

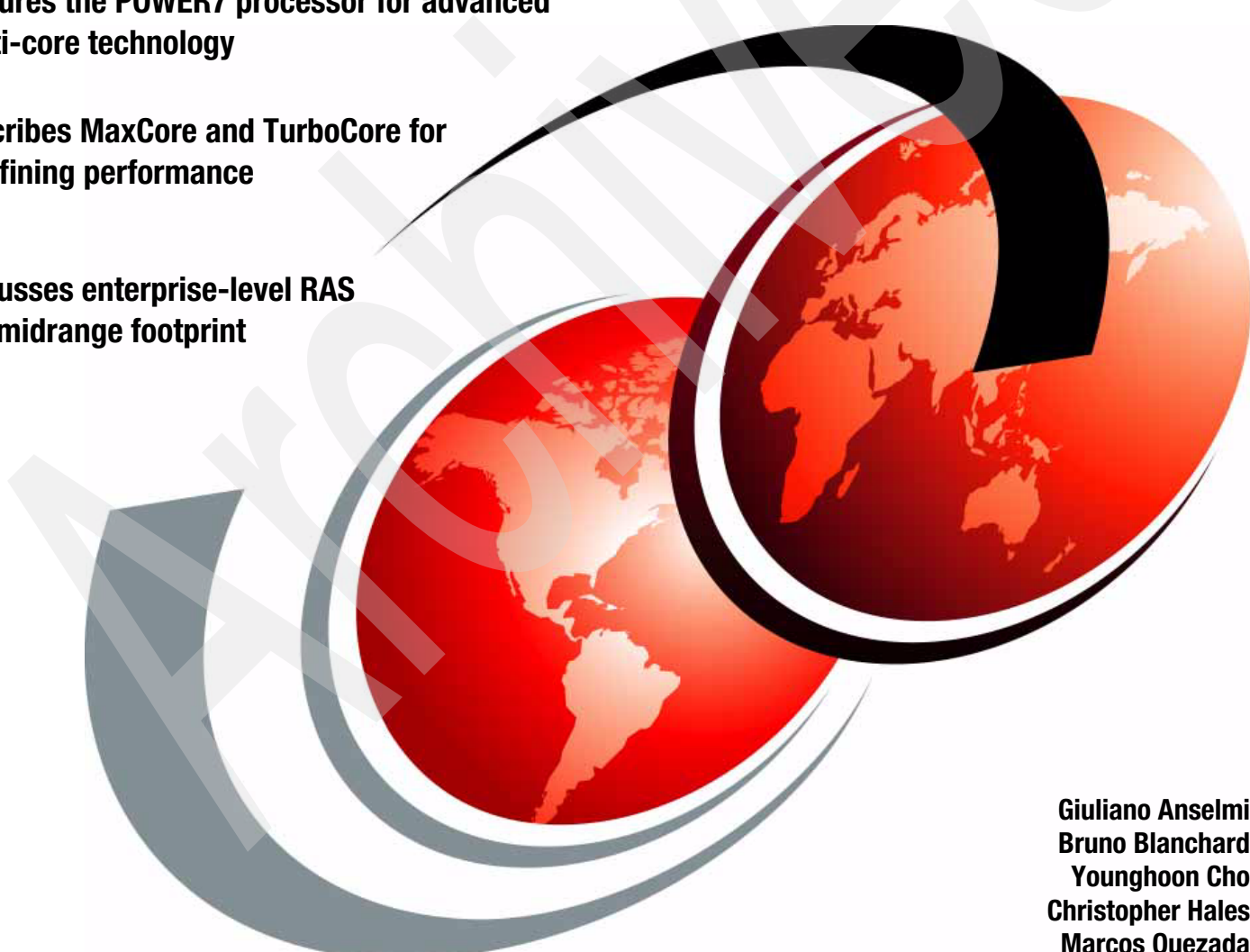
IBM Power 770 and 780

Technical Overview and Introduction

Features the POWER7 processor for advanced multi-core technology

Describes MaxCore and TurboCore for redefining performance

Discusses enterprise-level RAS in a midrange footprint



Giuliano Anselmi
Bruno Blanchard
Younghoon Cho
Christopher Hales
Marcos Quezada



International Technical Support Organization

IBM Power 770 and 780 Technical Overview and Introduction

March 2010

Archived

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

Archived

First Edition (March 2010)

This edition applies to the IBM Power 770 (9117-MMB) and IBM Power 780 (9179-MHB) Power Systems servers.

