on all later versions. (For more information, see 3.4.9, "Operating system support for PowerVM" on page 157). The virtual Ethernet is part of the base system configuration.

Virtual Ethernet has the following major features:

- The virtual Ethernet adapters can be used for both IPv4 and IPv6 communication and can transmit packets with a size up to 65,408 bytes. Therefore, the maximum MTU for the corresponding interface can be up to 65,394 (65,390 if VLAN tagging is used).
- The POWER Hypervisor presents itself to partitions as a virtual 802.1Q-compliant switch. The maximum number of VLANs is 4096. Virtual Ethernet adapters can be configured as either untagged or tagged (following the IEEE 802.1Q VLAN standard).
- A partition can support 256 virtual Ethernet adapters. Besides a default port VLAN ID, the number of additional VLAN ID values that can be assigned per virtual Ethernet adapter is 20, which implies that each virtual Ethernet adapter can be used to access 21 virtual networks.
- Each partition operating system detects the virtual local area network (VLAN) switch as an Ethernet adapter without the physical link properties and asynchronous data transmit operations.

Any virtual Ethernet can also have connectivity outside of the server if a layer-2 bridge to a physical Ethernet adapter is set in one Virtual I/O Server partition (see 3.4.4, "Virtual I/O Server (VIOS)" on page 145, for more details about shared Ethernet), also known as Shared Ethernet Adapter.

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

Virtual Fibre Channel

A virtual Fibre Channel adapter is a virtual adapter that provides client logical partitions with a Fibre Channel connection to a storage area network through the Virtual I/O Server logical partition. The Virtual I/O Server logical partition provides the connection between the virtual Fibre Channel adapters on the Virtual I/O Server logical partition and the physical Fibre Channel adapters on the managed system. Figure 3-1 depicts the connections between the client partition virtual Fibre Channel adapters and the external storage. For additional information, see “N_Port ID virtualization (NPIV)” on page 157.

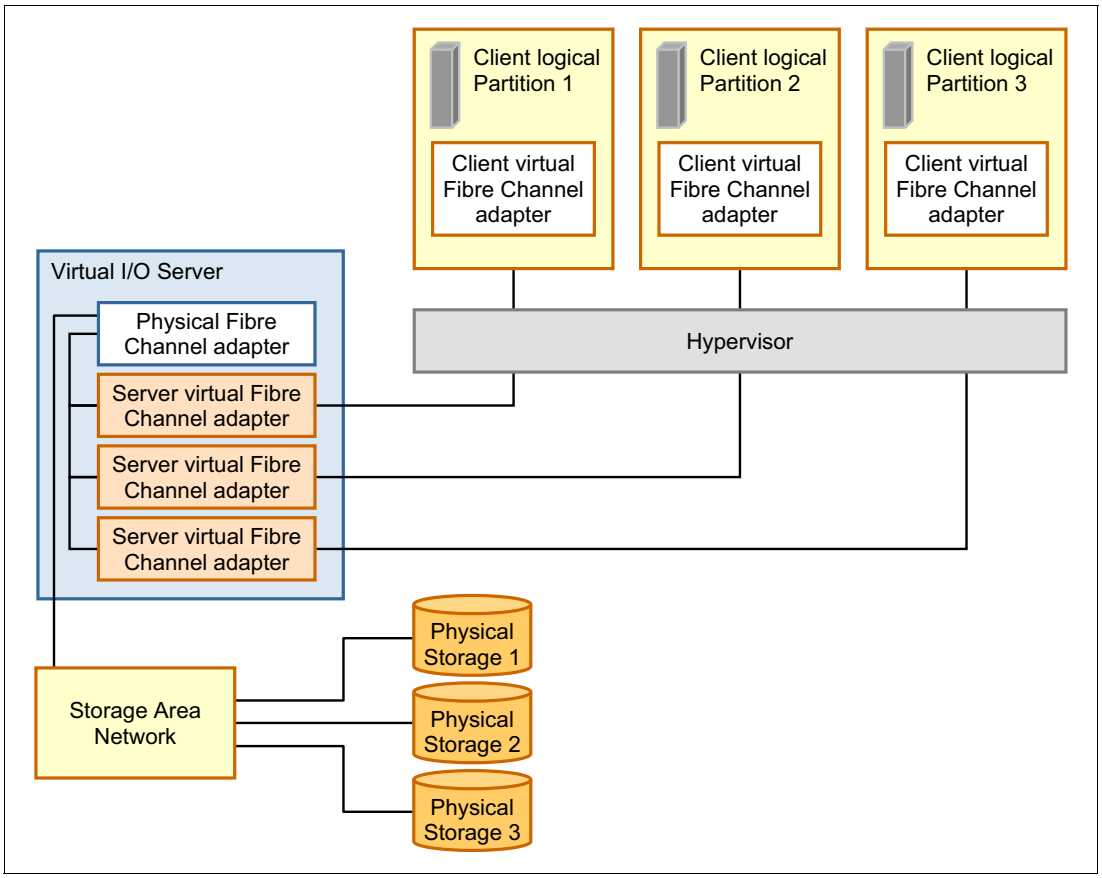

Figure 3-1 Connectivity between virtual Fibre Channels adapters and external SAN devices

Virtual (TTY) console

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

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

3.2 POWER processor modes

Although the POWER modes are not strictly a virtualization feature, they are described here because they affect various virtualization features.

On Power Systems servers, partitions can be configured to run in several modes, including the following modes:

- POWER6 compatibility mode: This execution mode is compatible with Version 2.05 of the Power Instruction Set Architecture (ISA). For more information, visit the following address:

  `http://power.org/wp-content/uploads/2012/07/PowerISA_V2.05.pdf`

- POWER6+ compatibility mode: This mode is similar to POWER6, with eight additional Storage Protection Keys.
- POWER7 mode: This is the native mode for POWER7 processors, implementing the v2.06 of the Power Instruction Set Architecture. For more information, visit the following address:

  `http://power.org/wp-content/uploads/2012/07/PowerISA_V2.06B_V2_PUBLIC.pdf`

The selection of the mode is made on a per-partition basis, from the managed console, by editing the partition profile (Figure 3-2).

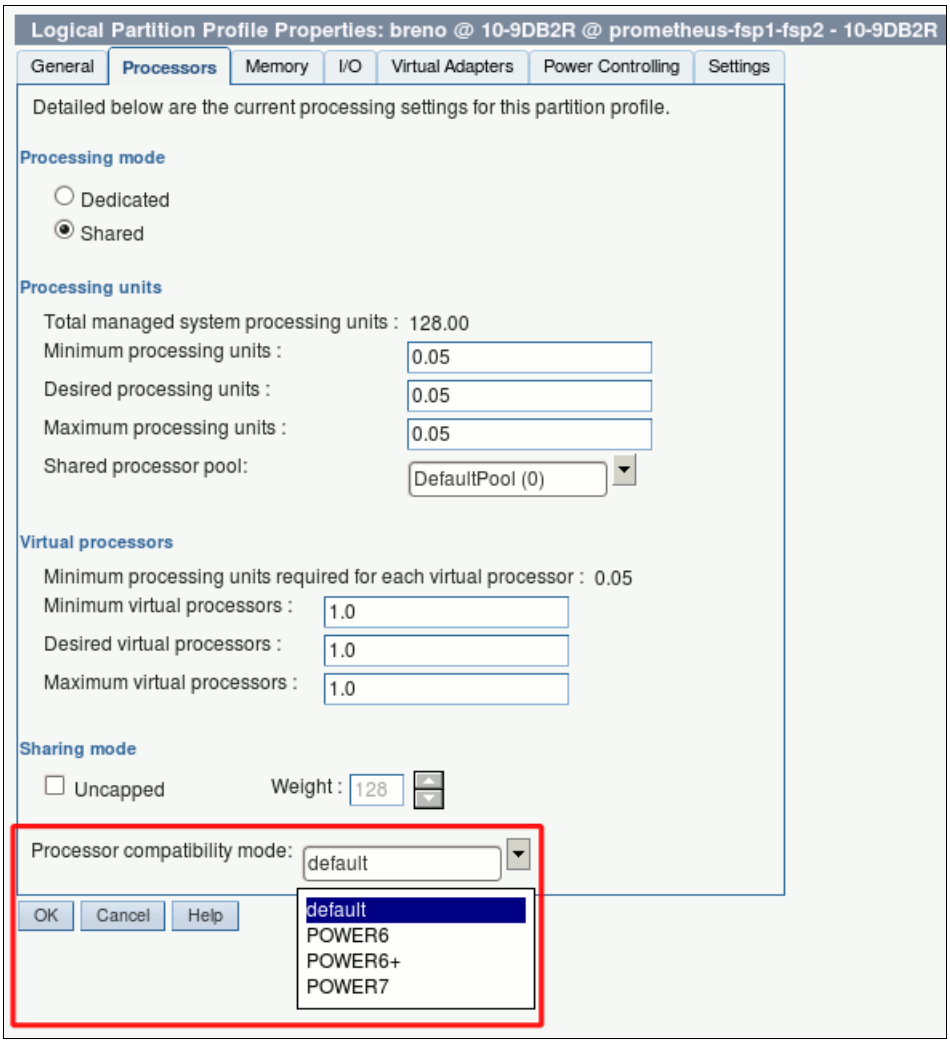

Figure 3-2 Configuring partition profile compatibility mode from the managed console

Table 3-2 lists the differences between the modes.

Table 3-2 Differences between POWER6 and POWER7 Compatibility mode

POWER6 and POWER6+ mode	POWER7 mode	Customer value
2-thread SMT	4-thread SMT	Throughput performance, processor core utilization
Vector Multimedia Extension/ AltiVec (VMX)	Vector Scalar Extension (VSX)	High-performance computing
Affinity OFF by default	3-tier memory, Micro partition Affinity, Dynamic Platform Optimizer	Improved system performance for system images spanning sockets and nodes
► Barrier Synchronization ► Fixed 128-byte array, Kernel Extension Access	► Enhanced Barrier Synchronization ► Variable Sized Array, User Shared Memory Access	High-performance computing parallel programming synchronization facility
► 64-core and 128-thread scaling	► 32-core and 128-thread scaling ► 64-core and 256-thread scaling ► 128-core and 512-thread scaling ► 256-core and 1024-thread scaling	Performance and scalability for large scale-up single system image workloads (such as OLTP, ERP scale-up, and WPAR consolidation)
EnergyScale CPU Idle	EnergyScale CPU Idle and Folding with Nap and Sleep	Improved energy efficiency

3.3 Active Memory Expansion

Active Memory Expansion enablement (FC 4790) is an optional feature of POWER7 processor-based servers that must be specified when creating the configuration in the e-Config tool for Power 795.

This feature enables memory expansion on the system. Using compression/decompression of memory content can effectively expand the maximum memory capacity, providing additional server workload capacity and performance.

Active Memory Expansion is a POWER technology that allows the effective maximum memory capacity to be much larger than the true physical memory maximum. Compression/decompression of memory content can allow memory expansion up to 100%, which in turn enables a partition to perform significantly more work or support more users with the same physical amount of memory. Similarly, it can allow a server to run more partitions and do more work for the same physical amount of memory.

Active Memory Expansion is available for partitions running AIX 6.1, Technology Level 4 with SP2, or later. Linux will support Active Memory Expansion on the next major versions.

Active Memory Expansion uses CPU resource of a partition to compress/decompress the memory contents of this same partition.The trade-off of memory capacity for processor cycles can be an excellent choice, but the degree of expansion varies based on how compressible

the memory content is, and it also depends on having adequate spare CPU capacity available for this compression/decompression.

Tests in IBM laboratories, using sample work loads, showed excellent results for many workloads in terms of memory expansion for each additional CPU used. Other test workloads had more modest results. The ideal scenario is when there are a lot of “cold pages,” that is, infrequently referenced pages. However, if a lot of memory pages are referenced frequently, the Active Memory Expansion might not be a good choice.

TIP: If the workload is Java-based, the garbage collector must be tuned, so that it does not access the memory pages so often, converting cold pages to hot.

Clients have much control over Active Memory Expansion usage. Each individual AIX partition can turn on or turn off Active Memory Expansion. Control parameters set the amount of expansion desired in each partition to help control the amount of CPU used by the Active Memory Expansion function. An initial program load (IPL) is required for the specific partition that is turning memory expansion on or off. After turned on, monitoring capabilities are available in standard AIX performance tools, such as **`lparstat`**, **`vmstat`**, **`topas`**, and **`svmon`**.

Figure 3-3 represents the percentage of CPU that is used to compress memory for two partitions with separate profiles. The green curve (indicated by 1 in the figure) corresponds to a partition that has spare processing power capacity. The blue curve (2) corresponds to a partition constrained in processing power.

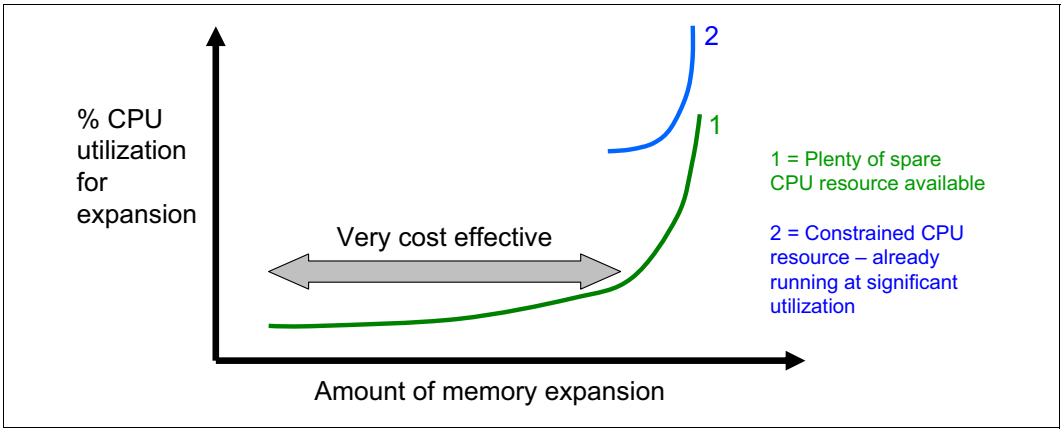

Figure 3-3 CPU usage versus memory expansion effectiveness

Both cases show that there is a knee-of-curve relationship for the CPU resource that is required for memory expansion:

- Busy processor cores do not have resources to spare for expansion.
- The more memory expansion that is done, the more CPU resource is required.

The knee varies depending on how compressible the memory contents are. This example demonstrates the need for a case-by-case study of whether memory expansion can provide a positive return on investment.

To help you do this study, a planning tool is included with AIX 6.1 Technology Level 4, allowing you to sample actual workloads and estimate how expandable the partition's memory is and how much CPU resource is needed. Any Power Systems model can run the planning tool. Figure 3-4 shows an example of the output that is returned by this planning tool. The tool outputs various real memory and CPU resource combinations to achieve the effective memory that you want. It also recommends one particular combination. In this example, the tool recommends that you allocate 13% of processing power (2.13 physical processors in this setup) to benefit from 119% extra memory capacity.

```
Active Memory Expansion Modeled Statistics:
-------------------------------------------
Modeled Expanded Memory Size :  52.00 GB
Achievable Compression ratio   :4.51

Expansion    Modeled True      Modeled              CPU Usage
Factor       Memory Size       Memory Gain          Estimate
---------    -------------     ------------------   -----------
    1.40         37.25 GB        14.75 GB [ 40%]   0.00 [  0%]
    1.80         29.00 GB        23.00 GB [ 79%]   0.87 [  5%]
    2.19         23.75 GB        28.25 GB [119%]   2.13 [ 13%]
    2.57         20.25 GB        31.75 GB [157%]   2.96 [ 18%]
    2.98         17.50 GB        34.50 GB [197%]   3.61 [ 23%]
    3.36         15.50 GB        36.50 GB [235%]   4.09 [ 26%]

Active Memory Expansion Recommendation:
---------------------------------------
The recommended AME configuration for this workload is to configure the LPAR
with a memory size of 23.75 GB and to configure a memory expansion factor
of 2.19.  This will result in a memory gain of 119%. With this
configuration, the estimated CPU usage due to AME is approximately 2.13
physical processors, and the estimated overall peak CPU resource required for
the LPAR is 11.65 physical processors.
```

Figure 3-4 Output from Active Memory Expansion planning tool

After you select the value of the memory expansion factor that you want to achieve, you can use this value to configure the partition from the managed console (Figure 3-5).

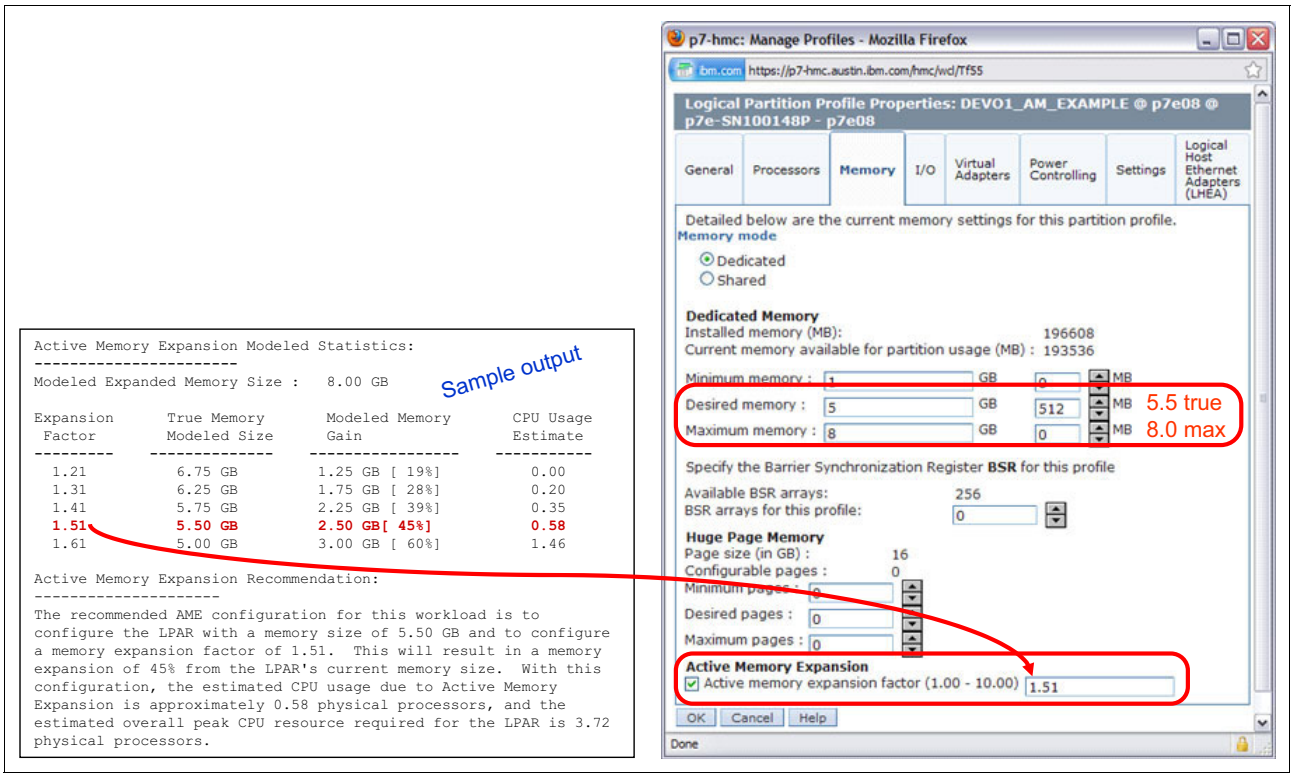

Figure 3-5 Using the planning tool result to configure the partition

On the HMC menu that describes the partition, select the **Active Memory Expansion** check box and enter the true and maximum memory, and the memory expansion factor. To turn off expansion, clear the check box. In both cases, to activate the change, rebooting the partition is necessary.

IA one-time, 60-day trial of Active Memory Expansion is available to provide more exact memory expansion and CPU measurements. The trial can be requested by using the Capacity on Demand web page:

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

Active Memory Expansion can be ordered with the initial order of the server or as an MES order. A software key is provided when the enablement feature is ordered that is applied to the server. To enable the physical server, rebooting is not necessary. The key is specific to an individual server and is permanent. It cannot be moved to a separate server. This feature is ordered per server, independently of the number of partitions using memory expansion.

From the HMC, you can view whether the Active Memory Expansion feature was activated (Figure 3-6).

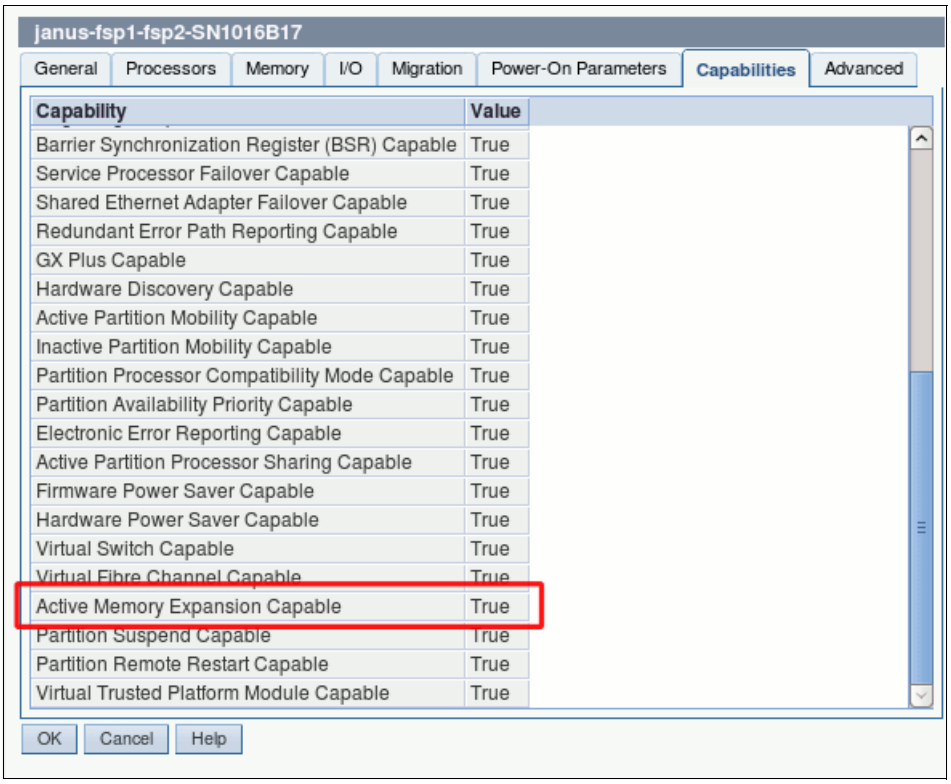

Figure 3-6 Server capabilities listed from the HMC

Target system: If you want to move an LPAR that uses Active Memory Expansion to another system that uses Live Partition Mobility, the target system must support Active Memory Expansion (the target system must have Active Memory Expansion activated with the software key). If the target system does not have Active Memory Expansion activated, the mobility operation fails during the pre-mobility check phase, and an appropriate error message displays to the user.

For detailed information regarding Active Memory Expansion, download the *Active Memory Expansion: Overview and Usage Guide:*

http://www.ibm.com/systems/power/hardware/whitepapers/am_exp.html

3.4 PowerVM

The PowerVM platform is the family of technologies, capabilities, and offerings that deliver industry-leading virtualization on the IBM Power Systems. It is the new umbrella branding term for PowerVM (Logical Partitioning, Micro-Partitioning, POWER Hypervisor, Virtual I/O Server, Live Partition Mobility, Workload Partitions, and more). As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and value-added software. The licensed features of each of the three editions of PowerVM are described in 3.4.1, “PowerVM editions” on page 137.

3.4.1 PowerVM editions

This section provides information about the virtualization capabilities of the PowerVM, and the three editions of PowerVM, which are suited for various purposes:

- PowerVM Express Edition

  PowerVM Express Edition is designed for customers looking for an introduction to more advanced virtualization features at a highly affordable price, generally in single-server projects.

- PowerVM Standard Edition

  This edition provides advanced virtualization functions and is intended for production deployments and server consolidation.

- PowerVM Enterprise Edition

  This edition is suitable for large server deployments such as multi-server deployments and cloud infrastructure. It includes unique features such as Active Memory Sharing and Live Partition Mobility.

Table 3-3 shows the editions of the PowerVM that are available for Power 795.

Table 3-3 Availability of PowerVM per POWER7 processor technology-based server model

PowerVM editions	Feature code
Express	N/A
Standard	FC 7943
Enterprise	FC 8002

For more information about the features included on each version of PowerVM, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940-04.

HMC: At the time of writing, the IBM Power 795 (9119-FHB) can be managed only by the Hardware Management Console.

3.4.2 Logical partitions (LPARs)

LPARs and virtualization increase utilization of system resources and add a new level of configuration possibilities. This section provides details and configuration specifications about this topic.

Logical partitioning

Logical partitioning was introduced with the POWER4 processor-based product line and the AIX Version 5.1, Red Hat Enterprise Linux 3.0 and SUSE SLES Linux 9.0 operating systems. This technology offered the capability to divide a pSeries system into separate logical systems, allowing each LPAR to run an operating environment on dedicated attached devices, such as processors, memory, and I/O components.

Later, dynamic logical partitioning increased the flexibility, allowing selected system resources, such as processors, memory, and I/O components, to be added and deleted from logical partitions while they were executing. AIX Version 5.2, with all the necessary enhancements to enable dynamic LPAR, was introduced in 2002. At the same time, Red Hat Enterprise Linux 5 and SUSE Enterprise Linux 9.0 were also able do support dynamic logical partitioning. The ability to reconfigure dynamic LPARs encourages system administrators to

dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

Micro-Partitioning technology

Micro-Partitioning technology allows you to allocate fractions of processors to a logical partition. This technology was introduced with POWER5 processor-based systems. A logical partition that uses fractions of processors is also known as a shared processor partition or micropartition. Micropartitions run over a set of processors called a shared processor pool, and virtual processors are used to let the operating system manage the fractions of processing power that is assigned to the logical partition. From an operating system perspective, a virtual processor cannot be distinguished from a physical processor, unless the operating system has been enhanced to be made aware of the difference. Physical processors are abstracted into virtual processors that are available to partitions. The meaning of the term *physical processor* in this section is a *processor core*. For example, a 2-core server has two physical processors.

When defining a shared processor partition, several options must be defined:

- The minimum, desired, and maximum processing units

  Processing units are defined as processing power, or the fraction of time that the partition is dispatched on physical processors. Processing units define the capacity entitlement of the partition.

- The shared processor pool

  Pick one from the list with the names of each configured shared processor pool. This list also displays the pool ID of each configured shared processor pool in parentheses. If the name of the desired shared processor pool is not available here, you must first configure the desired shared processor pool by using the shared processor pool Management window. Shared processor partitions use the default shared processor pool, which is called DefaultPool by default. See 3.4.3, “Multiple shared processor pools” on page 140, for details about multiple shared processor pools.

- Whether the partition will be able to access extra processing power to “fill up” its virtual processors above its capacity entitlement (selecting either to cap or uncap your partition)

  If spare processing power is available in the shared processor pool or other partitions are not using their entitlement, an uncapped partition can use additional processing units if its entitlement is not enough to satisfy its application processing demand.

- The weight (preference) in the case of an uncapped partition

- The minimum, desired, and maximum number of virtual processors

The POWER Hypervisor calculates partition processing power based on minimum, desired, and maximum values, processing mode, and is also based on requirements of other active partitions. The actual entitlement is never smaller than the processing unit’s desired value, but can exceed that value in the case of an uncapped partition and up to the number of virtual processors allocated.

On the POWER7 processors, a partition can be defined with a processor capacity as small as 0.05 processing units. This value represents 0.05 of a physical processor. Each physical processor can be shared by up to 20 shared processor partitions, and the partition’s entitlement can be incremented fractionally by as little as 0.01 of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC.

Cores and maximums: The IBM Power 795 machine supports up to 256 cores, and has the following maximums:

- Up to 256 dedicated partitions
- Up to 1000 micropartitions (20 micropartitions max. per physical active core)

An important point is that the maximums stated are supported by the hardware, but the practical limits depend on application workload demands.

Consider the following additional information about virtual processors:

- A virtual processor can be running (dispatched) either on a physical processor or as standby waiting for a physical processor to became available.
- Virtual processors do not introduce any additional abstraction level. They are only a dispatch entity. When running on a physical processor, virtual processors run at the same speed as the physical processor.
- Each partition's profile defines CPU entitlement that determines how much processing power any given partition should receive. The total sum of CPU entitlement of all partitions cannot exceed the number of available physical processors in a shared processor pool.
- The number of virtual processors can be changed dynamically through a dynamic LPAR operation.

Processing mode

When you create a logical partition, you can assign entire processors for dedicated use, or you can assign partial processing units from a shared processor pool. This setting defines the processing mode of the logical partition. The diagram in Figure 3-7 shows the concepts included in this section.

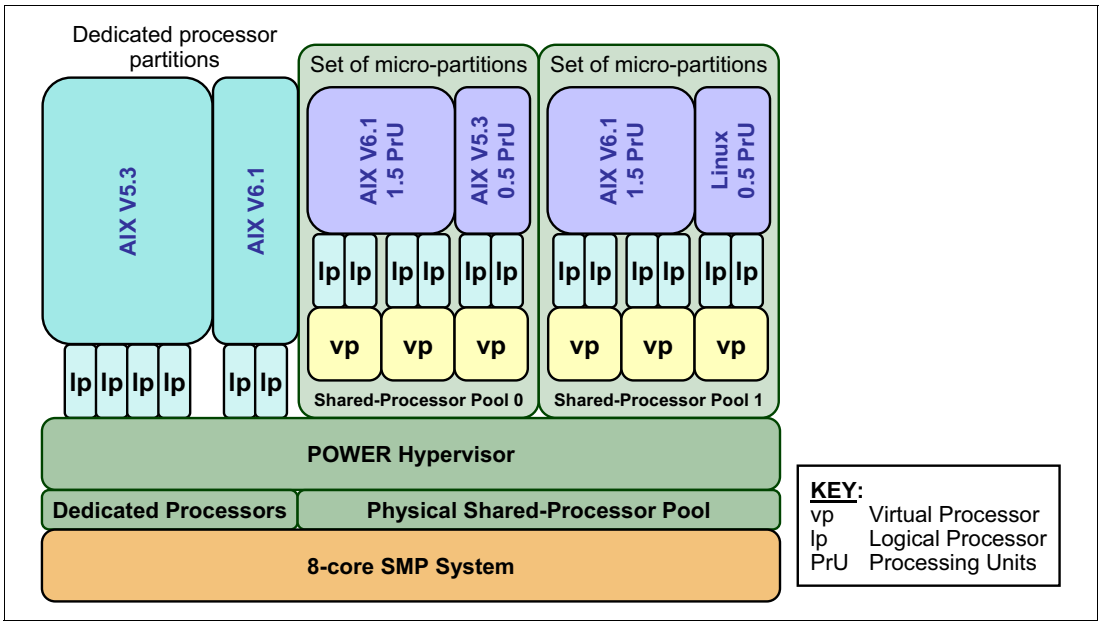

Figure 3-7 Logical partitioning concepts

Dedicated mode

In dedicated mode, physical processors are assigned as a whole to partitions. The simultaneous multithreading feature in the POWER7 processor core allows the core to execute instructions from two or four independent software threads simultaneously. To support this feature, we use the concept of *logical processors*. The operating system (AIX, IBM i, or Linux) sees one physical processor as two or four logical processors if the simultaneous multithreading feature is on. It can be turned off and on dynamically while the operating system is executing (for AIX, use the **smtctl** command; for Linux, use **ppc64_cpu --smt** command). If simultaneous multithreading is off, each physical processor is presented as one logical processor, and thus only one thread.

Shared dedicated mode

On POWER7 processor technology-based servers, you can configure dedicated partitions to become processor donors for idle processors that they own, allowing for the donation of spare CPU cycles from dedicated processor partitions to a shared processor pool. The dedicated partition maintains absolute priority for dedicated CPU cycles. Enabling this feature can help to increase system utilization without compromising the computing power for critical workloads in a dedicated processor.

Shared mode

In shared mode, logical partitions use virtual processors to access fractions of physical processors. Shared partitions can define any number of virtual processors (the maximum number is 10 times the number of processing units assigned to the partition). From the POWER Hypervisor point of view, virtual processors represent dispatching objects. The POWER Hypervisor dispatches virtual processors to physical processors according to the partition's processing units entitlement. One processing unit represents one physical processor's processing capacity. At the end of the POWER Hypervisor dispatch cycle (10 ms), all partitions receive total CPU time equal to their processing unit's entitlement. The logical processors are defined on top of virtual processors. So, even with a virtual processor, the concept of a logical processor exists and the number of logical processors depends whether the simultaneous multithreading is turned on or off.

3.4.3 Multiple shared processor pools

Multiple shared processor pools (MSPPs) capability is supported on POWER7 processor-based servers. With this capability, a system administrator can create a set of micropartitions with the purpose of controlling the processor capacity that can be consumed from the physical shared processor pool.

To implement MSPPs, a set of underlying techniques and technologies exists. Figure 3-8 shows an overview of the architecture of multiple shared processor pools.

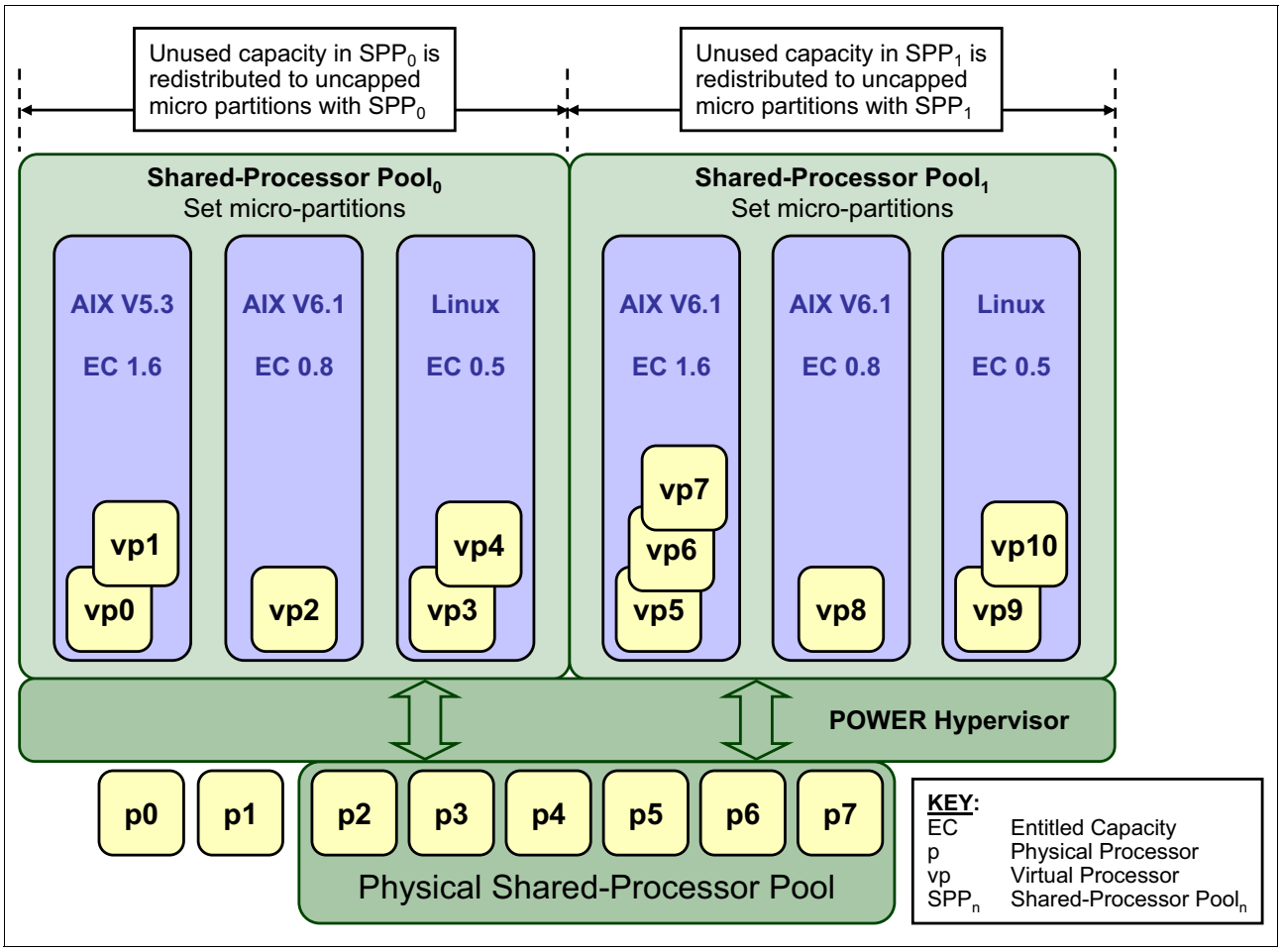

Figure 3-8 Overview of the architecture of multiple shared processor pools

Micropartitions are created and then identified as members of either the default shared processor pool$_0$ or a user-defined shared processor pool$_n$. The virtual processors that exist within the set of micropartitions are monitored by the POWER Hypervisor, and processor capacity is managed according to user-defined attributes.

If the Power Systems server is under heavy load, each micropartition within a shared processor pool is guaranteed its processor entitlement plus any capacity that it might be allocated from the reserved pool capacity if the micropartition is uncapped.

If certain micropartitions in a shared processor pool do not use their capacity entitlement, the unused capacity is ceded and other uncapped micropartitions within the same shared processor pool are allocated the additional capacity according to their uncapped weighting. In this way, the entitled pool capacity of a shared processor pool is distributed to the set of micropartitions within that shared processor pool.

All Power Systems servers that support the multiple shared processor pools capability will have a minimum of one (the default) shared processor pool and up to a maximum of 64 shared processor pools.

Default shared processor pool (SPP_0)

On any Power Systems server that supports multiple shared processor pools, a default shared processor pool is always automatically defined. The default shared processor pool has a pool identifier of zero (SPP-ID = 0) and can also be referred to as SPP_0. The default shared processor pool has the same attributes as a user-defined shared processor pool except that these attributes are not directly under the control of the system administrator. They have fixed values (Table 3-4).

Table 3-4 Attribute values for the default shared processor pool (SPP_0)

SPP_0 attribute	Value
Shared processor pool ID	0
Maximum pool capacity	The value is equal to the capacity in the physical shared processor pool.
Reserved pool capacity	0
Entitled pool capacity	Sum (total) of the entitled capacities of the micropartitions in the default shared processor pool.

Creating multiple shared processor pools

The default shared processor pool (SPP_0) is automatically activated by the system and is always present.

All other shared processor pools exist, but by default are inactive. By changing the maximum pool capacity of a shared processor pool to a value greater than zero, it becomes active and can accept micropartitions (either transferred from SPP_0 or newly created).

Levels of processor capacity resolution

The following two levels of processor capacity resolution are implemented by the POWER Hypervisor and multiple shared processor pools:

- $Level_0$

  The first level, $Level_0$, is the resolution of capacity within the same shared processor pool. Unused processor cycles from within a shared processor pool are harvested and then redistributed to any eligible micropartition within the same shared processor pool.

- $Level_1$

  This is the second level of processor capacity resolution. When all $Level_0$ capacity is resolved within the multiple shared processor pools, the POWER Hypervisor harvests unused processor cycles and redistributes them to eligible micropartitions regardless of the multiple shared processor pools structure. See “Capacity allocation above the entitled pool capacity (Level1)” on page 143

Figure 3-9 shows the levels of unused capacity redistribution that are implemented by the POWER Hypervisor.

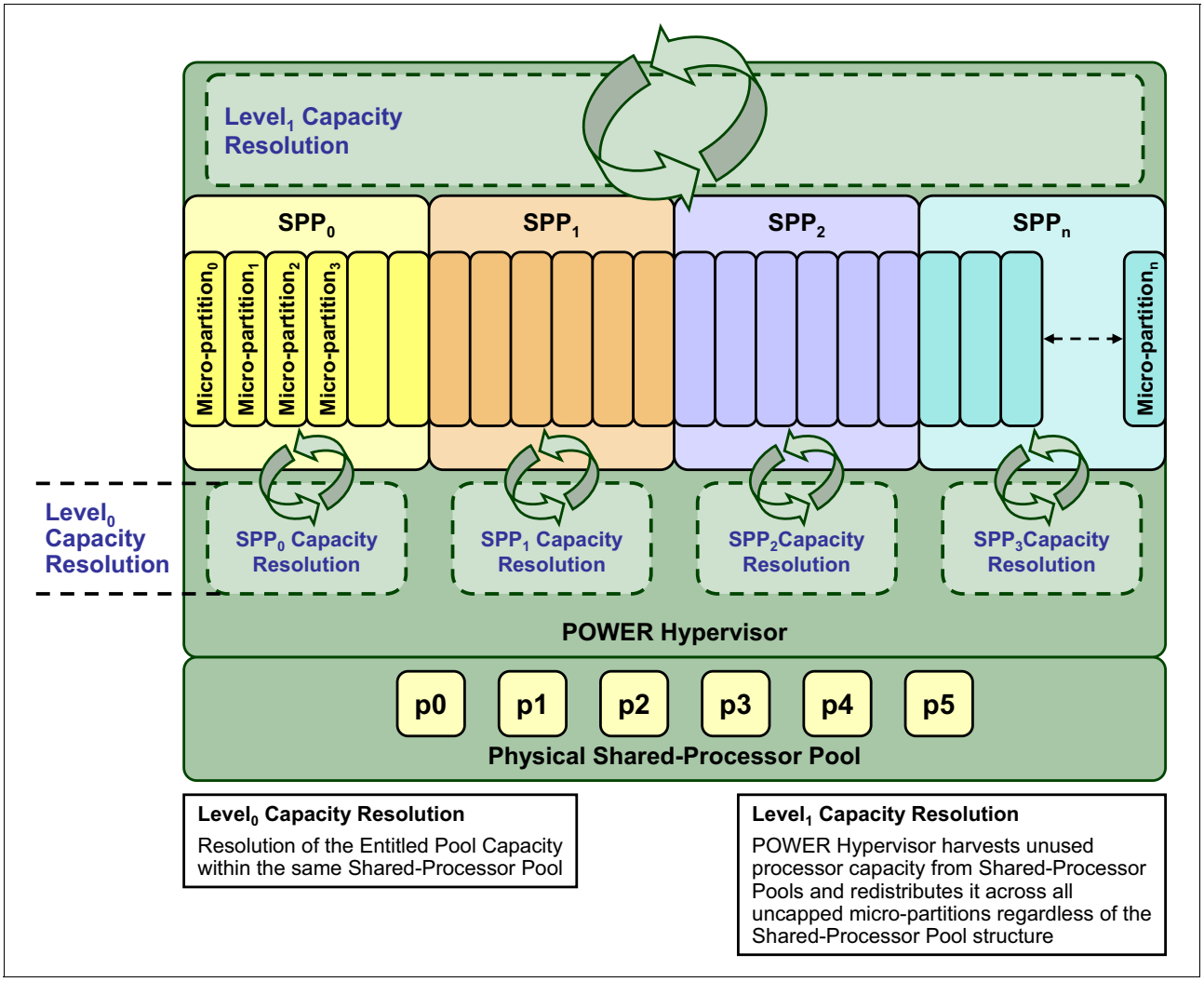

Figure 3-9 The levels of unused capacity redistribution

Capacity allocation above the entitled pool capacity ($Level_1$)

The POWER Hypervisor initially manages the entitled pool capacity at the shared processor pool level. This is where unused processor capacity within a shared processor pool is harvested and then redistributed to uncapped micropartitions within the same shared processor pool. This level of processor capacity management is sometimes referred to as $Level_0$ capacity resolution.