© Copyright International Business Machines Corporation 2010. All rights reserved.

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

Contents

Notices	vii
Trademarks	viii
Preface	ix
The team who wrote this paper	ix
Now you can become a published author, too!	xi
Comments welcome	xi
Stay connected to IBM Redbooks	xi
Chapter 1. General description	1
1.1 Overview of systems	2
1.2 Operating environment	3
1.3 Physical package	3
1.4 System features	5
1.4.1 Power 770 system features	5
1.4.2 Power 780 system features	6
1.4.3 Minimum features	7
1.4.4 Power supply features	9
1.4.5 Processor card features	9
1.4.6 Summary of processor features	10
1.4.7 Memory features	13
1.5 Disk and media features	16
1.6 I/O drawers	17
1.6.1 PCI-DDR 12X Expansion Drawers (#5796)	17
1.6.2 12X I/O Drawer PCIe (#5802 and #5877)	17
1.6.3 I/O drawers and usable PCI slot	17
1.7 Comparison between models	18
1.8 Build to Order	19
1.9 IBM Editions	19
1.10 Model upgrade	19
1.10.1 Upgrade considerations	19
1.11 Hardware Management Console models	20
1.12 System racks	21
1.12.1 IBM 7014 Model T00 rack	21
1.12.2 IBM 7014 Model T42 rack	22
1.12.3 IBM 7014 Model S25 rack	22
1.12.4 Feature number 0555 rack	23
1.12.5 Feature number 0551 rack	23
1.12.6 Feature number 0553 rack	23
1.12.7 The AC power distribution unit and rack content	23
1.12.8 Rack-mounting rules	25
1.12.9 Useful rack additions	25
Chapter 2. Architecture and technical overview	27
2.1 The IBM POWER7 processor	29
2.1.1 POWER7 processor overview	30
2.1.2 POWER7 processor core	31
2.1.3 Simultaneous multithreading	32
2.1.4 Memory access	33

2.1.5 Flexible POWER7 processor packaging and offerings	33
2.1.6 On-chip L3 cache innovation and Intelligent Cache	34
2.1.7 POWER7 processor and Intelligent Energy	36
2.1.8 Comparison of the POWER7 and POWER6 processors	36
2.2 POWER7 processor cards	36
2.3 Memory subsystem	38
2.3.1 Fully buffered DIMM	38
2.3.2 Memory placement rules.	38
2.3.3 Memory throughput.	43
2.4 Capacity on Demand.	44
2.4.1 Capacity Upgrade on Demand (CUoD).	44
2.4.2 On/Off Capacity on Demand (On/Off CoD).	44
2.4.3 Utility Capacity on Demand (Utility CoD)	45
2.4.4 Trial Capacity On Demand (Trial CoD).	45
2.4.5 Software licensing and CoD	46
2.5 Drawer interconnection cables	46
2.6 System bus	50
2.6.1 I/O buses and GX++ card	50
2.6.2 FSP bus	51
2.7 Internal I/O subsystem	51
2.7.1 Blind-swap cassettes	52
2.7.2 System ports	52
2.8 Integrated Virtual Ethernet adapter	52
2.8.1 IVE subsystem and feature codes	53
2.9 PCI adapters	54
2.9.1 LAN adapters	55
2.9.2 Graphics accelerators	55
2.9.3 SCSI and SAS adapters	56
2.9.4 iSCSI	57
2.9.5 Fibre Channel adapter	57
2.9.6 Fibre Channel over Ethernet (FCoE)	58
2.9.7 InfiniBand Host Channel adapter	59
2.9.8 Asynchronous adapter	59
2.10 Internal storage	59
2.10.1 Dual split backplane mode	61
2.10.2 Triple split backplane	62
2.10.3 Dual storage IOA configurations	63
2.10.4 DVD	64
2.11 External I/O subsystems	64
2.11.1 PCI-DDR 12X Expansion drawer (#5796).	65
2.11.2 12X I/O Drawer PCIe (#5802 and #5877).	66
2.11.3 Dividing SFF drive bays in 12X I/O drawer PCIe	68
2.11.4 12X I/O Drawer PCIe and PCI-DDR 12X Expansion Drawer 12X cabling	70
2.11.5 12X I/O Drawer PCIe and PCI-DDR 12X Expansion Drawer SPCN cabling	71
2.12 External disk subsystems	73
2.12.1 EXP 12S Expansion Drawer	73
2.12.2 EXP 24S Expansion Drawer	74
2.12.3 IBM System Storage	74
2.13 Hardware Management Console	76
2.13.1 HMC Functional overview	77
2.13.2 HMC connectivity to the POWER7 processor based systems	78
2.13.3 High availability using the HMC	80
2.13.4 HMC code level.	84

2.14 Operating system support	84
2.14.1 Virtual I/O Server	85
2.14.2 IBM AIX operating system	85
2.14.3 IBM i operating system	87
2.14.4 Linux operating system	87
2.15 Compiler technology	88
2.16 Energy management	89
2.16.1 IBM EnergyScale technology	89
2.16.2 Thermal power management device (TPMD) card	91
Chapter 3. Virtualization	93
3.1 POWER Hypervisor	94
3.2 POWER processor modes	96
3.3 Active Memory Expansion	98
3.4 PowerVM	102
3.4.1 PowerVM editions	102
3.4.2 Logical partitions (LPARs)	103
3.4.3 Multiple Shared-Processor Pools	106
3.4.4 Virtual I/O Server	111
3.4.5 PowerVM Lx86	115
3.4.6 PowerVM Live Partition Mobility	115
3.4.7 Active Memory Sharing	117
3.4.8 NPIV	118
3.4.9 Operating System support for PowerVM	119
3.4.10 POWER7 Linux programming support	119
3.5 System Planning Tool	121
Chapter 4. Continuous availability and manageability	123
4.1 Reliability	125
4.1.1 Designed for reliability	125
4.1.2 Placement of components	126
4.1.3 Redundant components and concurrent repair	126
4.2 Availability	126
4.2.1 Partition availability priority	127
4.2.2 General detection and deallocation of failing components	127
4.2.3 Memory protection	129
4.2.4 Cache protection	134
4.2.5 Special uncorrectable error handling	134
4.2.6 PCI enhanced error handling	135
4.3 Serviceability	136
4.3.1 Detecting	137
4.3.2 Diagnosing	140
4.3.3 Reporting	141
4.3.4 Notifying	143
4.3.5 Locating and servicing	144
4.4 Manageability	148
4.4.1 Service user interfaces	148
4.4.2 IBM Power Systems firmware maintenance	155
4.4.3 Electronic Services and Electronic Service Agent	158
4.5 Operating system support for RAS features	159
Related publications	163
IBM Redbooks	163
Online resources	163

How to get Redbooks 165
Help from IBM 165

Archived

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

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

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

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

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

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

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

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


COPYRIGHT LICENSE:

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

Trademarks

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 .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™	POWER Hypervisor™	POWER®
AIX 5L™	Power Systems™	pSeries®
AIX®	Power Systems Software™	Rational®
DB2®	POWER4™	Redbooks®
DS8000®	POWER4+™	Redpaper™
Electronic Service Agent™	POWER5™	Redbooks (logo)  ®
EnergyScale™	POWER5+™	RS/6000®
FlashCopy®	POWER6+™	System p®
Focal Point™	POWER6®	System Storage®
IBM Systems Director Active Energy Manager™	POWER7™	System z®
IBM®	PowerHA™	Tivoli®
Micro-Partitioning™	PowerPC®	TotalStorage®
	PowerVM™	XIV®

The following terms are trademarks of other companies:

InfiniBand Trade Association, InfiniBand, and the InfiniBand design marks are trademarks and/or service marks of the InfiniBand Trade Association.

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

SnapManager, and the NetApp logo are trademarks or registered trademarks of NetApp, Inc. in the U.S. and other countries.

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

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

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

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

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

Preface

This IBM® Redpaper™ publication is a comprehensive guide covering the IBM Power 770 and Power 780 servers supporting IBM AIX®, IBM i, and Linux operating systems. The goal of this paper is to introduce the major innovative Power 770 and 780 offerings and their prominent functions, including:

- ▶ Unique modular server packaging
- ▶ The specialized IBM POWER7™ Level 3 cache that provides greater bandwidth, capacity, and reliability
- ▶ The 1 Gb or 10 Gb Integrated Virtual Ethernet adapter that brings native hardware virtualization up to 64 logical ports on this server
- ▶ IBM PowerVM™ virtualization including PowerVM Live Partition Mobility and PowerVM Active Memory™ Sharing
- ▶ 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
- ▶ Enterprise-ready reliability, serviceability, and availability

Professionals who want to acquire a better understanding of IBM Power Systems™ products should read this Redpaper. The intended audience includes the following areas:

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

This Redpaper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the 770 and 780 systems.