At a higher level, the POWER Hypervisor harvests unused processor capacity from the multiple shared processor pools that do not consume all of their entitled pool capacity. If a particular shared processor pool is heavily loaded and several of the uncapped micropartitions within it require additional processor capacity (above the entitled pool capacity), then the POWER Hypervisor redistributes some of the extra capacity to the uncapped micropartitions. This level of processor capacity management is sometimes referred to as $Level_1$ capacity resolution.

To redistribute unused processor capacity to uncapped micropartitions in multiple shared processor pools above the entitled pool capacity, the POWER Hypervisor uses a higher level of redistribution, $Level_1$.

Level1 capacity resolution: When allocating additional processor capacity in excess of the entitled pool capacity of the shared processor pool, the POWER Hypervisor takes the uncapped weights of *all micropartitions in the system* into account, *regardless of the multiple shared processor pool structure*.

Where there is unused processor capacity in underutilized shared processor pools, the micropartitions within the shared processor pools cede the capacity to the POWER Hypervisor.

In busy shared processor pools, where the micropartitions use all of the entitled pool capacity, the POWER Hypervisor allocates additional cycles to micropartitions, in which *all* of the following statements are true:

- The maximum pool capacity of the shared processor pool that hosts the micropartition is not met.
- The micropartition is uncapped.
- The micropartition has enough virtual processors to take advantage of the additional capacity.

Under these circumstances, the POWER Hypervisor allocates additional processor capacity to micropartitions on the basis of their uncapped weights independent of the shared processor pool hosting the micropartitions. This can be referred to as $Level_1$ capacity resolution. Consequently, when allocating additional processor capacity in excess of the entitled pool capacity of the shared processor pools, the POWER Hypervisor takes the uncapped weights of all micropartitions in the system into account, regardless of the multiple shared processor pool structure.

Dynamic adjustment of maximum pool capacity

The maximum pool capacity of a shared processor pool, other than the default shared processor $pool_0$, can be adjusted dynamically from the managed console, using either the graphical interface or the command-line interface (CLI).

Dynamic adjustment of reserved pool capacity

The reserved pool capacity of a shared processor pool, other than the default shared processor $pool_0$, can be adjusted dynamically from the managed console, using either the graphical interface or the CLI.

Dynamic movement between shared processor pools

A micropartition can be moved dynamically from one shared processor pool to another using the managed console using either the graphical interface or the CLI. Because the entitled pool capacity is partly made up of the sum of the entitled capacities of the micropartitions, removing a micropartition from a shared processor pool reduces the entitled pool capacity for that shared processor pool. Similarly, the entitled pool capacity of the shared processor pool that the micropartition joins will increase.

Deleting a shared processor pool

Shared processor pools cannot be deleted from the system. However, they are deactivated by setting the maximum pool capacity and the reserved pool capacity to zero. The shared processor pool still exists but not active. Use the managed console interface to deactivate a shared processor pool. A shared processor pool cannot be deactivated unless all micropartitions that are hosted by the shared processor pool are removed.

Live Partition Mobility and multiple shared processor pools

A micropartition can leave a shared processor pool because of PowerVM Live Partition Mobility. Similarly, a micropartition can join a shared processor pool in the same way. When performing PowerVM Live Partition Mobility, you are given the opportunity to designate a destination shared processor pool on the target server to receive and host the migrating micropartition.

Because several simultaneous micropartition migrations are supported by PowerVM Live Partition Mobility, migrating the entire shared processor pool from one server to another is conceivable.

3.4.4 Virtual I/O Server (VIOS)

The Virtual I/O Server is part of all PowerVM Editions. It is a special-purpose partition that allows the sharing of physical resources between logical partitions to allow more efficient utilization (for example, consolidation). In this case, the Virtual I/O Server owns the physical resources (SCSI, Fibre Channel, network adapters, and optical devices) and allows client partitions to share access to them, thus minimizing the number of physical adapters in the system. The Virtual I/O Server eliminates the requirement that every partition owns a dedicated network adapter, disk adapter, and disk drive. The Virtual I/O Server supports OpenSSH for secure remote logins. It also provides a firewall for limiting access by ports, network services, and IP addresses. Figure 3-10 shows an overview of a Virtual I/O Server configuration.

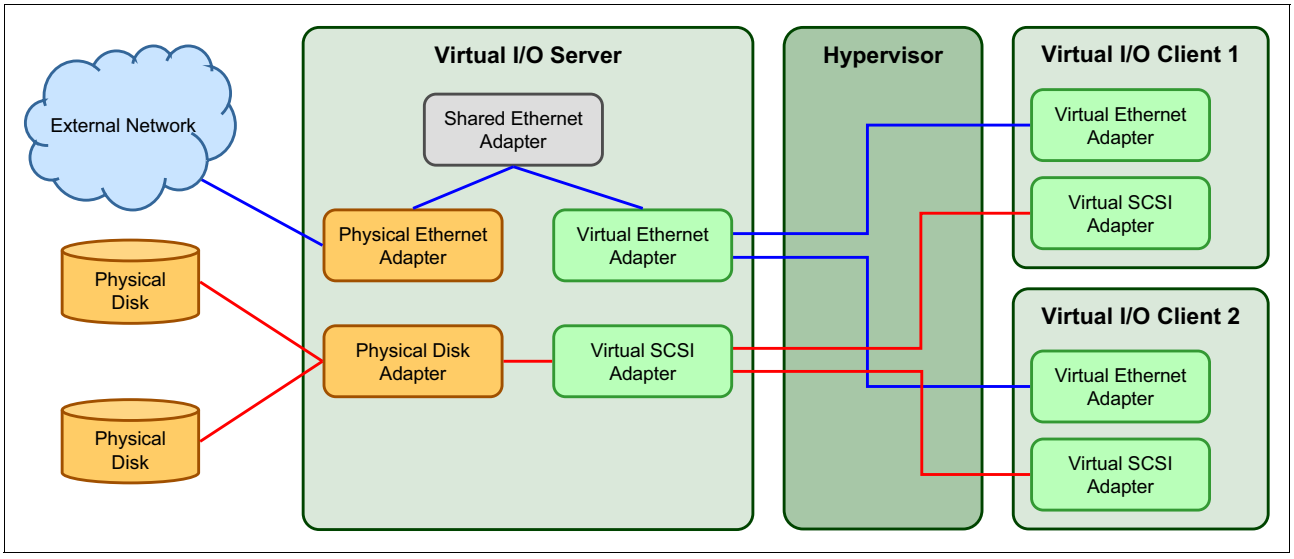

Figure 3-10 Architectural view of the Virtual I/O Server

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 who order any PowerVM edition. This dedicated software is only for the Virtual I/O Server (and IVM in case it is used) and is supported only in special Virtual I/O Server partitions.

Three major virtual devices are supported by the Virtual I/O Server:

- Shared Ethernet Adapter (see “Shared Ethernet Adapter” on page 146)
- Virtual SCSI (see “Virtual SCSI” on page 148)
- Virtual Fibre Channel adapter (this adapter is used with the NPIV feature, described in “N_Port ID virtualization (NPIV)” on page 157)

Shared Ethernet Adapter

A Shared Ethernet Adapter (SEA) can be used to connect a physical Ethernet network to a virtual Ethernet network. The Shared Ethernet Adapter provides this access by connecting the internal hypervisor VLANs with the VLANs on the external switches. Because the Shared Ethernet Adapter 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 Shared Ethernet Adapter also provides the ability for several client partitions to share one physical adapter. With a Shared Ethernet Adapter, you can connect internal and external VLANs using a physical adapter. The Shared Ethernet Adapter service can be hosted only in the Virtual I/O Server, not in a general-purpose AIX or Linux 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 partition can provide bridging function also, by using the **`brctl`** command.

Figure 3-11 shows a configuration example of a Shared Ethernet Adapter with one physical and two virtual Ethernet adapters. A Shared Ethernet Adapter can include up to 16 virtual Ethernet adapters on the Virtual I/O Server that share the same physical access.

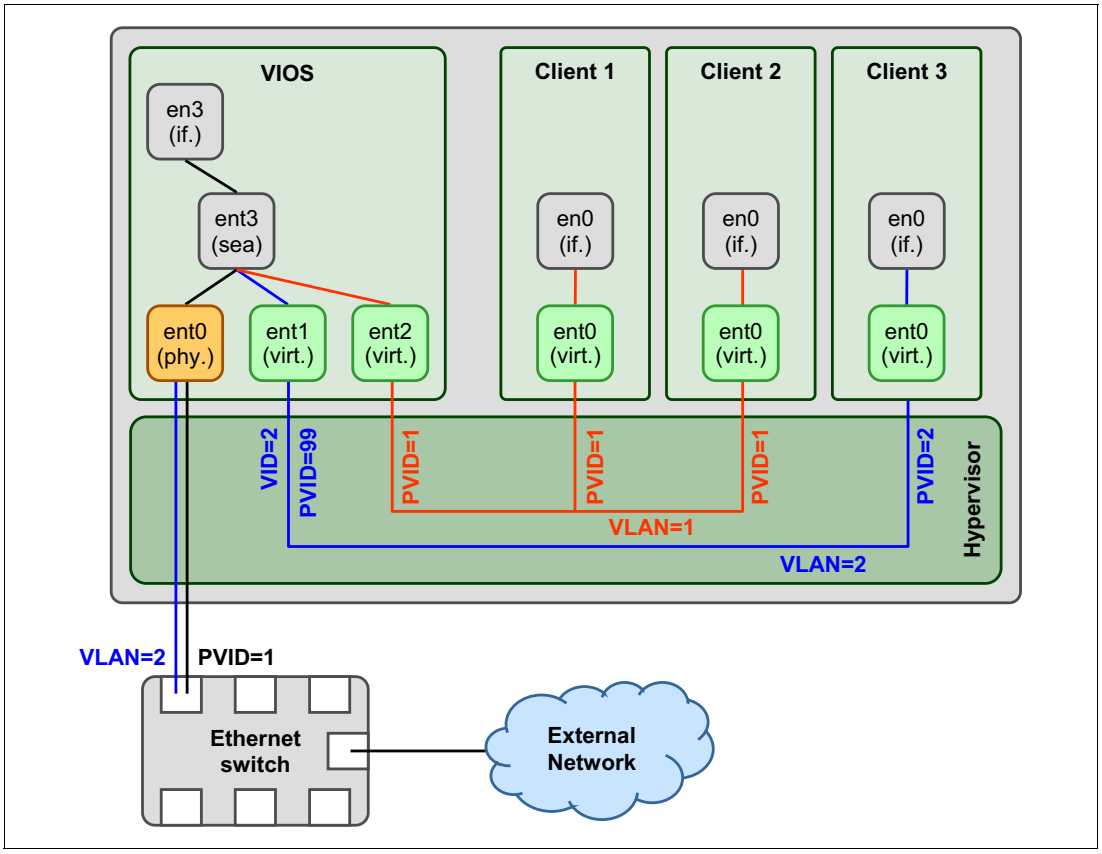

Figure 3-11 Architectural view of a Shared Ethernet Adapter

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

Unicast, broadcast, and multicast are 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 on a Shared Ethernet Adapter.

Configuration: A Shared Ethernet Adapter does not require a configured IP address to be able to perform the Ethernet bridging functionality. Configuring IP on the Virtual I/O Server is convenient because the Virtual I/O Server can then be reached by TCP/IP, for example, to perform dynamic LPAR operations or to enable remote login. This task can be done either by configuring an IP address directly on the Shared Ethernet Adapter device or on an additional virtual Ethernet adapter in the Virtual I/O Server. This approach leaves the Shared Ethernet Adapter without the IP address, allowing for maintenance on the Shared Ethernet Adapter without losing IP connectivity in case Shared Ethernet Adapter failover is configured.

Virtual SCSI

Virtual SCSI is used to refer 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 the physical resources and acts as a server or, in SCSI terms, a target device. The client logical partitions access the virtual SCSI backing storage devices provided by the Virtual I/O Server as clients.

The virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured by using a managed console or through the Integrated Virtualization Manager on smaller systems. The virtual SCSI server (target) adapter is responsible for executing any SCSI commands that it receives. It is owned by the Virtual I/O Server partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the Virtual I/O Server.

Physical disks that are presented to the Virtual/O Server can be exported and assigned to a client partition in a number of 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 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 can be assigned to separate partitions. Therefore, virtual SCSI enables sharing of adapters and disk devices.

Figure 3-12 shows an example where one physical disk is divided into two logical volumes by the Virtual I/O Server. Each client partition 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*.

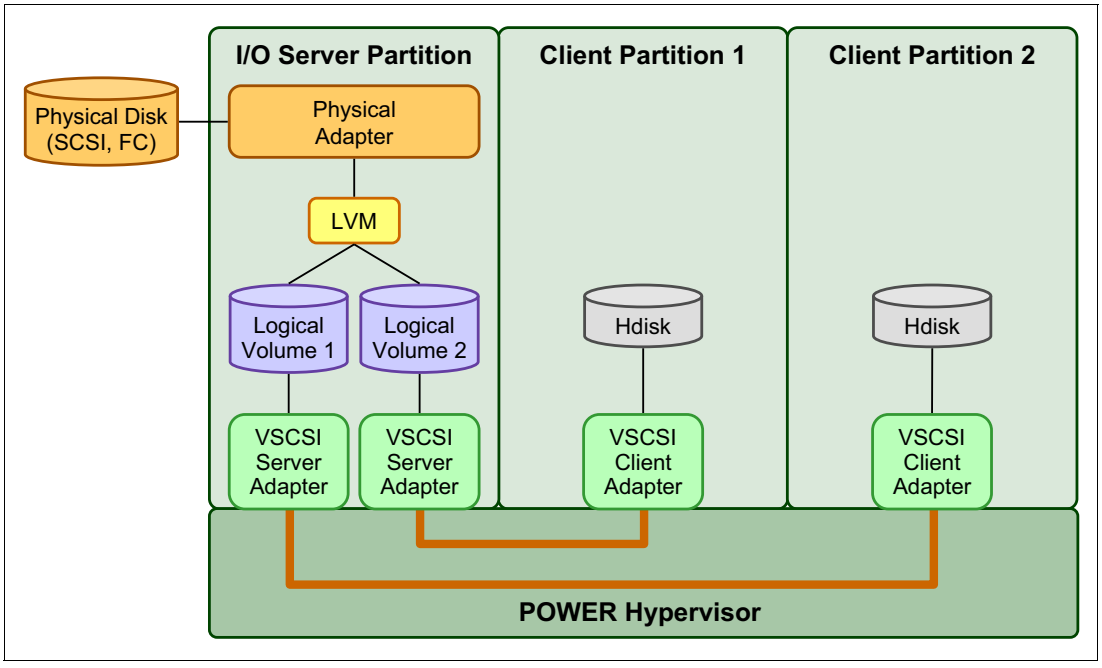

Figure 3-12 Architectural view of virtual SCSI

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 the specific storage devices supported for Virtual I/O Server, see the following location:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html

Virtual I/O Server functions

The Virtual I/O Server has a number of features, including monitoring solutions:

- Support for Live Partition Mobility starting on POWER6 processor-based systems with the PowerVM Enterprise Edition.

  For more information about Live Partition Mobility, see 3.4.5, “PowerVM Live Partition Mobility” on page 150.

- Support for virtual SCSI devices backed by a file that are then accessed as standard SCSI-compliant LUNs
- Support for virtual Fibre Channel devices that are used with the NPIV feature
- Virtual I/O Server Expansion Pack with additional security functions such as Kerberos (Network Authentication Service for users and client and server applications), Simple Network Management Protocol (SNMP) v3, and Lightweight Directory Access Protocol (LDAP) client functionality
- System Planning Tool (SPT) and Workload Estimator, which are designed to ease the deployment of a virtualized infrastructure. For more information about the System Planning Tool, see 3.5, “System Planning Tool” on page 160.
- Includes IBM Systems Director agent and a number of preinstalled IBM Tivoli® agents:
  - Tivoli Identity Manager, to allow easy integration into an existing Tivoli Systems Management infrastructure
  - Tivoli Application Dependency Discovery Manager (ADDM), which creates and automatically maintains application infrastructure maps including dependencies, change-histories, and deep configuration values
- vSCSI eRAS
- Additional CLI statistics in **svmon**, **vmstat**, **fcstat**, and **topas**.
- Monitoring solutions to help manage and monitor the Virtual I/O Server and shared resources

  New commands and views provide additional metrics for memory, paging, processes, Fibre Channel HBA statistics, and virtualization.

For more information about the Virtual I/O Server and its implementation, see *PowerVM Virtualization on IBM System p: Introduction and Configuration Fourth Edition*, SG24-7940.

3.4.5 PowerVM Live Partition Mobility

With PowerVM Live Partition Mobility, you can move a running logical partition, including its operating system and running applications, from one system to another without any shutdown or without disrupting the operation of that logical partition. Inactive partition mobility allows you to move a powered-off logical partition from one system to another.

Partition mobility provides systems management flexibility and improves system availability, as follows:

- 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 maintenance because you can use it to work around scheduled maintenance activities.
- Avoid downtime for a server upgrade by moving logical partitions to another server and then performing the upgrade. With this approach, your users can continue their work without disruption.
- Avoid unplanned downtime. With preventive failure management, if a server indicates a potential failure, you can move its logical partitions to another server before the failure occurs. Partition mobility can help avoid unplanned downtime.
- Take advantage of server optimization:
  - Consolidation: You can consolidate workloads running on several small, under-used servers onto a single large 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 operating system requirements

The operating system running in the mobile partition has to be AIX or Linux. The Virtual I/O Server partition itself cannot be migrated. All versions of AIX and Linux that are supported on the IBM POWER7 processor-based servers also support partition mobility.

Source and destination system requirements

The source partition must be one that has only virtual devices. If any physical devices are in its allocation, they must be removed before the validation or migration is initiated. An N_Port ID virtualization (NPIV) device is considered virtual and is compatible with partition migration.

The hypervisor must support the partition mobility functionality (also called migration process) that is available on POWER6 and POWER7 processor-based hypervisors. Firmware must be at firmware level eFW3.2 or later. All POWER7 processor-based hypervisors support Live Partition Mobility. Source and destination systems can have separate firmware levels, but they must be compatible with each other.

A possibility is to migrate partitions back and forth between POWER6 and POWER7 processor-based servers. Partition mobility uses the POWER6 compatibility modes that are provided by POWER7 processor-based servers. On the POWER7 processor-based server, the migrated partition is then executing in POWER6 compatibility mode.

If you want to move an active logical partition from a POWER6 processor-based server to a POWER7 processor-based server so that the logical partition can take advantage of the additional capabilities available with the POWER7 processor, complete the following steps:

1. Set the partition-preferred processor compatibility mode to the default mode.

   When you activate the logical partition on the POWER6 processor-based server, it runs in the POWER6 mode.

2. Move the logical partition to the POWER7 processor-based server.

   Both the current and preferred modes remain unchanged for the logical partition until you restart the logical partition.

3. Restart the logical partition on the POWER7 processor-based server.

   The hypervisor evaluates the configuration. Because the preferred mode is set to default and the logical partition now runs on a POWER7 processor-based server, the highest mode available is the POWER7 mode. The hypervisor determines that the most fully featured mode that is supported by the operating environment installed in the logical partition is the POWER7 mode and changes the current mode of the logical partition to the POWER7 mode.

Now the current processor compatibility mode of the logical partition is the POWER7 mode, and the logical partition runs on the POWER7 processor-based server.

Tip: The “Migration combinations of processor compatibility modes for active Partition Mobility” web page offers presentations of the supported migrations:

`http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3pcmcombosact.htm`

The Virtual I/O Server on the source system provides the access to the client resources and must be 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 managed console and will be configured on both the source and destination Virtual I/O Servers, which are designated as the mover service partitions for the mobile partition, to participate in active mobility. Other requirements include a similar time-of-day on each server, systems must not be running on battery power, and shared storage (external hdisk with `reserve_policy=no_reserve`). In addition, all logical partitions must be on the same open network with RMC established to the managed console.

The managed console is used to configure, validate, and orchestrate. Use it to configure the Virtual I/O Server as an MSP and to configure the VASI device. A managed console wizard validates your configuration and identifies issues that can cause the migration to fail. During the migration, the managed console controls all phases of the process.

Improved Live Partition Mobility benefits

The possibility to move partitions between POWER6 and POWER7 processor-based servers greatly facilitates the deployment of POWER7 processor-based servers, as follows:

- Installation of the new server can be performed while the application is executing on a POWER6. After the POWER7 processor-based server is ready, the application can be migrated to its new hosting server without application down time.
- When adding POWER7 processor-based servers to a POWER6 environment, you get the additional flexibility to perform workload balancing across the entire set of POWER6 and POWER7 processor-based servers.
- When performing server maintenance, you get the additional flexibility to use POWER6 Servers for hosting applications usually hosted on POWER7 processor-based servers, and vice versa, allowing you to perform this maintenance with no application planned down time.

For more information about Live Partition Mobility and how to implement it, see *IBM PowerVM Live Partition Mobility*, SG24-7460.

3.4.6 Active Memory Sharing

Active Memory Sharing is an IBM PowerVM advanced memory virtualization technology that provides system memory virtualization capabilities to IBM Power Systems, allowing multiple partitions to share a common pool of physical memory.

Active Memory Sharing is available only with the Enterprise version of PowerVM.

The physical memory of an IBM Power Systems can be assigned to multiple partitions in either dedicated or shared mode. The system administrator has the capability to assign some physical memory to a partition and some physical memory to a pool that is shared by other partitions. A single partition can have either dedicated or shared memory:

- With a pure dedicated memory model, the system administrator's task is to optimize available memory distribution among partitions. When a partition suffers degradation because of memory constraints and other partitions have unused memory, the administrator can manually issue a dynamic memory reconfiguration.
- With a shared memory model, the system automatically decides the optimal distribution of the physical memory to partitions and adjusts the memory assignment based on partition load. The administrator reserves physical memory for the shared memory pool, assigns partitions to the pool, and provides access limits to the pool.

Active Memory Sharing can be exploited to increase memory utilization on the system either by decreasing the global memory requirement or by allowing the creation of additional partitions on an existing system. Active Memory Sharing can be used in parallel with Active Memory Expansion on a system running a mixed workload of several operating system. For example, AIX partitions can take advantage of Active Memory Expansion. Other operating systems take advantage of Active Memory Sharing also.

For additional information regarding Active Memory Sharing, see *PowerVM Virtualization Active Memory Sharing*, REDP-4470.

3.4.7 Active Memory Deduplication

In a virtualized environment, the systems might have a considerable amount of duplicated information stored on RAM after each partition has its own operating system, and some of them might even share the same kind of applications. On heavily loaded systems this might lead to a shortage of the available memory resources, forcing paging by the Active Memory Sharing partition operating systems, the Active Memory Deduplication pool, or both, which might decrease overall system performance.

Figure 3-13 shows the standard behavior of a system without Active Memory Deduplication enabled on its Active Memory Sharing shared memory pool. Identical pages within the same or different LPARs each require their own unique physical memory page, consuming space with repeated information.

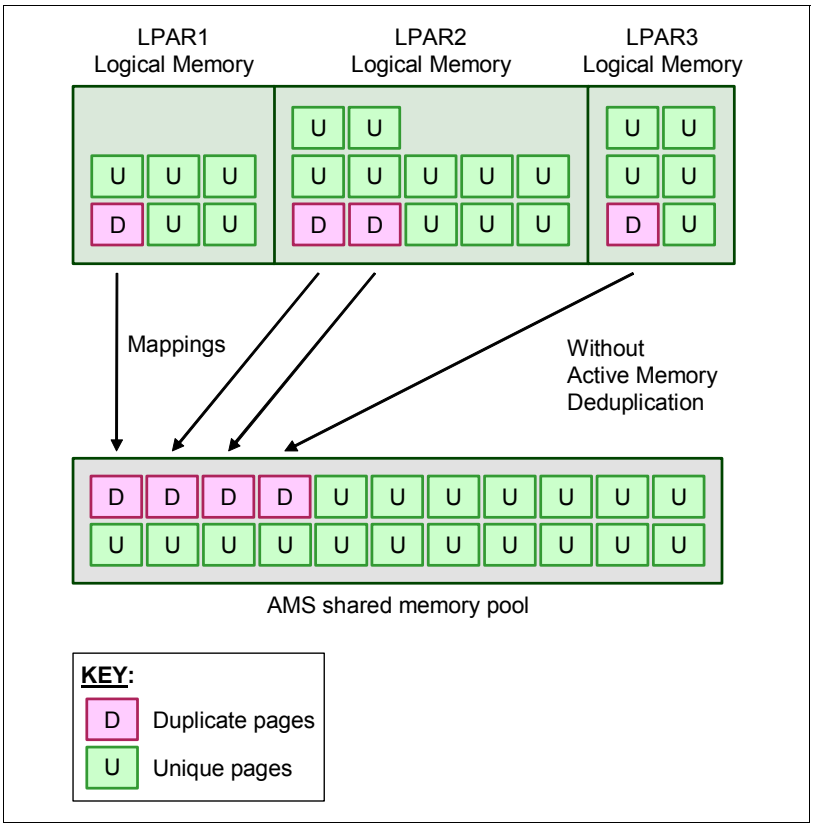

Figure 3-13 Active Memory Sharing shared memory pool without Active Memory Deduplication enabled

Active Memory Deduplication allows the hypervisor to dynamically map identical partition memory pages to a single physical memory page within a shared memory pool. This feature enables a better utilization of the Active Memory Sharing shared memory pool, increasing the system's overall performance by avoiding paging. Deduplication can cause the hardware to incur fewer cache misses, which will also lead to improved performance.

Figure 3-14 shows the behavior of a system with Active Memory Deduplication enabled on its Active Memory Sharing shared memory pool. Duplicated pages from different LPARs are stored just once, providing the Active Memory Sharing pool with more free memory.

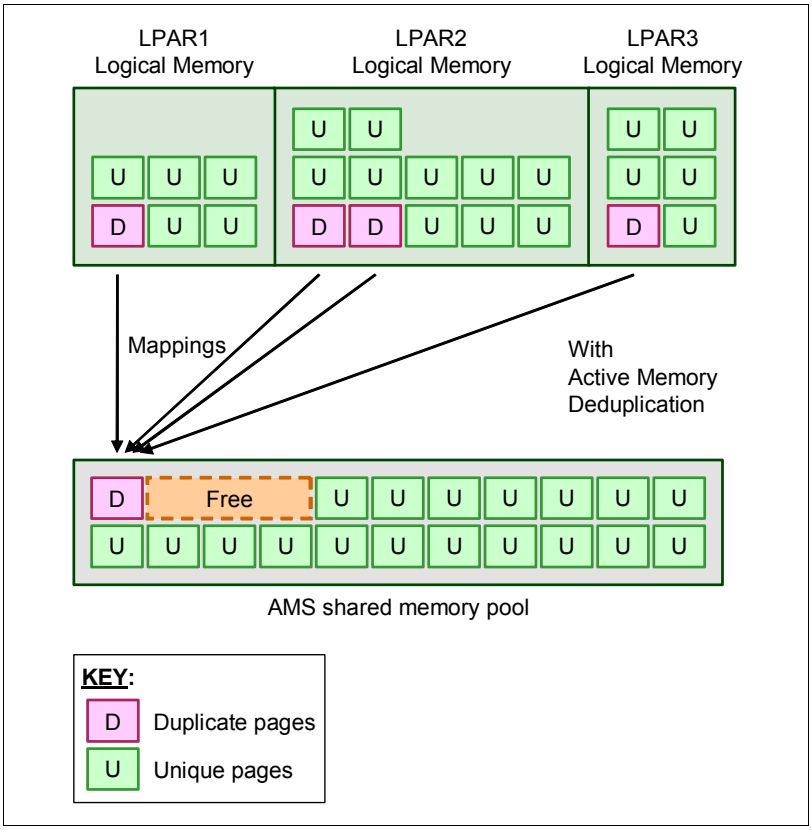

Figure 3-14 Identical memory pages mapped to a single physical memory page with Active Memory Duplication enabled

Active Memory Deduplication depends on the Active Memory Sharing feature to be available, and consumes CPU cycles donated by the Active Memory Sharing pool's VIOS partitions to identify deduplicated pages. The operating systems running on the Active Memory Sharing partitions can hint to the PowerVM Hypervisor that some pages (such as frequently referenced read-only code pages) are particularly good for deduplication.

To perform deduplication, the hypervisor cannot compare every memory page in the Active Memory Sharing pool with every other page. Instead, it computes a small signature for each page that it visits and stores the signatures in an internal table. Each time that a page is inspected, its signature is looked up against the known signatures in the table. If a match is found, the memory pages are compared to be sure that the pages are really duplicates. When a duplicate is found, the hypervisor remaps the partition memory to the existing memory page and returns the duplicate page to the Active Memory Sharing pool.

Figure 3-15 shows two pages being written in the Active Memory Sharing memory pool and having their signatures matched on the deduplication table.

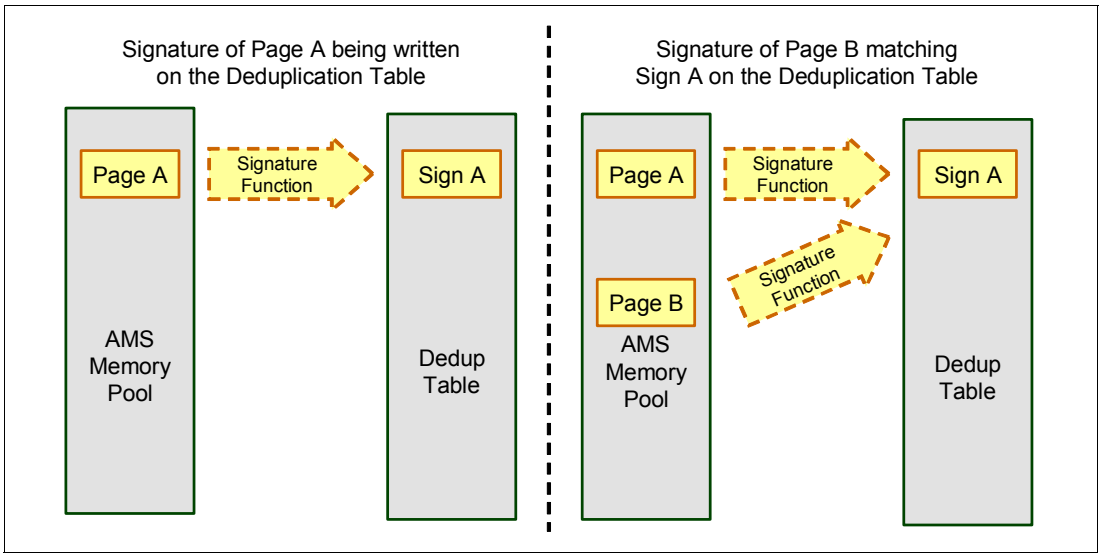

Figure 3-15 Memory pages having their signatures matched by Active Memory Deduplication

From the LPAR point of view, the Active Memory Deduplication feature is completely transparent. If an LPAR attempts to modify a deduplicated page, the hypervisor grabs a free page from the Active Memory Sharing pool, copies the duplicate page contents into the new page, and maps the LPAR's reference to the new page so that the LPAR can modify its own unique page.

System administrators can dynamically configure the size of the deduplication table, ranging from 1/8192 up to 1/256 of the configured maximum Active Memory Sharing memory pool size. Having this table too small might lead to missed deduplication opportunities. Conversely, having a table that is too large might waste a small amount of overhead space.

The management of the Active Memory Deduplication feature is done through a managed console, allowing administrators to take the following steps:

- Enable and disable Active Memory Deduplication at an Active Memory Sharing Pool level.
- Display deduplication metrics.
- Display and modify the deduplication table size.

Figure 3-16 shows the Active Memory Deduplication being enabled to a shared memory pool.

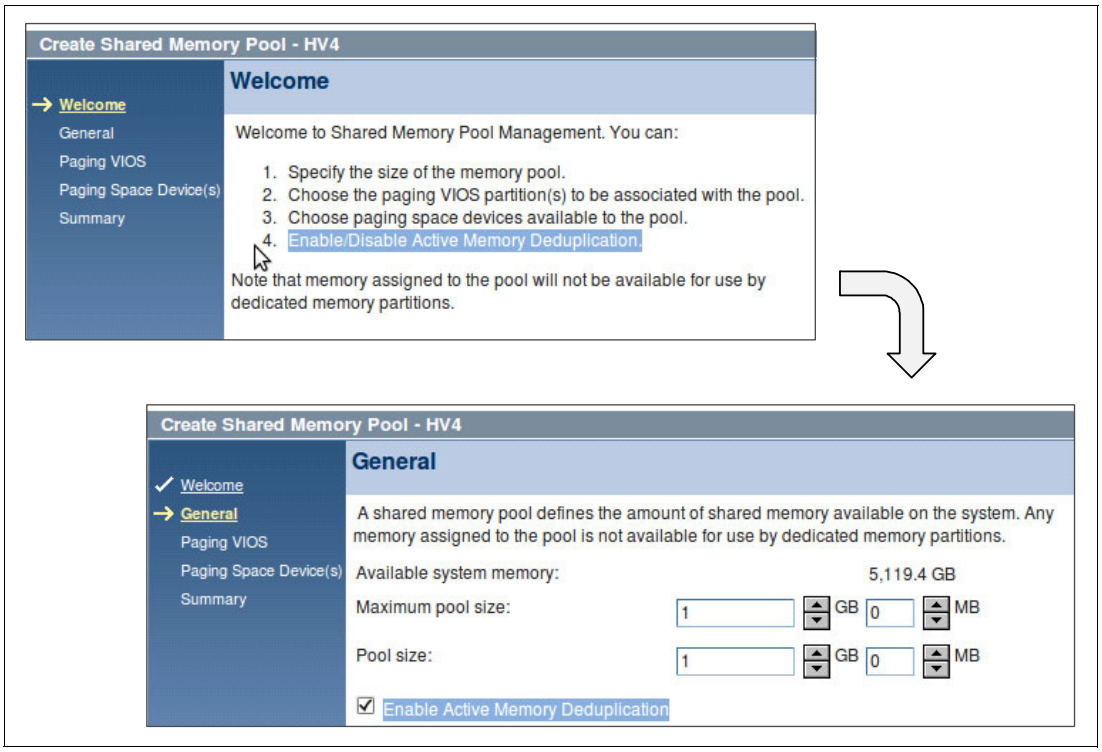

Figure 3-16 Enabling the Active Memory Deduplication for a shared memory pool

The Active Memory Deduplication feature requires the following minimum components:

- PowerVM Enterprise edition
- System firmware level 740
- AIX Version 6: AIX 6.1 TL7 or later
- AIX Version 7: AIX 7.1 TL1 SP1 or later
- IBM i: 7.14 or 7.2 or later
- SLES 11 SP2 or later
- RHEL 6.2 or later

3.4.8 Dynamic Platform Optimizer

Dynamic Platform Optimizer (DPO, FC EB33) is an IBM PowerVM feature that helps the user to configure the Logical Partition memory and CPU affinity on the new POWER7 processor-based servers, thus, improve performance under some workload scenarios.

In a nonuniform memory access (NUMA) context, the main goal of the DPO is to assign a local memory to the CPUs, thus, reducing the memory access time, because a local memory access is much faster than a remote access.

Accessing remote memory on a NUMA environment is expensive, although common, mainly if the system did a partition migration, or even, if logical partitions are created, suspended, and destroyed frequently, as it happens frequently in a cloud environment. In this context, DPO tries to swap remote memory by local memory to the CPU.

Dynamic Platform Optimizer should be launched and stopped by using the HMC command-line interface with the **`optmem`** tool (see Example 3-1 on page 157). Also, the

`lsoptmem` tool is able to show important information about logical partition CPU and memory affinity, as monitoring the status of a running optimization process.

Example 3-1 Launching DPO for a LPAR 1

```
#optmem -m managed_system -t affinity --id 1 -o start
```

TIP: While the DPO process is running, the specified LPAR can have 10 - 20% of performance degradation.

N_Port ID virtualization (NPIV)

NPIV technology allows multiple logical partitions to access independent physical storage through the same physical Fibre Channel adapter. This adapter is attached to a Virtual I/O Server partition that acts only as a pass-through, managing the data transfer through the POWER Hypervisor.

Each partition using NPIV is identified by a pair of unique worldwide port names, enabling you to connect each partition to independent physical storage on a SAN. Unlike virtual SCSI, only the client partitions see the disk.

For additional information and requirements for NPIV, see the following resources:

- *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

Support: NPIV is supported in PowerVM Standard and Enterprise Editions on the IBM Power 795 servers.

3.4.9 Operating system support for PowerVM

Table 3-5 summarizes the PowerVM features supported by the operating systems compatible with the POWER7 processor-based servers.

Table 3-5 Virtualization features supported by AIX, IBM i and Linux

Feature	AIX 5.3	AIX 6.1	AIX 7.1	IBM i 6.1.1	IBM i 7.1	RHEL 5.8	RHEL 6.3	SLES 10 SP4	SLES 11 SP2
Virtual SCSI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Ethernet	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared Ethernet Adapter	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Fibre Channel	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Tape	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Logical partitioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dynamic LPAR I/O adapter add or remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dynamic LPAR processor add or remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dynamic LPAR memory add	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Feature	AIX 5.3	AIX 6.1	AIX 7.1	IBM i 6.1.1	IBM i 7.1	RHEL 5.8	RHEL 6.3	SLES 10 SP4	SLES 11 SP2
Dynamic LPAR memory remove	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Micro-Partitioning	Yes	Yes	Yes	Yes	Yes	Yes[a]	Yes[b]	Yes[a]	Yes
Shared dedicated capacity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Shared Processor Pools	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual I/O Server	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Integrated Virtualization Manager	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Suspend and resume	No	Yes	Yes	No	Yes[c]	Yes	Yes	No	No
Shared Storage Pools	Yes	Yes	Yes	Yes	Yes[d]	Yes	Yes	Yes	No
Thin provisioning	Yes	Yes	Yes	Yes[e]	Yes[e]	Yes	Yes	Yes	No
Active Memory Sharing	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Active Memory Deduplication	No	Yes[f]	Yes[g]	No	Yes[h]	No	Yes	No	Yes
Live Partition Mobility	Yes	Yes	Yes	No	Yes[i]	Yes	Yes	Yes	Yes
Simultaneous multithreading (SMT)	Yes[j]	Yes[k]	Yes	Yes[l]	Yes	Yes[j]	Yes	Yes[j]	Yes
Active Memory Expansion	No	Yes[m]	Yes	No	No	No	No	No	No
Capacity on Demand[n]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIX Workload Partitions	No	Yes	Yes	No	No	No	No	No	No

a. This version can only support 10 virtual machines per core.
b. Need RHEL 6.3 Errata upgrade to support 20 virtual machines per core.
c. Requires IBM i 7.1 TR2 with PTF SI39077 or later.
d. Requires IBM i 7.1 TR1.
e. Will become a fully provisioned device when used by IBM i.
f. Requires AIX 6.1 TL7 or later.
g. Requires AIX 7.1 TL1 or later.
h. Requires IBM i 7.1.4 or later.
i. Requires IBM i 7.1 TR4 PTF group or later. Access the following link for more details:
http://www-912.ibm.com/s_dir/SLKBase.nsf/1ac66549a21402188625680b0002037e/e1877ed7f3b0cfa8862579ec0048e067?OpenDocument#_Section1
j. Supports only two threads.
k. AIX 6.1 up to TL4 SP2 only supports two threads, and supports four threads as of TL4 SP3.
l. IBM i 6.1.1 and up support SMT4.
m. On AIX 6.1 with TL4 SP2 and later.
n. Available on selected models.

3.4.10 Linux support

IBM Linux Technology Center (LTC) contributes to the development of Linux by providing support for IBM hardware in Linux distributions. In particular, the LTC makes tools and code available to the Linux communities to take advantage of the POWER7 technology and develop POWER7 optimized software.

Table 3-6 lists the support of specific programming features for various versions of Linux.

Table 3-6 Linux support for POWER7 features

Features	Linux releases				Comments
	SLES 10 SP4	SLES 11 SP2	RHEL 5.8	RHEL 6.3	
POWER6 compatibility mode	Yes	Yes	Yes	Yes	—
POWER7 mode	No	Yes	No	Yes	Take advantage of the POWER7 features.
Strong Access Ordering	No	Yes	No	Yes	Can improve emulation performance
Scale to 256 cores/ 1024 threads	No	Yes	No	Yes	Base OS support available
4-way SMT	No	Yes	No	Yes	Better hardware usage
VSX support	No	Yes	No	Yes	Full exploitation requires Advance Toolchain.
Distro Toolchain mcpu/mtune=p7	No	Yes	No	Yes	SLES11/GA Toolchain has minimal P7 enablement necessary to support kernel build
Advance Toolchain support	Yes, execution is restricted to Power6 instructions	Yes	Yes, execution is restricted to Power6 instructions	Yes	Alternative GNU Toolchain that explores the new technologies available on POWER architecture
64k base page size	No	Yes	Yes	Yes	Better memory utilization, and smaller footprint.
Tickless idle	No	Yes	No	Yes	Improved energy utilization and virtualization of partially to fully idle partitions

See the following sources of information:

- Advance Toolchain:

  `http://www.ibm.com/developerworks/wikis/display/hpccentral/How+to+use+Advance+Toolchain+for+Linux+on+POWER`

- University of Illinois Linux on Power Open Source Repository:

  `http://ppclinux.ncsa.illinois.edu`

- Release notes:
  - `ftp://linuxpatch.ncsa.uiuc.edu/toolchain/at/at05/suse/SLES_11/release_notes.at05-2.1-0.html`
  - `ftp://linuxpatch.ncsa.uiuc.edu/toolchain/at/at05/redhat/RHEL5/release_notes.at05-2.1-0.html`

3.5 System Planning Tool

The IBM System Planning Tool (SPT) helps you design systems to be partitioned with logical partitions. You can also plan for and design non-partitioned systems by 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 reasons:

- To create reports
- As input to the IBM configuration tool (e-Config)
- To create and deploy partitions on your system (or systems) automatically

System plans that are generated by the SPT can be deployed on the system by the Hardware Management Console (HMC), or Integrated Virtualization Manager (IVM).

Automatically deploy: Ask your IBM representative or IBM Business Partner to use the Customer Specified Placement manufacturing option if you want to automatically deploy your partitioning environment on a new machine. SPT verifies that the resource's allocation is the same as that specified in your `.sysplan` file.

You can create an entirely new system configuration, or you can create a system configuration based on any of these items:

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

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

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

Using the SPT is an effective way of documenting and backing up key system settings and partition definitions. With it, the user can create records of systems and export them to the personal workstation or backup system of choice. These same backups can then be imported back onto the same managed console when needed. This step can be useful when cloning systems, enabling the user to import the system plan to any managed console multiple times.

The SPT and its supporting documentation is on the IBM System Planning Tool site:

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

3.6 POWER Version 2.2 enhancements

The latest available PowerVM Version 2.2 contains the following enhancements:

- Up to 20 logical partitions per core.
- Role Based Access Control (RBAC):

  RBAC brings an added level of security and flexibility in the administration of VIOS. With RBAC, you can create a set of authorizations for the user management commands. You can assign these authorizations to the *UserManagement* role, and this role can be given to

any other user. So a normal user with the role UserManagement can manage the users on the system but will not have any further access.

With RBAC, the Virtual I/O Server can split management functions that presently can be done only by the *padmin* user, provide better security by providing only the necessary access to users, and easy management and auditing of system functions.

- Support for concurrent adding of VLANs
- Support for USB tape

  The Virtual I/O Server now supports a USB DAT-320 Tape Drive and its use as a virtual tape device for VIOS clients.
- Support for USB Blu-ray

  The Virtual I/O Server now supports USB Blu-ray optical devices. AIX does not support mapping these as virtual optical devices to clients. However, you can import the disk in to the virtual optical media library and map the created file to the client as a virtual DVD drive.

The IBM PowerVM IBM Workload Partitions Manager™ for AIX, Version 2.2 has the following enhancements:

- When used with AIX 6.1 Technology Level 6, the following support applies:
  - Support for exporting VIOS SCSI disk into a WPAR. Compatibility analysis and mobility of WPARs with VIOS SCSI disk. In addition to Fibre Channel devices, now VIOS SCSI disks can be exported into a workload partition (WPAR).
  - WPAR Manager command-line interface (CLI). The WPAR Manager CLI allows federated management of WPARs across multiple systems by command line.
  - Support for workload partition definitions. The WPAR definitions can be preserved after WPARs are deleted. These definitions can be deployed at a later time to any WPAR-capable system.
- In addition to the feature supported on AIX 6.1 Technology Level 6, the following support applies to AIX 7.1:
  - Support for AIX 5.2 Workload Partitions for AIX 7.1. Lifecycle management and mobility enablement for AIX 5.2 Technology Level 10 SP8 Version WPARs.
  - Support for trusted kernel extension loading and configuration from WPARs. Enables exporting a list of kernel extensions that can then be loaded inside a WPAR, yet maintaining isolation.

4

Continuous availability and manageability

This chapter provides information about IBM reliability, availability, and serviceability (RAS) design and features. This set of technologies, implemented on IBM Power Systems servers, provides the possibility to improve your architecture's total cost of ownership (TCO) by reducing unplanned down time.

The elements of RAS can be described as follows:

- Reliability: Indicates how infrequently a defect or fault in a server manifests itself.
- Availability: Indicates how infrequently the functionality of a system or application is affected by a fault or defect.
- Serviceability: Indicates how well faults and their effects are communicated to users and services, and how efficiently and nondisruptively the faults are repaired.

Each successive generation of IBM servers is designed to be more reliable than the previous server family. POWER7 processor-based servers have new features to support new levels of virtualization, help ease administrative burden and increase system utilization.

Reliability starts with components, devices, and subsystems designed to be fault-tolerant. POWER7 uses lower voltage technology improving reliability with stacked latches to reduce soft error susceptibility. During the design and development process, subsystems go through rigorous verification and integration testing processes. During system manufacturing, systems go through a thorough testing process to help ensure high product quality levels.

The processor and memory subsystem contain features designed to avoid or correct environmentally induced, single-bit, intermittent failures and also handle solid faults in components, including selective redundancy to tolerate certain faults without requiring an outage or parts replacement.

4.1 Reliability

Highly reliable systems are built with highly reliable components. On IBM POWER processor-based systems, this basic principle is expanded upon 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 designed with fewer components and interconnects have fewer opportunities to fail. Simple design choices such as integrating processor cores on a single POWER chip can dramatically reduce the opportunity for system failures. In this case, an 8-core server can include one-fourth as many processor chips (and chip socket interfaces) as with a double CPU-per-processor design. Not only does this case reduce the total number of system components, it reduces the total amount of generated heat in the design, resulting in an additional reduction in required power and cooling components. POWER7 processor-based servers also integrate L3 cache into the processor chip for a higher integration of parts.

Parts selection also plays a critical role in overall system reliability. IBM uses three grades of components (1, 3, and 5); grade 3 is defined as the 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.

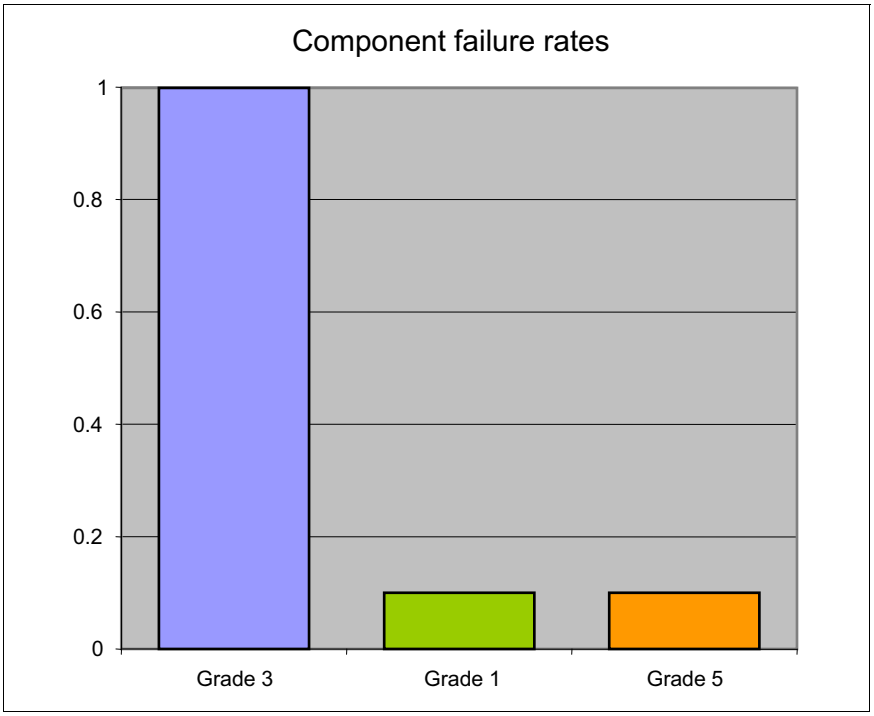

Figure 4-1 Component failure rates

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; large decreases in component reliability are directly correlated with relatively small increases in temperature; POWER processor-based systems are carefully packaged to ensure adequate cooling. Critical system components, such as the POWER7 processor chips, are positioned on printed circuit cards so they receive fresh air during operation. In addition, POWER 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.

4.1.3 Redundant components and concurrent repair

High opportunity components or, those that most affect system availability, are protected with redundancy and the ability to be repaired concurrently. The Power 795 uses redundancy in many areas of the design. The redundant hardware is not optional, but rather is base function of the design.

The Power 795 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)
- Universal power interface connect (UPIC) power distribution and communication cables
- CEC cage:
  - System controller, oscillator card (hot failover)
  - Vital product data (VPD) card (active redundancy)
- Processor books:
  - Node controller (hot failover)
  - VPD and CoD modules; memory DIMM VPD (active redundancy)
  - DCAs (active redundancy)
  - Voltage regulator modules (active redundancy, hot-replace)
  - Book TPMD function (active redundancy)
- Ethernet cabling to node controllers and system controllers
- Blowers (active redundancy, hot-replace)
- All out of band service interfaces (active redundancy)
- All LED indicator drive circuitry (active redundancy)
- Thermal sensors (active redundancy) Processor book temperature sensors, Processor temperature sensors, memory DIMM sensors

Furthermore, Power 795 denotes redundancy in other components:

- POWER7 cores, which include redundant bits in L1 instruction and data caches, L2 caches, and in L2 and L3 directories
- Power 795 main memory DIMMs, which use an innovative ECC algorithm from IBM research that improves bit error correction and memory failures, and also contain extra DRAM chips for improved redundancy
- Redundant, hot-swappable 12X I/O drawer connections

For maximum availability, be sure to connect power cords from the same system to two separate and independent power sources.

Redundant components: Check your configuration for optional redundant components before ordering your system.

4.2 Availability

IBM hardware and microcode capability to continuously monitor execution of hardware functions is generally described as the process of first-failure data capture (FFDC). This process includes the strategy of predictive failure analysis, which refers to the ability to track intermittent correctable errors and to vary components offline before they reach the point of hard failure causing a system outage, and without the need to recreate the problem.

The POWER7 family of systems continues to introduce significant enhancements that are designed to increase system availability and ultimately a high availability objective with hardware components that can do the following functions:

- Self-diagnose and self-correct during run time
- Automatically reconfigure to mitigate potential problems from suspect hardware
- Self-heal or automatically substitute good components for failing components

Note: POWER7 processor-based servers are independent of the operating system for error detection and fault isolation within the CEC.

Throughout this chapter, we describe IBM POWER7 processor-based systems technologies that are focused on keeping a system running. For a specific set of functions that are focused on detecting errors before they become serious enough to stop computing work, see 4.3.1, “Detecting” on page 178.

4.2.1 Partition availability priority

Also available is the ability to assign availability priorities to partitions. If an alternate processor recovery event requires spare processor resources and there are no other means of obtaining the spare resources, the system determines which partition has the lowest priority and attempts to claim the needed resource. On a properly configured POWER processor-based server, this approach allows that capacity to first be obtained from a low priority partition instead of a high priority partition.

This capability is relevant to the total system availability because it gives the system an additional stage before an unplanned outage. In the event that insufficient resources exist to maintain full system availability, these servers attempt to maintain partition availability by user-defined priority.

Partition availability priority is assigned to partitions by using a *weight value* or integer rating. The lowest priority partition is rated at 0 (zero) and the highest priority partition is valued at 255. The default value is set at 127 for standard partitions, and set to 192 for Virtual I/O Server (VIOS) partitions. You can vary the priority of individual partitions.

Partition availability priorities can be set for both dedicated and shared processor partitions. The POWER Hypervisor uses the relative partition weight value among active partitions to favor higher priority partitions for processor sharing, adding and removing processor capacity, and favoring higher priority partitions for normal operation.

Note that the partition specifications for *minimum*, *desired*, and *maximum* capacity are also taken into account for capacity-on-demand options, and if total system-wide processor capacity becomes disabled because of deconfigured failed processor cores. For example, if total system-wide processor capacity is sufficient to run all partitions, at least with the minimum capacity, the partitions are allowed to start or continue running. If processor capacity is insufficient to run a partition at its minimum value, then starting that partition results in an error condition that must be resolved.

4.2.2 General detection and deallocation of 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 POWER processor-based system is flagged for persistent deallocation. Component removal can occur either dynamically (while the system is running) or at boot-time (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 way prevents the same faulty hardware from affecting system operation again; the repair action is deferred to a more convenient, less critical time.

The following persistent deallocation functions are included:

- Processor
- L2/L3 cache lines (cache lines are dynamically deleted)
- Memory
- Deconfigure or bypass failing I/O adapters

The auto-restart (reboot) option must be enabled from the Advanced System Management Interface (ASMI) or the Control (Operator) Panel. Figure 4-2 shows this option in the ASMI.

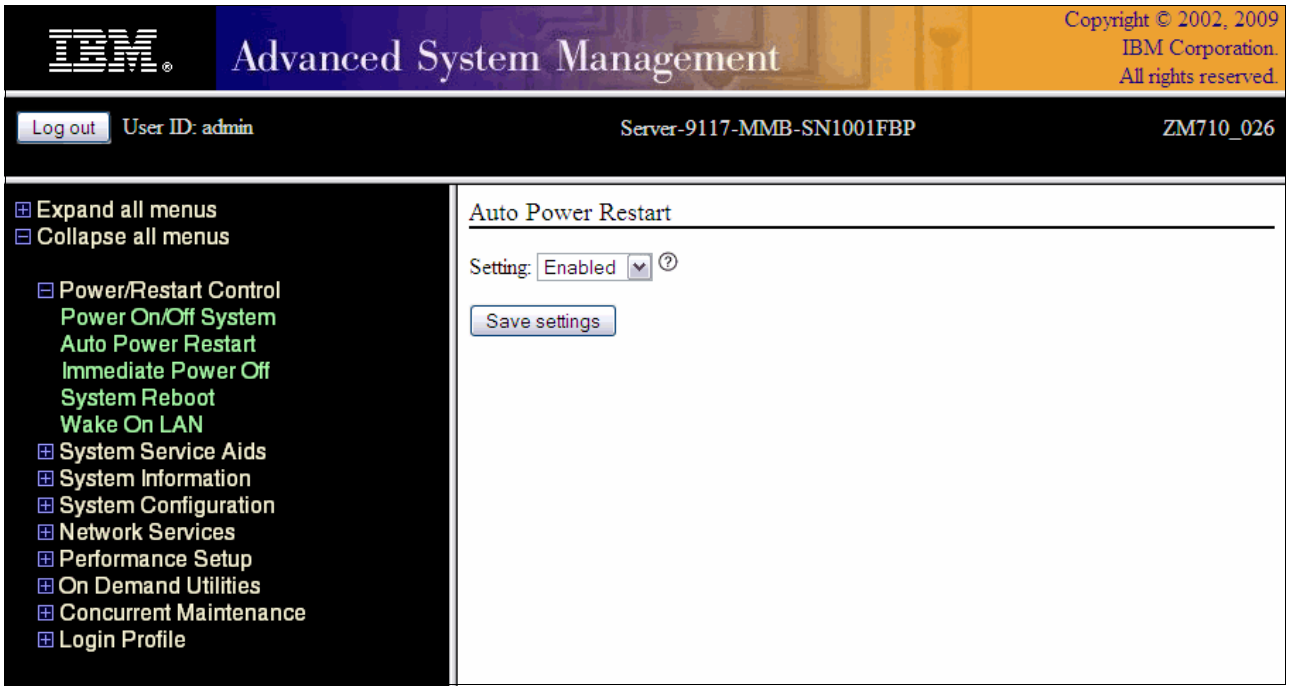

Figure 4-2 AMSI Auto Power Restart setting panel

Processor instruction retry

As in POWER6, the POWER7 processor has the ability to retry processor instruction and alternate processor recovery for a number of core-related faults. Doing this significantly reduces exposure to both permanent and intermittent errors in the processor core.

Intermittent errors, often because of cosmic rays or other sources of radiation, are generally not repeatable.

With the instruction retry function, when an error is encountered in the core, in caches and certain logic functions, the POWER7 processor first automatically retries the instruction. If the source of the error was truly transient, the instruction succeeds and the system can continue as before.

On IBM systems prior to POWER6, such an error typically caused a checkstop.

Alternate processor retry

Hard failures are more difficult, because they are permanent 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.

As in POWER6, POWER7 processors have the ability to extract the failing instruction from the faulty core and retry it elsewhere in the system for a number of faults, after which the failing core is dynamically deconfigured and scheduled for replacement.

Dynamic processor deallocation

Dynamic processor deallocation enables automatic deconfiguration of processor cores when patterns of recoverable core-related faults are detected. Dynamic processor deallocation prevents a recoverable error from escalating to an unrecoverable system error, which might otherwise 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 dynamically deconfigures the failing core and is called out for replacement. The entire process is transparent to the partition owning the failing instruction.

If inactivated processor cores or CoD processor cores are available, the system effectively puts a CoD processor into operation after an activated processor is determined to no longer be operational. In this way, the server remains with its total processor power.

If no CoD processor cores are available system-wide total processor capacity is lowered below the licensed number of cores.

Single processor checkstop

As in the POWER6 processor, the POWER7 processor provides single-core check-stopping for certain processor logic, command, or control errors that cannot be handled by the availability enhancements in the preceding section.

This way significantly reduces the probability of any one processor affecting total system availability by containing most processor checkstops to the partition that was using the processor at the time that the full checkstop goes into effect.

Even with all these availability enhancements to prevent processor errors from affecting system-wide availability, errors might result on a system-wide outage.

4.2.3 Memory protection

A memory protection architecture that provides good error resilience for a relatively small L1 cache might be inadequate for protecting the much larger system main store. Therefore, a variety of protection methods are used in POWER processor-based systems to avoid uncorrectable errors in memory.

Memory protection plans must take into account many factors, including the following items:

- Size
- Desired performance
- Memory array manufacturing characteristics

POWER7 processor-based systems have a number of protection schemes designed to prevent, protect, or limit the effect of errors in main memory:

- Chipkill

  Chipkill is an enhancement that enables a system to sustain the failure of an entire DRAM chip. An ECC word uses 18 DRAM chips from two DIMM pairs, and a failure on any of the DRAM chips can be fully recovered by the ECC algorithm. The system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced.

- 72-byte ECC

  In POWER7 an ECC word consists of 72 bytes of data. Of these, 64 are used to hold application data. The remaining eight bytes are used to hold check bits and additional information about the ECC word.

  This innovative ECC algorithm from IBM research works on DIMM pairs on a rank basis. (a rank is a group of 10 DRAM chips on the Power 795). With this ECC code, the system can dynamically recover from an entire DRAM failure (Chipkill) but also correct an error even if another *symbol* (a byte, accessed by a 2-bit line pair) experiences a fault (this way is an improvement from the Double Error Detection/Single Error Correction ECC implementation that is on the POWER6 processor-based systems).

- DRAM sparing

  The Power 795 system has a spare DRAM chip per rank on each DIMM that can be used to replace a failed DIMM in a rank (Chipkill event). Effectively, this protection means that a DIMM pair can sustain two, and in some cases three DRAM chip failures and correct the errors without performance degradation.

- Hardware scrubbing

  Hardware scrubbing is a method that is used to deal with intermittent errors. IBM POWER processor-based systems periodically address all memory locations; any memory locations with a correctable error are rewritten with the correct data.

- CRC

  The bus that is transferring data between the processor and the memory uses CRC error detection with a failed operation-retry mechanism and the ability to dynamically retune bus parameters when a fault occurs. In addition, the memory bus has spare capacity to substitute a data bit-line, when it is determined to be faulty.

POWER7 memory subsystem

The POWER7 processor chip contains two memory controllers with four channels per memory controller. Each channel connects to a single DIMM, but because the channels work in pairs, a processor chip can address four DIMM pairs, two pairs per memory controller.

The bus transferring data between the processor and the memory uses CRC error detection with a failed operation-retry mechanism and the ability to dynamically retune bus parameters when a fault occurs. In addition, the memory bus has spare capacity to substitute a spare data bit-line, which is determined to be faulty.

Figure 4-3 shows a POWER7 processor chip, with its memory interface, consisting of two controllers and four DIMMs per controller. Advanced memory buffer chips are exclusive to IBM and help to increase performance, acting as read/write buffers. On the Power 770 and 780, the advanced memory buffer chips are integrated to the DIMM that they support. Power 750 and 755 uses only one memory controller, Advanced memory buffer chips are on the system planar and support two DIMMs each.

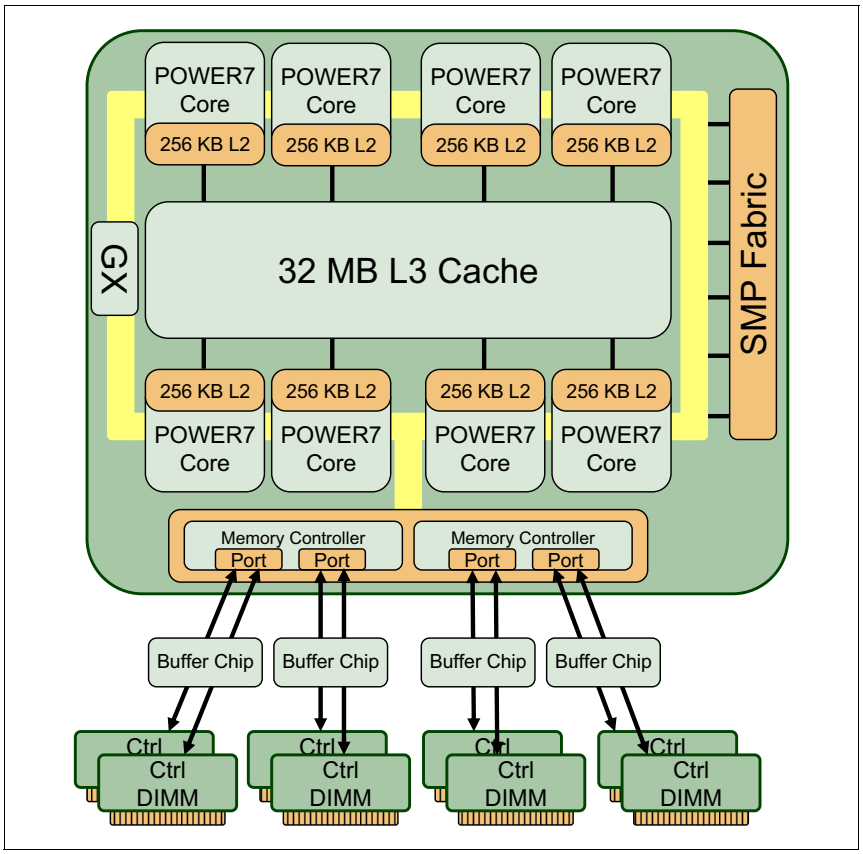

Figure 4-3 POWER7 memory subsystem

Memory page deallocation

Although coincident cell errors in separate memory chips are a statistic rarity, IBM POWER processor-based systems can contain these errors by using a memory page deallocation scheme for partitions that are running IBM AIX and the IBM i operating systems, and also for memory pages that are 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.

Pages that are used by the POWER Hypervisor are deallocated as soon as the page is released.

In other cases, the POWER Hypervisor notifies the owning partition that the page must be deallocated. Where possible, the operating system moves any data that is currently contained in that memory area to another memory area, and removes the page (or pages) that are associated with this error from its memory map, no longer addressing these pages. The operating system performs memory page deallocation without any user intervention and is transparent to users and applications.

The POWER Hypervisor maintains a list of pages that are 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 when memory pages are included that need to be deallocated.

Finally, if an uncorrectable error in memory is discovered, the logical memory block that is associated with the address with the uncorrectable error is marked for deallocation by the POWER Hypervisor. This deallocation takes effect on a partition reboot if the logical memory block is assigned to an active partition at the time of the fault.

In addition, the system will deallocate the entire memory group that is associated with the error on all subsequent system reboots until the memory is repaired. This precaution is intended to guard against future uncorrectable errors while waiting for parts replacement.

Memory persistent deallocation

Defective memory that is discovered at boot time is automatically switched off. 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.

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 must be scheduled as soon as is convenient.

Upon reboot, if not enough memory is available to meet minimum partition requirements, the POWER Hypervisor will reduce the capacity of one or more partitions.

Depending on the configuration of the system, the HMC Service Focal Point, OS Service Focal Point, or service processor receives a notification of the failed component, and will trigger a service call.

4.2.4 Active Memory Mirroring for Hypervisor

Active Memory Mirroring for Hypervisor is an option that mirrors the main memory used by the firmware. With this option, an uncorrectable error resulting from failure of main memory used by system firmware will not cause a system-wide outage. This option, enabled by default, efficiently guards against system-wide outages as a result of any such uncorrectable error that is associated with firmware.

With this option, uncorrectable errors in data that is owned by a partition or application is handled by the existing special uncorrectable error handling feature, described previously.

Active Memory Mirroring For Hypervisor is enabled or disabled by the ASMI. Figure 4-4 shows this option as Selective Memory Mirroring in the Advanced System Management interface.

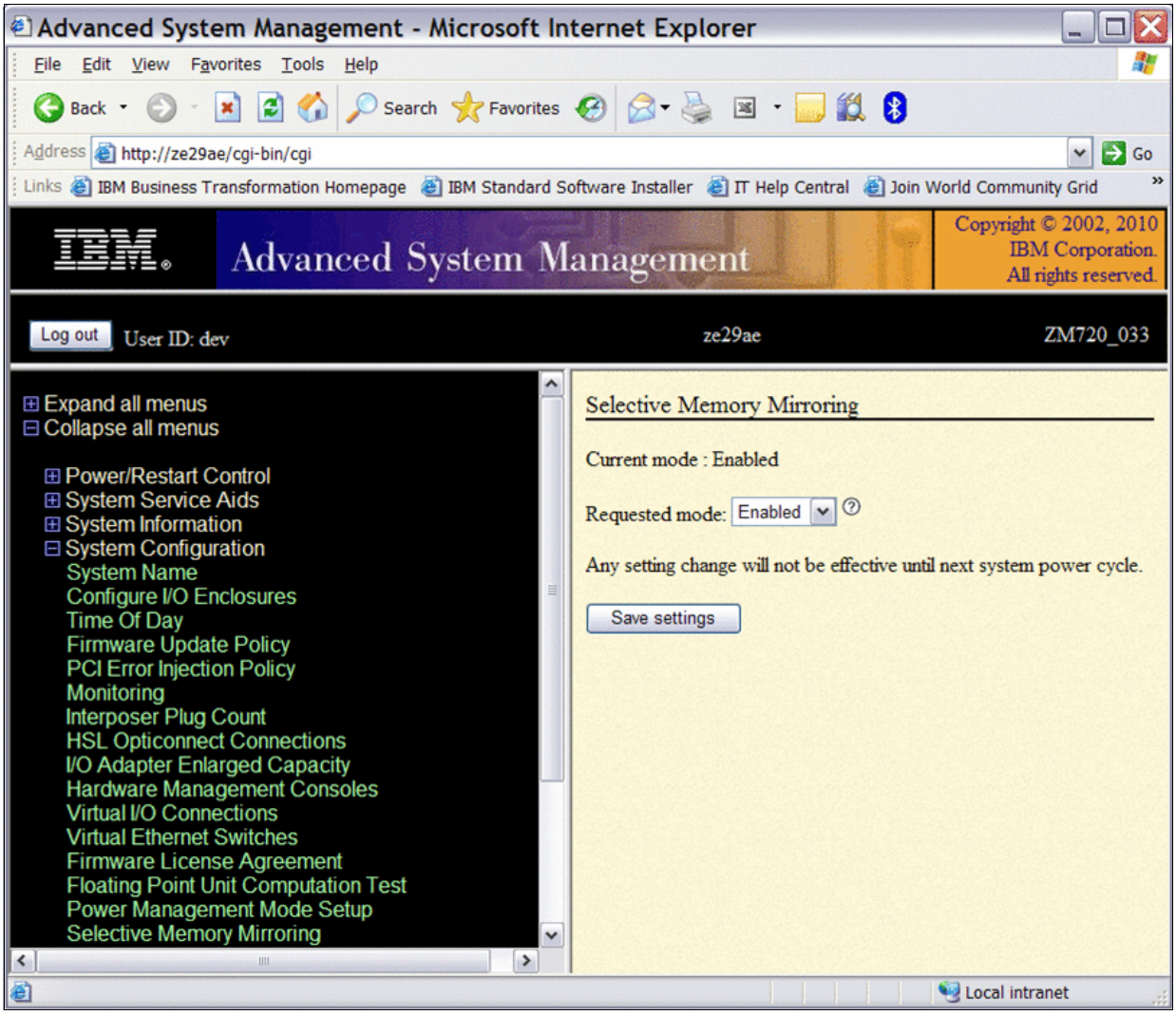

Figure 4-4 ASMI Selective Memory Mirroring (formerly Active Memory Mirroring) panel

4.2.5 Cache protection

POWER7 processor-based systems are designed with cache protection mechanisms, including cache-line delete in both L2 and L3 arrays, processor instruction retry and alternate processor recovery protection on L1-I and L1-D, and redundant *repair* bits in L1-I, L1-D, and L2 caches, and in L2 and L3 directories.

L1 instruction and data array protection

The POWER7 processor instruction and data caches are protected against intermittent errors by using processor instruction retry and against permanent errors by Alternate processor recovery, both mentioned previously. L1 cache is divided into sets. POWER7 processor can deallocate all but one set before doing a processor instruction retry.

In addition, faults in the segment-lookaside buffer (SLB) array are recoverable by the POWER Hypervisor. The SLB is used in the core to do address translation calculations.

L2 and L3 array protection

The L2 and L3 caches in the POWER7 processor are protected with double-bit detect single-bit correct error detection code (ECC). Single-bit errors are corrected before being forwarded to the processor, and subsequently written back to L2 and L3.

In addition, the caches maintain a cache-line delete capability. A threshold of correctable errors detected on a cache line can result in the data in the cache line being purged and the cache line removed from further operation without requiring a reboot. An ECC uncorrectable error that is detected in the cache can also trigger a purge and deleting of the cache line. This behavior results in no loss of operation because an unmodified copy of the data can be held on system memory to reload the cache line from main memory. Modified data is handled through special uncorrectable error handling.

L2- and L3-deleted cache lines are marked for persistent deconfiguration on subsequent system reboots until the processor card can be replaced.

4.2.6 Special uncorrectable error handling

Although rare, an uncorrectable data error can occur in memory or a cache. IBM POWER processor-based systems attempt to limit, to the least possible disruption, the impact of an uncorrectable error by 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. Consider the following examples:

- Data in the instruction L1 cache is never modified within the cache itself. Therefore, an uncorrectable error that is discovered in the cache is treated like an ordinary cache-miss, and correct data is loaded from the L2 cache.
- The L2 and L3 cache of the POWER7 processor-based systems can hold an unmodified copy of data in a portion of main memory. In this case, an uncorrectable error simply triggers a reload of a cache line from main memory.

In cases where the data cannot be recovered from another source, a technique called special uncorrectable error (SUE) handling is used to prevent an uncorrectable error in memory or cache from immediately causing the system to terminate. Instead, the system tags the data and determines whether it can ever be used again:

- If the error is irrelevant, it does not force a checkstop.
- If the data is used, termination can be limited to the program or kernel, or hypervisor that owns the data, or a freezing of the I/O adapters that are controlled by an I/O hub controller if data is to be transferred to an I/O device.

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 later) or Linux, and a process attempts to use the data, the operating system is informed of the error and might terminate, or only terminate a specific process associated with the corrupt data, depending on the operating system and firmware level and whether the data was associated with a kernel or non-kernel process.

Only when the corrupt data is being used by the POWER Hypervisor must the entire system be rebooted, thereby preserving overall system integrity.

Depending on system configuration and source of the data, errors that are encountered during I/O operations might not result in a machine check. Instead, the incorrect data is handled by the PCI 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, resulting in the loss of all I/O operations that use the frozen hardware until a power-on reset of the PHB. The impact to partitions depends on how the I/O is configured for redundancy. In a server that is configured for failover availability, redundant adapters spanning multiple PHB chips can enable the system to recover transparently, without partition loss.

4.2.7 PCI-enhanced error handling

IBM estimates that PCI adapters can account for a significant portion of the hardware-based errors on a large server. Whereas servers that rely on boot-time diagnostics can identify failing components to be replaced by hot-swap and reconfiguration, runtime errors pose a more significant problem.

PCI adapters are generally complex designs that involve extensive on-board instruction processing, often on embedded microcontrollers. They tend to use industry standard grade components with an emphasis on product cost that is relative to high reliability. In certain cases, they might be more likely to encounter internal microcode errors, or 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 certain cases, an error in the adapter can 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.

PCI-enhanced error-handling-enabled adapters respond to a special data packet that is generated from the affected PCI slot hardware by calling system firmware. The firmware examines the affected bus, allows the device driver to reset it, and continues without a system reboot. For Linux, enhanced error handling (EEH) support extends to the majority of frequently used devices, although various third-party PCI devices might not provide native EEH support.

To detect and correct PCIe bus errors, POWER7 processor-based systems use CRC detection and instruction retry correction; for PCI-X, it uses ECC.

Figure 4-5 shows the location and mechanisms used throughout the I/O subsystem for PCI enhanced error handling.

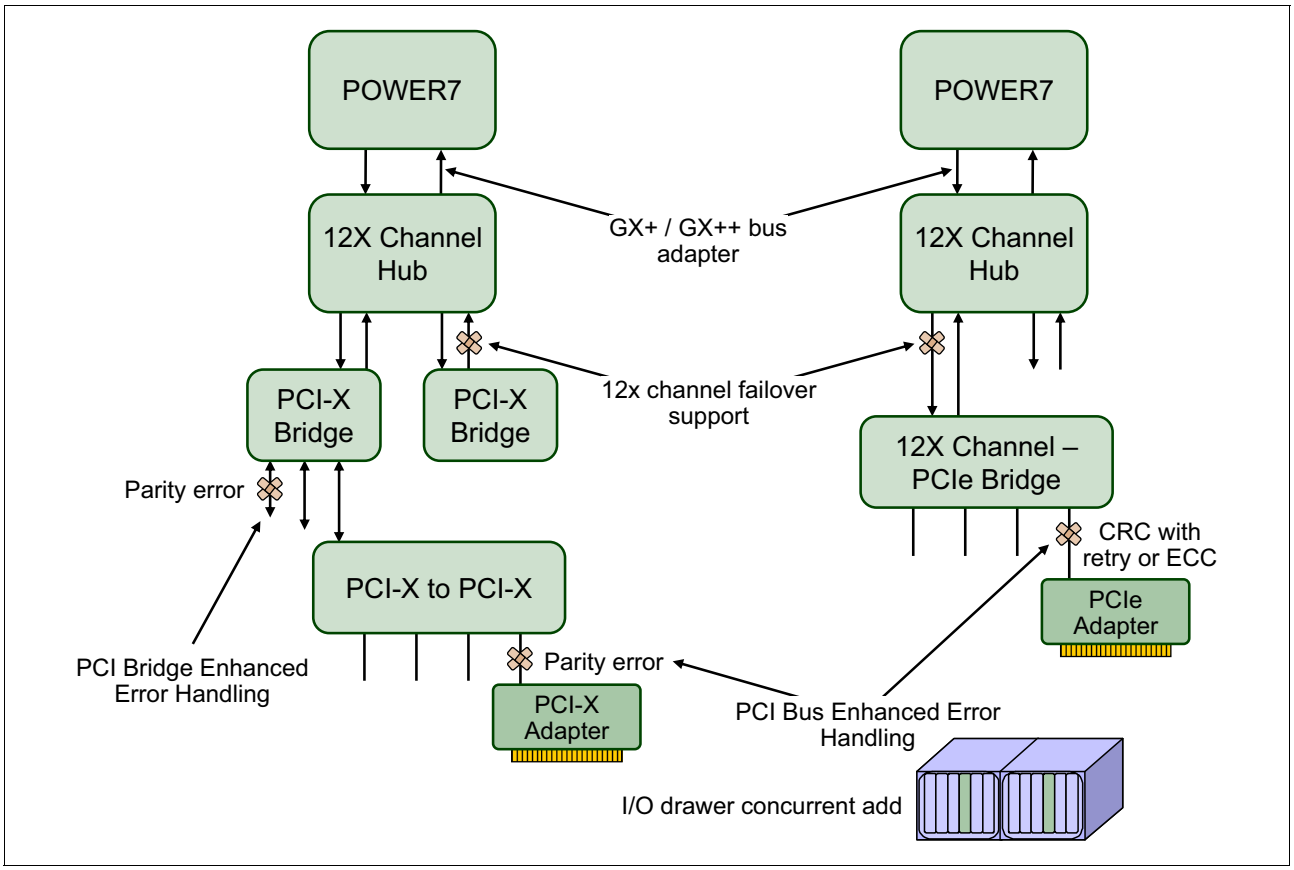

Figure 4-5 PCI-enhanced error handling

4.3 Serviceability

IBM Power Systems design considers both IBM and client needs. The IBM Serviceability Team enhanced the base service capabilities and continues to implement a strategy that incorporates best-of-its-kind service characteristics from diverse IBM systems offerings.

Serviceability includes system installation, system upgrades and downgrades (MES), and system maintenance and repair.

The goal of the IBM Serviceability Team is to design and provide the most efficient system service environment, which includes the following benefits:

- Easy access to service components; design for customer setup (CSU), customer installed features (CIF), and customer-replaceable units (CRU)
- On demand service education
- Error detection and fault isolation (ED/FI)
- First-failure data capture (FFDC)
- 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, IBM Power Systems servers enable faster and more accurate repair, and reduce the possibility of human error.

Client control of the service environment extends to firmware maintenance on all of the POWER processor-based systems. This strategy contributes to higher systems availability with reduced maintenance costs.

The term *servicer*, when used in the context of this document, denotes the person whose task is to do service-related actions on a system. For an item designated as a CRU, the servicer might be the client. In other cases, for field-replaceable unit (FRU) items, the servicer might be an IBM representative or an authorized warranty service provider.

Service can be divided into three main categories:

- Service components: The basic service-related building blocks
- Service functions: Service procedures or processes containing one or more service components
- Service operating environment: The specific system operating environment that specifies how service functions are provided by the various service components

The basic component of service is a *serviceable event*.

Serviceable events are platform, regional, and local error occurrences that require a service action (repair). The event can include a *call home* function to report the problem so that the repair can be assessed by a trained service representative. In all cases, the client is notified of the event. Event notification includes a clear indication of when servicer intervention is required to rectify the problem. The intervention might be a service action that the client can perform or it might require a service provider.

Serviceable events are classified as follows:

- Recoverable: This event is a correctable resource or function failure. The server remains available, but there might be some amount of decrease in operational performance available for the client's workload (applications).
- Unrecoverable: This event is an uncorrectable resource or function failure. In this instance, there is potential degradation in availability and performance, or loss of function to the client's workload.
- Predictable (using thresholds in support of Predictive Failure Analysis): This event is a determination that continued recovery of a resource or function might lead to degradation of performance or failure of the client's workload. Although the server remains fully available, if the condition is not corrected, an unrecoverable error might occur.
- Informational: This event is notification about a resource or function in the following circumstances:
  - It is *out-of* or *returned-to* specification and might require user intervention.
  - It requires user intervention to complete a system tasks.

Platform errors are faults that affect all partitions in various ways. They are detected in the CEC by the service processor, the system power control network (SPCN), or the POWER Hypervisor. When a failure occurs in these components, the POWER Hypervisor notifies each partition's operating system to execute any required precautionary actions or recovery methods. The OS is required to report these kinds of errors as serviceable events to the Service Focal Point application because, by the definition, they affect the partition.

Platform errors are faults that are related to the following items:

- The central electronics complex (CEC), which is that part of the server that is composed of the central processing units, memory, storage controls, and the I/O hubs
- The power and cooling subsystems
- The firmware that is used to initialize the system and diagnose errors

Regional errors are faults that affect some, but not all partitions. They are detected by the POWER Hypervisor or the service processor. Examples include RIO bus, RIO bus adapter, PHB, multi-adapter bridges, I/O hub, and errors on I/O units (except adapters, devices, and their connecting hardware).

Local errors are faults that are detected in a partition (by the partition firmware or the operating system) for resources that are owned only by that partition. The POWER Hypervisor and service processor are not aware of these errors. Local errors might include “secondary effects” that result from platform errors preventing partitions from accessing partition-owned resources. Examples include PCI adapters or devices assigned to a single partition. If a failure occurs to one of these resources, only a single operating system partition need be informed.

This section provides an overview of the progressive steps of error detection, analysis, reporting, notifying, and repairing that are found in all POWER processor-based systems.

4.3.1 Detecting

The first and most crucial component of a solid serviceability strategy is the ability to accurately and effectively detect errors when they occur. Although 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. POWER processor-based systems employ IBM System z® server-inspired error detection mechanisms that extend from processor cores and memory to power supplies and hard drives.

Service processor

The *service processor* is a microprocessor that is powered separately from the main instruction processing complex. The service processor provides the capabilities for the following features:

- POWER Hypervisor (system firmware) and Hardware Management Console (HMC) connection surveillance
- Various remote power control options
- Reset and boot features
- 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 environmentally related problems 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 under the following circumstances:

- The operating temperature exceeds the critical level (for example, failure of air conditioning or air circulation around the system).
- The system fan speed is out of operational specification, for example, because of multiple fan failures.
- The server input voltages are out of operational specification.

The service processor can immediately shut down a system when the following situations occur:

- Temperature exceeds the critical level or remains above the warning level for too long.
- Internal component temperatures reach critical levels.
- Non-redundant fan failures occur.

- Placing calls

  On systems without an HMC, the service processor can place calls to report surveillance failures with the POWER Hypervisor, critical environmental faults, and critical processing faults even when the main processing unit is inoperable.

- 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, cache, memory, and associated hardware that is required for proper booting of the operating system, when the system is powered on at the initial installation or after a hardware configuration change (for example, 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 to 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

- Concurrent access to the service processors menus of the Advanced System Management Interface (ASMI):

  This access allows nondisruptive abilities to change system default parameters, interrogate service processor progress and error logs, set and reset server indicators, (Guiding Light for midrange and high-end servers, and Light Path for low-end servers), indeed, accessing all service processor functions without having to power-down the system to the standby state. This way allows the administrator or service representative to dynamically access the menus from any web browser-enabled console that is attached to the Ethernet service network, concurrently with normal system operation.

- Managing the interfaces for connecting uninterruptible power source systems to the POWER processor-based systems, performing timed power-on (TPO) sequences, and interfacing with the power and cooling subsystem

Error checkers

IBM POWER processor-based systems contain specialized hardware detection circuitry that is used to 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:

- Continuous monitoring of system operations to detect potential calculation errors
- Attempts to isolate physical faults based on runtime detection of each unique failure
- Ability to initiate a wide variety of recovery mechanisms designed to correct the problem

  The POWER processor-based systems include extensive hardware and firmware recovery logic.

Fault isolation registers (FIR)

Error checker signals are captured and stored in hardware FIRs. The associated 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)

The FFDC error isolation technique 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 will be captured without the need to re-create 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 is detected automatically without intervention by a service representative. Pertinent error data that is related to the fault is captured and saved for analysis. In hardware, FFDC data is collected from the fault isolation registers (FIRs) and from the associated logic. In firmware, this data consists of return codes, function calls, and so forth.

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 a field-replaceable unit (FRU).

This proactive diagnostic strategy is a significant improvement over the classic, less accurate *reboot and diagnose* service approaches.

Figure 4-6 shows a diagram of a fault isolation register implementation.

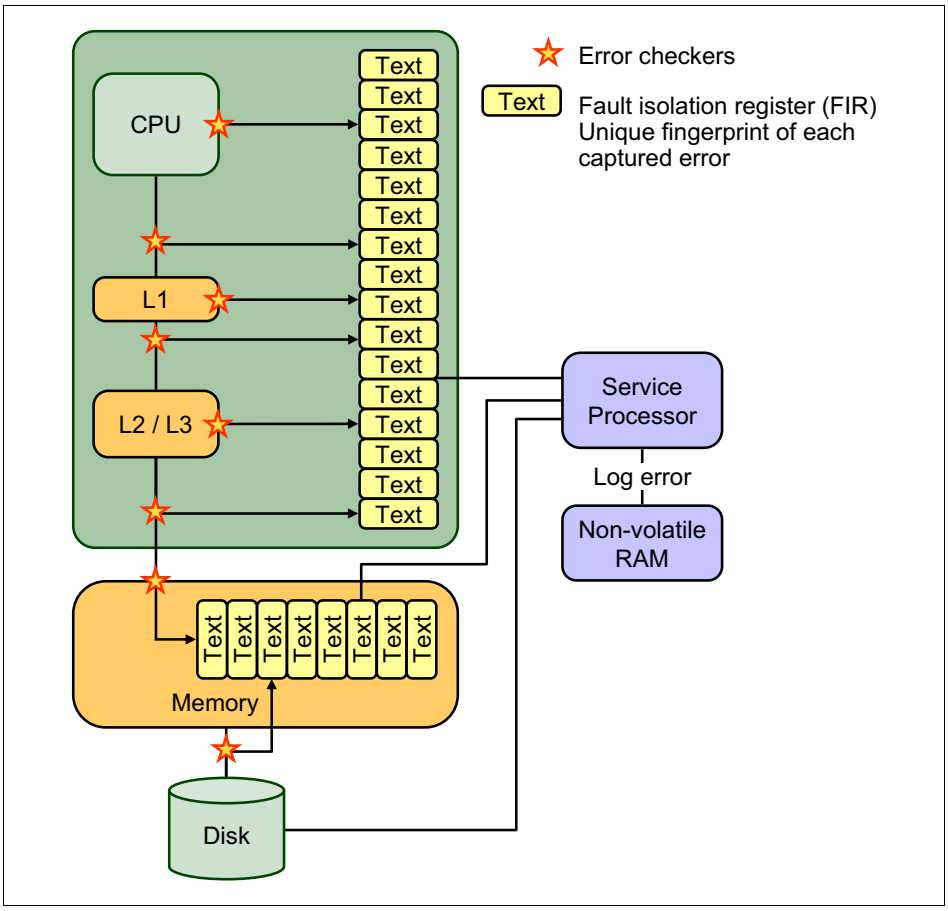

Figure 4-6 Schematic of FIR implementation

Analysis: FFDC plays a critical role in delivering servers that can self-diagnose and self-heal. Using thousands of checkers (diagnostic probes) deployed at critical junctures throughout the server, the system effectively “traps” hardware errors at system run time.

The separately powered service processor is then used to analyze the checkers and perform problem determination. Using this method, IBM no longer has to rely on an intermittent “reboot and retry” error detection strategy, but knows with a degree of certainty which part is having problems. In this automated approach, 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 FRU (part) call, and manual interpretation of the data is not normally required.

This architecture is also the basis for the IBM predictive failure analysis, because the service processor can now count, and log, intermittent component errors, and can deallocate or take other corrective actions when an error threshold is reached.

Fault isolation

The service processor interprets error data that is captured by the FFDC checkers (saved in the FIRs or other firmware-related data capture methods) to determine the root cause of the error event.

Root cause analysis might indicate that the event is recoverable, meaning that a service action point or need for repair was not reached. Alternatively, it might indicate that a service action point was reached, where the event exceeded a predetermined threshold or was unrecoverable. Based on the isolation analysis, recoverable error threshold counts might 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 that a service action point has been reached, a request for service is initiated through an error logging component.

4.3.2 Diagnosing

Using the extensive network of advanced and complementary error detection logic that is built directly into hardware, firmware, and operating systems, the IBM Power Systems servers can perform considerable self-diagnosis.

Boot time

When an IBM Power Systems server powers up, the service processor initializes system hardware. Boot-time diagnostic testing uses a multitier approach for system validation, starting with managed low-level diagnostics that are supplemented with system firmware initialization and configuration of I/O hardware, followed by OS-initiated software test routines. The following boot-time diagnostic routines are included:

- Built-in self-tests (BISTs) for both logic components and arrays ensure the internal integrity of components. Because the service processor assists in performing these tests, the system is enabled to perform fault determination and isolation, whether or not the system processors are operational. Boot-time BISTs can also find faults undetectable by processor-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.

Run time

All Power Systems servers can monitor critical system components during run time, and they can take corrective actions when recoverable faults occur. IBM 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 runtime diagnostic capabilities originate with the service processor. Extensive diagnostic and fault analysis routines were developed and improved over many generations of POWER processor-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 engineering expertise to count recoverable errors (called *thresholds*) and predict when corrective actions must be automatically initiated by the system.

Such actions can include the following possibilities:

- Requests for a part to be replaced
- Dynamic 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.

4.3.3 Reporting

In the unlikely event that a system hardware or environmentally induced failure is diagnosed, IBM Power Systems servers report the error through a number of mechanisms. The analysis result is stored in system NVRAM. Error log analysis (ELA) can be used to display the failure cause and the physical location of the failing hardware.

With the integrated service processor, the system has the ability to automatically send an alert through a phone line to a pager, or call for service in the event of a critical system failure. A hardware fault also illuminates the amber system fault LED, located on the system unit to alert the user of an internal hardware problem.

On POWER7 processor-based servers, hardware and software failures are recorded in the system log. When an HMC is attached, an ELA routine analyzes the error, forwards the event to the Service Focal Point (SFP) application running on the HMC, and notifies the system administrator that it isolated a likely cause of the system problem. The service processor event log also records unrecoverable checkstop conditions, forwards them to the SFP application, and notifies the system administrator.

After the information is logged in the SFP application, if the system is properly configured, a call-home service request is initiated and the pertinent failure data with service parts information and part locations is sent to an IBM service organization. Client contact information and specific system-related data such as the machine type, model, and serial number, along with error log data related to the failure are sent to IBM Service.

Error logging and analysis

When the root cause of an error has been identified by a fault isolation component, an error log entry is created with basic data such as the following items:

- An error code that uniquely describes the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data such as engineering and manufacturing levels
- Return codes
- Resource identifiers
- FFDC data

Data containing information about the effect that the repair will have on the system is also included. Error log routines in the operating system can then use this information and decide to call home to contact service and support, send a notification message, or continue without an alert.

Remote support

The Resource Monitoring and Control (RMC) application is delivered as part of the base operating system, including the operating system running on the Hardware Management Console (HMC). The RMC application 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 also, but these are not used for the service infrastructure.

Service Focal Point

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 must only be reported once, regardless of how many logical partitions experience the potential effect of the error. The Manage Serviceable Events task on the Hardware Management Console (HMC) is responsible for aggregating duplicate error reports, and ensures 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 Resource Monitoring and Control Subsystem (RMC) to relay error information to the Hardware Management Console. For global events (platform unrecoverable errors, for example) the service processor will also forward error notification of these events to the Hardware Management Console, providing a redundant error-reporting path in case of errors in the RMC network.

The first occurrence of each failure type is recorded in the Manage Serviceable Events task on the Hardware Management Console. This task then filters and maintains a history of duplicate reports from other logical partitions on the service processor. It then looks at all active service event requests, analyzes the failure to ascertain the root cause and, if enabled, initiates a call home for service. This methodology 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 that is collected either automatically at the time of a failure or manually at a later time. The data collected is dependent on the invocation method but includes information such as 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 certain circumstances, an error might require a dump to be automatically or manually created. In this event, it is off-loaded to the HMC upon the 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 becomes necessary to view the dump remotely, the HMC dump record notifies the IBM support center regarding on which HMC the dump is located.

4.3.4 Notifying

After a Power Systems server detects, diagnoses, and reports an error to an appropriate aggregation point, it then takes steps to notify the client, and if necessary, the IBM support organization. Depending upon the assessed severity of the error and support agreement, this might range from a simple notification to having field service personnel automatically dispatched to the client site with the correct replacement part.

Client 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 *Client Notify*. Clients are notified because these events might 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 are as follows:

- Network events such as the loss of contact over a local area network (LAN)
- Environmental events such as ambient temperature warnings
- Events that need further examination by the client (although these events do not necessarily require a part replacement or repair action)

Client Notify events are serviceable events, by definition, because they indicate that something has happened that requires client awareness in the event the client wants to take further action. These events can always be reported back to IBM at the client's discretion.

Call home

A correctly configured POWER processor-based system can initiate an automatic or manual call from a client location to the IBM service and support organization with error data, server status, or other service-related information. The *call-home* feature invokes the service organization in order for the appropriate service action to begin, automatically opening a problem report, and in certain 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

Power Systems store VPD internally, keeping a record of how much memory is installed, how many processors are 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 of the information that is related to the error, along with 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.

4.3.5 Locating and servicing

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Packaging for service

The following service enhancements are included in the physical packaging of the systems to facilitate service:

- Color coding (touch points):
  - Terra-cotta-colored touch points indicate that a component (FRU or CRU) can be concurrently maintained.
  - Blue-colored touch points delineate components that are not concurrently maintained (those that require the system to be turned off for removal or repair).
- Tool-less design: Selected IBM systems support tool-less or simple tool designs. These designs require no tools or simple tools such as flathead screw drivers to service the hardware components.
- Positive retention: Positive retention mechanisms help to assure proper connections between hardware components, such as from cables to connectors, and between two cards that attach to each other. Without positive retention, hardware components run the risk of becoming loose during shipping or installation, preventing a good electrical connection. Positive retention mechanisms such as latches, levers, thumb-screws, pop Nylatches (U-clips), and cables are included to help prevent loose connections and aid in installing (seating) parts correctly. These positive retention items do not require tools.

Light Path

The Light Path LED feature is for low-end systems, including Power Systems up to models 750 and 755, that can be repaired by clients. In the Light Path LED implementation, when a fault condition is detected on the POWER7 processor-based system, an amber FRU fault LED is illuminated, which is then rolled up to the system fault LED. The Light Path system pinpoints the exact part by turning on the amber FRU fault LED that is associated with the part to be replaced.

The system can clearly identify components for replacement by using specific component-level LEDs, and can also guide the servicer directly to the component by signaling (staying on solid) the system fault LED, enclosure fault LED, and the component FRU fault LED.

After the repair, the LEDs shut off automatically if the problem is fixed.

Guiding Light

The enclosure and system identify LEDs that turn on solid and that can be used to follow the path from the system to the enclosure and down to the specific FRU.

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 might be necessary in certain complex high-end configurations.

In these situations, Guiding Light waits for the servicer's indication of what failure to attend first and then illuminates the LEDs to the failing component.

Data centers 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.

POWER7 high-end Power 795 server uses a front and back light strip for service. The front light strip is new with the Power 795, but the rear light strip is the same as the one on POWER6 595. The light strips use the Firefly LED controller. Figure 4-7 shows a front light strip in IBM Power 795 server.

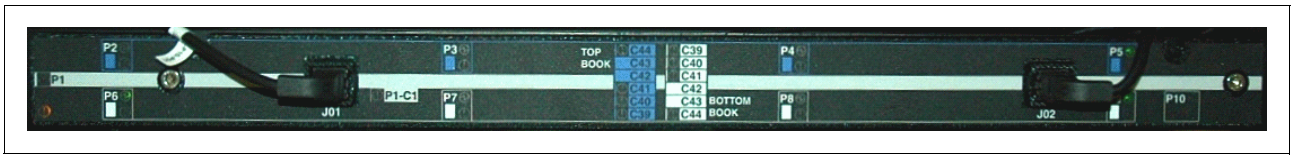

Figure 4-7 Front light strip in Power 795

Service labels

Service providers use service labels to assist them in performing maintenance actions. Service labels are found in various formats and positions, and are intended to transmit readily available information to the servicer during the repair process.

Various service labels and the purpose of each are described in the following list:

- Location diagrams are strategically located on the system hardware, relating information regarding the placement of hardware components. Location diagrams can include location codes, drawings of physical locations, concurrent maintenance status, or other data that is pertinent to a repair. Location diagrams are especially useful when multiple components are installed such as DIMMs, CPUs, processor books, fans, adapter cards, LEDs, and power supplies.
- Remove or replace procedure labels contain procedures often found on a cover of the system or in other spots that are accessible to the servicer. These labels provide systematic procedures, including diagrams, detailing how to remove and replace certain serviceable hardware components.
- Numbered arrows indicate the order of operation and serviceability direction of components. Various serviceable parts such as latches, levers, and touch points must be pulled or pushed in a certain direction and certain order so that the mechanical mechanisms can engage or disengage. Arrows generally improve the ease of serviceability.

The operator panel

The operator panel on a POWER processor-based system is a four-row by 16-element LCD display that is 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. It includes various buttons, enabling a service support representative or client to change various boot-time options and other limited service functions.

The CEC Hot Add & Repair Maintenance (CHARM)

The IBM POWER7 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 wearing down or burning out; other devices such as I/O adapters can begin to wear from

repeated plugging and unplugging. For these reasons, these devices have been specifically designed to be concurrently maintainable, when properly configured.

In other cases, a client might be in the process of moving or redesigning a data center, or planning a major upgrade. At times like these, flexibility is crucial. The most recent members of the IBM Power Systems family, based on the POWER7 processor, continue to support concurrent maintenance of power, cooling, media devices, I/O drawers, GX adapter, and the operator panel. In addition, they support concurrent firmware fixpack updates when possible. The determination of whether a firmware fixpack release can be updated concurrently is identified in the *readme* file that is released with the firmware.

Concurrent add and repair capabilities for the Power Systems have been introduced incrementally since 1997, starting with power supply, fan, I/O device, PCI adapter, and I/O enclosure or drawer. In 2008, IBM introduced significant enhancements to the enterprise Power Systems 595 and 570 that highlighted the ability to add and upgrade system capacity and repair the CEC, or the heart of a large computer system, which includes the processors, memory and I/O hubs (GX adapters), without powering down the system.

The IBM Power 795 servers continue to improve on the CEC hot add and repair functions that were introduced with the Power Systems 595 and 570. Best practice experience from client engagements, in 2008 - 2010, calls for a new level of minimum enablement criteria that includes proper planning during system order, configuration, installation, I/O optimization for RAS.

With proper advanced planning and minimum criteria being met, the IBM Power 795 CEC Hot Add & Repair Maintenance (CHARM) function allows the expansion of the system processors, memory, and I/O hub capacity, and their repair also, with limited disruption to the system operation.

The CEC hardware includes the processors, memory, I/O hubs (GX adapters), system clock, service processor and associated CEC support hardware. The CHARM functions consist of special hardware design, service processor, hypervisor, and HMC firmware. The HMC provides the user interfaces for the system administrator and systems service representative to perform the tasks for the CHARM operation.

The CHARM functions use the existing Resource Monitoring and Control (RMC) connection between the HMC and the LPARs to determine the impact on the resources (processor, memory and I/O) of the partitions in preparation for a hot node/GX repair or hot node upgrade operation. The HMC communicates with the affected partitions to perform the dynamic LPAR remove and add operations during the CHARM operation. The HMC communicates with the service processor and hypervisor to perform CHARM tasks such as a system “readiness check,” administrative failover (for redundant service processor and clock functions), workload evacuation, hardware deactivation, power off, power on, CEC hardware diagnostics, hardware activation, resource configuration, resource integration, and error handling and reporting, during the CHARM operation.

During a hot-node repair, I/O hub repair, or node upgrade operation, the node or I/O hub is powered off for safe removal and insertion. Consequently, all resources that depend on the node (CEC processor book for the Power 795) or I/O hub including processors, memory, internal I/O, and externally attached I/O are not functional during the CHARM operation. The CHARM functions identify the resources that will be affected, and guide the system administrator through the process of gracefully deactivating or freeing up resources as necessary to complete the operation. The CHARM functions ensure that all resources that will be affected by the operation are explicitly deactivated by the user or workloads that use those resources, and can be evacuated to available resources elsewhere in the system, before allowing the operation to proceed.

Hot-node add, hot-node repair, and memory upgrade

With the proper configuration and required protective measures, the Power 795 servers are designed for node add, node repair, concurrent GX adapter add, concurrent GX adapter repair, concurrent system controller repair or memory upgrade without powering down the system.

The Power 795 servers support the addition of another CEC enclosure (node) to a system (hot-node add), adding GX adapter or adding more memory (memory upgrade) to an existing node. The additional Power 795 enclosure or memory can be ordered as a system upgrade (MES order) and added to the original system. The additional resources of the newly added CEC enclosure (node) or memory can then be assigned to existing OS partitions or new partitions as required. Hot-node add and memory upgrade enable the upgrading of a server by integrating a second, third, or fourth CEC enclosure or additional memory into the server, with reduced impact to the system operation.

In an unlikely event that CEC hardware (for example, processor or memory) experienced a failure, the hardware can be repaired by freeing the processors and memory in the node and its attached I/O resources (node evacuation).

The Prepare for Hot Repair or Upgrade (PHRU) utility is a tool on the HMC, and is used by the system administrator to prepare the system for a hot-node repair, hot-node upgrade, or hot-plug GX repair operation. Among other tasks, the utility identifies the following information:

- I/O resources that will be affected by the operation
- I/O resources that might be in use by operating systems and must be deactivated or released by the operating system
- Additional processor and memory capacity that must be made available
- System configuration issues that preclude CHARM operations

Table 4-1 details the PHRU utility usage.

Table 4-1 PHRU utility usage

CHARM operation	Minimum nodes to use operation	PHRU usage	System administrator	Service representative
Hot-node add	1	No	Planning only	Yes
Hot-node repair	2	Yes	Yes	Yes
Hot-node upgrade (memory)	2	Yes	Yes	Yes
Concurrent GX adapter add	1	No	Planning only	Yes
Hot GX adapter repair	1	Yes	Yes	Yes
Concurrent system controller repair	1	No	Planning only	Yes

Minimum Enablement Criteria

CHARM operations are complex; they involve numerous steps that are performed by service personnel or the system administrator while the system is powered on. Because the likelihood of failure increases with the complexity of the operation, the following minimum enablement criteria offers the best protection against any unforeseeable situations:

- Be sure that all scheduled hot-adds, hot-repairs, or hot-upgrades are done during "non-peak" operational hours.
- Be sure prerequisites are met:
  - Critical I/O resources must be configured with redundant paths for hot-node repair, hot-node upgrade, and hot-plug GX repair.
  - IBM Electronic Service Agent™ must be enabled to ensure timely resolution of all hardware failures. This step minimizes the opportunity for a situation in which multiple hardware failures occur during a CHARM operation.
  - Critical business applications must be moved to another server using Live Partition Mobility (LPM), if available, OR critical applications are quiesced for hot node add, hot node repair, hot node upgrade and hot GX adapter repair.
- The system administrator must *not* dynamically change the size of the 16 MB large page pool in AIX partitions with the **vmo** command while a CHARM operation is in progress.

For more details, see *IBM Power 770/780 and 795 Servers CEC Hot Add & Repair Maintenance Technical Overview*:

http://ibm.co/XjqeZ4

Blind-swap PCI adapters

Blind-swap PCI adapters represent significant service and ease-of-use enhancements in I/O subsystem design, 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. With *blind-swap*, PCI adapters can be concurrently replaced without having to put the I/O drawer into a service position.

Firmware updates

Firmware updates for Power 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.

When an HMC is connected to the system, the new firmware image is loaded with any of the following methods:

- IBM distributed media, such as a CD-ROM
- A Problem Fix distribution from the IBM Service and Support repository
- Use of FTP from another server
- Download from the IBM Fix Central website:

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

IBM supports multiple firmware releases in the field. Therefore, under expected circumstances, a server can operate on an existing firmware release, using concurrent firmware fixes to stay up-to-date with the current patch level. Because changes to various server functions (for example, changing initialization values for chip controls) cannot occur during system operation, a patch in this area requires a system reboot for activation. Under normal operating conditions, IBM provides patches for an individual firmware release level for

up to two years after first making the release code generally available. After this period, clients must plan to update to stay on a supported firmware release.

Activation of new firmware functions, as opposed to patches, 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 concurrent patches, which include functions that are not activated until a subsequent server reboot. A server with these patches can operate normally. The additional concurrent fixes can be installed and activated when the system reboots after the next scheduled outage.

Additional capability is being added to the firmware to be able to view the status of a system power control network background firmware update. This subsystem can update as necessary, as migrated nodes or I/O drawers are added to the configuration. The new firmware provides an interface to be able to view the progress of the update, and 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 that is used to guide a service provider, step-by-step, through the process of repairing a system and verifying that the problem is repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being repaired. The following repair scenarios are covered by repair and verify:

- Replacing a defective field-replaceable unit (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

Repair and verify procedures can be used by both service representative providers who are familiar with the task and those who are not. Education on demand content is placed in the procedure at the appropriate locations. Throughout the R&V 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.

Clients can subscribe through the subscription services to obtain the notifications about the latest updates available for service-related documentation. The latest version of the documentation is accessible through the Internet; a CD-ROM version is also available.

4.4 Manageability

Several functions and tools help manageability, and help you to efficiently and effectively manage your system.

4.4.1 Service user interfaces

The service interface allows support personnel or the client to communicate with the service support applications in a server using a console, interface, or terminal. Delivering a clear, concise view of available service applications, the service interface allows the support team to manage system resources and service information in an efficient and effective way.

Applications that are available through the service interface are carefully configured and placed to give service providers access to important service functions.

Various service interfaces are used, depending on the state of the system and its operating environment. The primary service interfaces are as follows:

- Light Path and Guiding Light

  For more information, see “Light Path” on page 186 and “Guiding Light” on page 186.
- Service processor, Advanced System Management Interface (ASMI)
- Operator panel
- Operating system service menu
- Service Focal Point on the Hardware Management Console
- Service Focal Point Lite on Integrated Virtualization Manager

Service processor

The service processor is a controller that is 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 that is connected to the POWER 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

Functions

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 Advanced System Management Interface (ASMI) Secure Sockets Layer (SSL) network connections. The service processor provides the ability to view and manage the machine-wide settings by using the ASMI, and enables complete system and partition management from the HMC.

Analyze system that does not boot: The service processor enables a system that does not boot to be analyzed. Error log analysis can be done from either the ASMI or HMC.

The service processor uses two Ethernet 10/100 Mbps ports. Note the following information:

- Both Ethernet ports are visible only 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.
- Both Ethernet ports support only auto-negotiation. Customer selectable media speed and duplex settings are not available.
- Both Ethernet ports have a default IP address, as follows:
  - 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
- When a redundant service processor is present, the default IP addresses are as follows:
  - Service processor Eth0 or HMC1 port is configured as 169.254.2.146
  - Service processor Eth1 or HMC2 port is configured as 169.254.3.146

The following functions are available through service processor:

- Call home
- Advanced System Management Interface (ASMI)
- Error Information (error code, PN, Location Codes) menu
- View of guarded components
- Limited repair procedures
- Generate dump
- LED Management menu
- Remote view of ASMI menus
- Firmware update through USB key

Advanced System Management Interface (ASMI)

ASMI is the interface to the service processor that enables you to manage the operation of the server, such as auto-power restart, and to view information about the server, such as the error log and vital product data. Various repair procedures require connection to the ASMI.

The ASMI is accessible through the HMC. It is also accessible by using 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. ASMI can also be accessed from an ASCII terminal. Use the ASMI to change the service processor IP addresses or to apply certain security policies and prevent access from undesired IP addresses or ranges.

You might be able to use the service processor's default settings. In that case, accessing the ASMI is not necessary. To access ASMI, use one of the following methods:

- Access the ASMI by 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 Advanced System Management interface from an HMC, use the following steps:

  a. Open Systems Management from the navigation pane.

  b. From the work pane, select one or more managed systems to work with.

  c. From the System Management tasks list, select **Operations Advanced System Management** (ASM).

- Access the ASMI by using 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 notebook that is connected to the service processor. The web interface is available during all phases of system operation, including the initial program load (IPL) and run time.

  However, several 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 ASMI provides a Secure Sockets Layer (SSL) web connection to the service processor. To establish an SSL connection, open your browser and use the `https://` format.

  Note: To make the connection through Internet Explorer, click **Tools Internet Options**. Clear the **Use TLS 1.0** check box, and click **OK**.

- Access the ASMI by using an ASCII terminal.

  The ASMI on an ASCII terminal supports a subset of the functions that are 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 several phases of system operation, such as the IPL and run time.

Operating system service menu

The system diagnostics consist of IBM i service tools, stand-alone diagnostics that are loaded from the DVD drive, and online diagnostics (available in AIX).

Online diagnostics, when installed, are a part of the AIX or IBM i 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. IBM i has a service tools problem log, IBM i history log (QHST), and IBM i problem log.

The modes are as follows:

- Service mode

  This mode requires a service mode boot of the system, enables the checking of system devices and features. Service mode provides the most complete self-check of the system resources. All system resources, except the SCSI adapter and the disk drives that are used for paging, can be tested.

- Concurrent mode

  This mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, certain devices might require additional actions by the user or diagnostic application before testing can be done.

- Maintenance mode

  This 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.

Alternate method: When you order a Power Systems, a DVD-ROM or DVD-RAM might be optional. An alternate method for maintaining and servicing the system must be available if you do not order the DVD-ROM or DVD-RAM.

The IBM i operating system and associated machine code provide Dedicated Service Tools (DST) as part of the IBM i licensed machine code (Licensed Internal Code) and System Service Tools (SST) as part of the IBM i operating system. DST can be run in dedicated mode (no operating system loaded). DST tools and diagnostics are a superset of those available under SST.

The IBM i **`End Subsystem`** (ENDSBS *ALL) command can shut down all IBM and customer applications subsystems except the controlling subsystem QTCL. The **`Power Down System`** (PWRDWNSYS) command can be set to power down the IBM i partition and restart the partition in DST mode.

You can start SST during normal operations, which leaves all applications running, by using the IBM i **`Start Service Tools`** (STRSST) command (when signed onto IBM i with the appropriately secured user ID).

With DST and SST you can look at various logs, run various diagnostics, or take several kinds of system dumps or other options.

Depending on the operating system, you typically see the following service-level functions when you use the operating system service menus:

- Product activity log
- Trace Licensed Internal Code
- Work with communications trace
- Display/Alter/Dump
- Licensed Internal Code log
- Main storage dump manager
- Hardware service manager
- Call Home/Customer Notification
- Error information menu
- LED management menu
- Concurrent/Non-concurrent maintenance (within scope of the OS)
- Managing firmware levels
  - Server
  - Adapter
- Remote support (access varies by OS)

Service Focal Point on the Hardware Management Console

Service strategies become more complicated in a partitioned environment. The Manage Serviceable Events task in the HMC can help to streamline this process.

Each logical partition reports errors that it detects and forwards the event to the Service Focal Point (SFP) application that is running on the HMC, 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.

In addition, you can use the Manage Serviceable Events task to initiate service functions on systems and logical partitions, including the exchanging of parts, configuring connectivity, and managing dumps.

The following functions are available through the Service Focal Point on the Hardware Management Console:

- Service Focal Point
  - Managing serviceable events and service data
  - Managing service indicators
- Error information
  - OS diagnostic
  - Service processor
  - Service Focal Point
- LED Management menu
- Serviceable events analysis
- Repair and verify
  - Concurrent maintenance
  - Deferred maintenance
  - Immediate maintenance
- Hot-node add, hot-node repair, and memory upgrade
- FRU replacement
- Managing firmware levels
  - HMC
  - Server
  - Adapter
  - Concurrent firmware updates
- Call home/customer notification
- Virtualization
- I/O Topology view
- Generate dump
- Remote support (full access)
- Virtual operator panel

Service Focal Point Lite on the Integrated Virtualization Manager

The following functions are available through the Service Focal Point Lite on the Integrated Virtualization Manager:

- Service Focal Point-Lite
  - Managing serviceable events and service data
  - Managing service indicators
- Error Information menu
  - OS Diagnostic
  - Service Focal Point lite
- LED Management menu
- Managing firmware levels
  - Server
  - Adapter
- Virtualization
- Generate dump (limited capability)
- Remote support (limited access)

4.4.2 IBM Power Systems firmware maintenance

The IBM Power Systems Client-Managed Microcode is a methodology that enables you to manage and install microcode updates on Power Systems and associated I/O adapters.

The system firmware consists of service processor microcode, Open Firmware microcode, SPCN microcode, and the POWER Hypervisor.

The firmware and microcode can be downloaded and installed either from an HMC, from a running partition, or from USB port number one (1) on the rear of Power Systems, if that system is not managed by an HMC.

Power Systems has a permanent firmware boot side (A side), and a temporary firmware boot side (B side). New levels of firmware must be installed on the temporary side first to test the update's compatibility with existing applications. When the new level of firmware is approved, it can be copied to the permanent side.

For access to the initial web pages that address this capability, see Support for IBM Systems web page:

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

For Power Systems, select the **Power** link. Figure 4-8 shows an example.

Figure 4-8 Support for Power servers web page

Although the content under the Popular links section can change, click the **Firmware and HMC updates** link to go to the resources for keeping your system's firmware current.

If there is an HMC to manage the server, the HMC interface can be used to view the levels of server firmware and power subsystem firmware that are installed and are available to download and install.

Each IBM Power Systems server has the following levels of server firmware and power subsystem firmware:

- Installed level

  This level of server firmware or power subsystem firmware is installed and will be installed into memory after the managed system is powered off and then powered on. It is installed on the temporary side of system firmware.

- Activated level

  This level of server firmware or power subsystem firmware is active and running in memory.

- Accepted level

  This level 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 permanent side of system firmware.

IBM provides the Concurrent Firmware Maintenance (CFM) function on selected Power Systems. This function supports applying non disruptive system firmware service packs to the system concurrently (without requiring a reboot operation 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 that are known as *deferred*. These deferred fixes can be installed concurrently but are not activated until the next IPL. Deferred fixes, if any, will be identified in the Firmware Update Descriptions table of the firmware documentation. For deferred fixes within a service pack, only the fixes in the service pack which cannot be concurrently activated are deferred. Table 4-2 shows the file-naming convention for system firmware.

Table 4-2 Firmware naming convention

PPNNSSS_FFF_DDD			
Firmware component ID	**Description**	**Definitions**	
PP	Package identifier	01	-
		02	-
NN	Platform and class	AL	Low End
		AM	Mid Range
		AS	IH Server
		AH	High End
		AP	Bulk Power for IH
		AB	Bulk Power for High End
SSS	Release indicator		
FFF	Current fixpack		
DDD	Last disruptive fixpack		

The following example uses the convention:

```
01AM710_086_063 = Managed System Firmware for 9117-MMB Release 710 Fixpack 086
```

An installation is disruptive if the following statements are true:

- The release levels (SSS) of currently installed and new firmware differ.
- The service pack level (FFF) and the last disruptive service pack level (DDD) are equal in new firmware.

Otherwise, 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 conditions for disruptive installation are not met.

4.4.3 Electronic Services and Electronic Service Agent

IBM transformed its delivery of hardware and software support services to help you achieve higher system availability. Electronic Services is a web-enabled solution that offers an exclusive, no-additional-charge enhancement to the service and support available for IBM servers. These services provide the opportunity for greater system availability with faster problem resolution and preemptive monitoring. The Electronic Services solution consists of two separate, but complementary, elements:

- Electronic Services news page

  The Electronic Services news page is a single Internet entry point that replaces the multiple entry points that are traditionally used to access IBM Internet services and support. The news page enables you to gain easier access to IBM resources for assistance in resolving technical problems.

- IBM Electronic Service Agent

  The Electronic Service Agent is software that resides on your server. It monitors events and transmits system inventory information to IBM on a periodic, client-defined timetable. The Electronic Service Agent automatically reports hardware problems to IBM.

Early knowledge about potential problems enables IBM to deliver proactive service that can result in higher system availability and performance. In addition, information that is collected through the Service Agent is made available to IBM service support representatives when they help answer your questions or diagnose problems. Installation and use of IBM Electronic Service Agent for problem reporting enables IBM to provide better support and service for your IBM server.

To learn how Electronic Services can work for you, visit the following site; an IBM ID is required:

https://www.ibm.com/support/electronic/portal

4.5 Operating system support for RAS features

Table 4-3 gives an overview of several features for continuous availability that are supported by the various operating systems that run on the POWER7 processor-based systems.

Table 4-3 Operating system support for RAS features

RAS feature	AIX 5.3	AIX 6.1	AIX 7.1	IBM i 6.1.1	IBM i 7.1	RHEL 5	SLES 10 SP 3	SLES 11 SP 1
System deallocation of failing components								
Dynamic processor deallocation	✓	✓	✓	✓	✓	✓	✓	✓
Dynamic processor sparing	✓	✓	✓	✓	✓	✓	✓	✓
Processor instruction retry	✓	✓	✓	✓	✓	✓	✓	✓
Alternate processor recovery	✓	✓	✓	✓	✓	✓	✓	✓
Partition contained checkstop	✓	✓	✓	✓	✓	✓	✓	✓
Persistent processor deallocation	✓	✓	✓	✓	✓	✓	✓	✓
GX++ bus persistent deallocation	✓	✓	✓	✓	✓	—	—	✓
PCI bus extended error detection	✓	✓	✓	✓	✓	✓	✓	✓
PCI bus extended error recovery	✓	✓	✓	✓	✓	Most	Most	Most
PCI-PCI bridge extended error handling	✓	✓	✓	✓	✓	—	—	—
Redundant RIO or 12X channel link	✓	✓	✓	✓	✓	✓	✓	✓
PCI card hot-swap	✓	✓	✓	✓	✓	✓	✓	✓
Dynamic SP failover at run time	✓	✓	✓	✓	✓	✓	✓	✓
Memory sparing with CoD at IPL time	✓	✓	✓	✓	✓	✓	✓	✓
Clock failover run time or IPL	✓	✓	✓	✓	✓	✓	✓	✓
Memory availability								
64-byte ECC code	✓	✓	✓	✓	✓	✓	✓	✓
Hardware scrubbing	✓	✓	✓	✓	✓	✓	✓	✓
CRC	✓	✓	✓	✓	✓	✓	✓	✓
Chipkill	✓	✓	✓	✓	✓	✓	✓	✓
L1 instruction and data array protection	✓	✓	✓	✓	✓	✓	✓	✓
L2/L3 ECC & cache line delete	✓	✓	✓	✓	✓	✓	✓	✓
Special uncorrectable error handling	✓	✓	✓	✓	✓	✓	✓	✓
Fault detection and isolation								
Platform FFDC diagnostics	✓	✓	✓	✓	✓	✓	✓	✓
Run-time diagnostics	✓	✓	✓	✓	✓	Most	Most	Most
Storage protection keys	—	✓	✓	✓	✓	—	—	—

RAS feature	AIX 5.3	AIX 6.1	AIX 7.1	IBM i 6.1.1	IBM i 7.1	RHEL 5	SLES 10 SP 3	SLES 11 SP 1
Dynamic trace	✓	✓	✓	✓	✓	—	—	✓
Operating system FFDC	—	✓	✓	✓	✓	—	—	—
Error log analysis	✓	✓	✓	✓	✓	✓	✓	✓
Service processor support for:								
► Built-in-self-tests (BIST) for logic and arrays	✓	✓	✓	✓	✓	✓	✓	✓
► Wire tests	✓	✓	✓	✓	✓	✓	✓	✓
► Component initialization	✓	✓	✓	✓	✓	✓	✓	✓
Serviceability								
Boot-time progress indicators	✓	✓	✓	✓	✓	Most	Most	Most
Firmware error codes	✓	✓	✓	✓	✓	✓	✓	✓
Operating system error codes	✓	✓	✓	✓	✓	Most	Most	Most
Inventory collection	✓	✓	✓	✓	✓	✓	✓	✓
Environmental and power warnings	✓	✓	✓	✓	✓	✓	✓	✓
Hot-plug fans, power supplies	✓	✓	✓	✓	✓	✓	✓	✓
Extended error data collection	✓	✓	✓	✓	✓	✓	✓	✓
SP “call home” on non-HMC configurations	✓	✓	✓	✓	✓	✓	✓	✓
I/O drawer redundant connections	✓	✓	✓	✓	✓	✓	✓	✓
I/O drawer hot add and concurrent repair	✓	✓	✓	✓	✓	✓	✓	✓
Concurrent RIO/GX adapter add	✓	✓	✓	✓	✓	✓	✓	✓
Concurrent cold-repair of GX adapter	✓	✓	✓	✓	✓	✓	✓	✓
Concurrent add of powered I/O rack to Power 595	✓	✓	✓	✓	✓	✓	✓	✓
SP mutual surveillance with POWER Hypervisor	✓	✓	✓	✓	✓	✓	✓	✓
Dynamic firmware update with HMC	✓	✓	✓	✓	✓	✓	✓	✓
Service Agent Call Home Application	✓	✓	✓	✓	✓	✓	✓	✓
Guiding light LEDs	✓	✓	✓	✓	✓	✓	✓	✓
Lightpath LEDs	✓	✓	✓	✓	✓	✓	✓	✓
System dump for memory, POWER Hypervisor, SP	✓	✓	✓	✓	✓	✓	✓	✓
Information center / systems support site service publications	✓	✓	✓	✓	✓	✓	✓	✓

RAS feature	AIX 5.3	AIX 6.1	AIX 7.1	IBM i 6.1.1	IBM i 7.1	RHEL 5	SLES 10 SP 3	SLES 11 SP 1
Systems support site education	✓	✓	✓	✓	✓	✓	✓	✓
Operating system error reporting to HMC SFP	✓	✓	✓	✓	✓	✓	✓	✓
RMC secure error transmission subsystem	✓	✓	✓	✓	✓	✓	✓	✓
Health check scheduled operations with HMC	✓	✓	✓	✓	✓	✓	✓	✓
Operator panel (real or virtual)	✓	✓	✓	✓	✓	✓	✓	✓
Concurrent operator panel maintenance	✓	✓	✓	✓	✓	✓	✓	✓
Redundant HMCs	✓	✓	✓	✓	✓	✓	✓	✓
Automated server recovery/restart	✓	✓	✓	✓	✓	✓	✓	✓
High availability clustering support	✓	✓	✓	✓	✓	✓	✓	✓
Repair and verify guided maintenance	✓	✓	✓	✓	✓	Most	Most	Most
Concurrent kernel update	—	✓	✓	✓	✓	—	—	—

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 publications” on page 207. Note that some of the documents referenced here might be available in softcopy only.

- *IBM Power 710 and 730 Technical Overview and Introduction*, REDP-4636
- *IBM Power 720 and 740 Technical Overview and Introduction*, REDP-4637
- *IBM Power Systems: SDMC to HMC Migration Guide (RAID1)*, REDP-4872
- *IBM PowerVM Live Partition Mobility*, SG24-7460
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *IBM PowerVM Virtualization on IBM System p: Introduction and Configuration Fourth Edition*, SG24-7940
- *IBM System p Advanced POWER Virtualization (PowerVM) Best Practices*, REDP-4194
- *IBM System Storage DS8000: Copy Services in Open Environments*, SG24-6788
- *IBM System Storage DS8700 Architecture and Implementation*, SG24-8786
- *PowerVM and SAN Copy Services*, REDP-4610
- *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- *SAN Volume Controller V4.3.0 Advanced Copy Services*, SG24-7574

Other publications

These publications are also relevant as further information sources:

- IBM Power Systems Facts and Features POWER7 Blades and Servers:
  ftp://public.dhe.ibm.com/common/ssi/ecm/en/pob03022usen/POB03022USEN.PDF
- Specific storage devices supported for Virtual I/O Server:
  http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html
- IBM Power 710 server Data Sheet:
  ftp://public.dhe.ibm.com/common/ssi/ecm/en/pod03048usen/POD03048USEN.PDF
- IBM Power 720 server Data Sheet:
  ftp://public.dhe.ibm.com/common/ssi/ecm/en/pod03049usen/POD03049USEN.PDF
- IBM Power 730 server Data Sheet:
  ftp://public.dhe.ibm.com/common/ssi/ecm/en/pod03050usen/POD03050USEN.PDF

- IBM Power 740 server Data Sheet:

  ftp://public.dhe.ibm.com/common/ssi/ecm/en/pod03051usen/POD03051USEN.PDF
- IBM Power 750 server Data Sheet:

  http://ibm.co/TDUIHs
- IBM Power 755 server Data Sheet:

  http://ibm.co/1OEuJVA
- IBM Power 770 server Data Sheet:

  http://ibm.co/1OEuR7q
- IBM Power 780 server Data Sheet:

  http://ibm.co/WGz5FO
- IBM Power 795 server Data Sheet:

  ftp://public.dhe.ibm.com/common/ssi/ecm/en/pod03053usen/POD03053USEN.PDF
- *Active Memory Expansion: Overview and Usage Guide*:

  http://www.ibm.com/systems/power/hardware/whitepapers/am_exp.html
- Migration combinations of processor compatibility modes for active Partition Mobility:

  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7hc3/iphc3pcmcombosact.htm
- Advance Toolchain for Linux website:

  http://www.ibm.com/developerworks/wikis/display/hpccentral/How+to+use+Advance+Toolchain+for+Linux+on+POWER

Online resources

These websites are also relevant as further information sources:

- IBM Power Systems Hardware Information Center:

  http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp
- IBM System Planning Tool website:

  http://www.ibm.com/systems/support/tools/systemplanningtool/
- Download from the IBM Fix Central website:

  http://www.ibm.com/support/fixcentral/
- Power Systems Capacity on Demand website:

  http://www.ibm.com/systems/power/hardware/cod/
- Support for IBM Systems website:

  http://www.ibm.com/systems/support
- IBM Electronic Services portal:

  https://www.ibm.com/support/electronic/portal
- IBM Storage website:

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

How to get Redbooks publications

You can search for, view, or download Redbooks publications, Redpapers publications, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks publications, at this website:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM Power 795 (9119-FHB)

Technical Overview and Introduction

Explores class-leading performance, energy efficiency, and scalability

Shows how the design can offer maximum availability

Features information about new I/O cards

This IBM Redpaper publication is a comprehensive guide that covers the IBM Power 795 server that supports IBM AIX, IBM i, and Linux operating systems. The goal of this paper is to introduce the innovative Power 795 offering and its major functions:

- IBM POWER7 processor available at frequencies of 3.7 GHz and 4.0 GHz with TurboCore options of 4.25 GHz and 4.31 GHz
- Specialized POWER7 Level 3 cache that provides greater bandwidth, capacity, and reliability
- IBM PowerVM virtualization, including PowerVM Live Partition Mobility and PowerVM IBM Active Memory Sharing
- TurboCore mode that delivers the highest performance per core
- Enhanced reliability, accessibility, and serviceability (RAS) features that are designed for maximum availability
- Active Memory Expansion that provides more usable memory than what is physically installed on the system
- IBM EnergyScale technology that provides features such as power trending, power-saving, capping of power, and thermal measurement

Professionals who want to acquire a better understanding of IBM Power Systems products can benefit from reading this paper.

This paper complements the available set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power 795 system.

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

REDP-4640-01

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