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, may 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, Austin Center.

Giuliano Anselmi has worked with IBM Power Systems for 18 years. He was previously a pSeries® Systems Product Engineer for seven years, supporting various IBM organizations, Business Partners, and Technical Support Organizations. He joined the IBM Technical Sales Support group in 2004 and was certified as an IT Specialist in 2009 after he was an IBM System Architect with the IBM Systems and Technology Group (STG) for three years. Giuliano currently works in Italy for Makram Srl, a company that offers IT Management, Business Continuity and Disaster Recovery adding value services that focus on IBM Power Systems and IBM Storage platforms.

Bruno Blanchard is a Certified IT Specialist with IBM in France, working in Integrated Technology Delivery. He has been with IBM for 26 years, and has 20 years of experience in AIX and IBM pSeries. He has written several IBM Redbooks® publications. He is currently involved as an IT Architect in projects that deploy Power Systems in on-demand data centers, server consolidation environments, and large server farms. His areas of expertise also include virtualization, clouds, and operating system provisioning.

Younghoon Cho is a Power Systems Top Gun with the post-sales Technical Support Team for IBM in Korea. He has nine years of experience working on RS/6000®, System p®, and Power Systems products. He is an IBM Certified Specialist in System p and AIX 5L™. He provides second-line technical support to Field Engineers working on Power Systems and system management.

Christopher Hales is a Consulting IT Specialist based in the U.K. Chris has been designing IT solutions with customers for over 25 years and he specializes in virtualization technologies. In 2007, he attended an internship at the IBM development labs in Austin, working on the Power 595 servers. He delivered the POWER7 technology lecture to the European STG Technical Conference in 2009 and was a keynote speaker at the announcement of POWER7 processors and Power Systems in London in February 2010. He has recently been given an Outstanding Technology Achievement Award and an Invention Achievement Award by IBM for his work on Multiple Shared-Processor Pools. Chris holds an Honors degree in computer science.

Marcos Quezada is a Brand Development Manager for Power Systems in Argentina. He is a Certified IT Specialist with 12 years of IT experience as a UNIX systems Pre-sales Specialist and as a Web Project Manager. He holds a degree in Informatics Engineering from Fundación Universidad de Belgrano. His areas of expertise include POWER® processor-based servers under the AIX operating system and pre-sales support of IBM Software, SAP, and Oracle architecture solutions that run on IBM UNIX Systems, with a focus on competitive accounts.

The project that produced this publication was managed by **Scott Vetter**, PMP

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

George Ahrens, Mark Applegate, Ron Arroyo, Gail Belli, Terri Brennan, Herve de Caceres, Anirban Chatterjee, Ben Gibbs, Marianne Golden, Stephen Hall, Daniel J. Henderson, David Hepkin, John Hock, Craig G. Johnson, Deanna M. Johnson, Roxette Johnson, Ronald Kalla, Bob Kovacs, Phil N. Lewis, Casey McCreary, Michael Middleton, Bill Moran, Michael J Mueller, Jeff Meute, Steve Munroe, Luiz Gustavo Nascimento, Duc Nguyen, Thoi Nguyen, Jonathan Van Niewaal, Mark Olson, Patrick O'Rourke, Jan Palmer, Amartey Pearson, David Pirnik, Audrey Romonosky, Jeffrey Scheel, Kimberly Schmid, Helena Sunny, Joel Tandler, Jeff Van Heuklon, Jez Wain, Steve Will, Ian Wills

Tamikia Barrow, Emma Jacobs, Diane Sherman
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 e-mail to:
redbooks@us.ibm.com
- ▶ Mail your comments to:
IBM Corporation, International Technical Support Organization
Dept. HYTD Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:
<http://www.facebook.com/pages/IBM-Redbooks/178023492563?ref=ts>
- ▶ 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 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>

Archived



General description

The IBM Power 770 (9117-MMB) and IBM Power 780 servers (9179-MHB) utilize the latest POWER7 processor technology designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads.

The innovative IBM Power 770 and IBM Power 780 servers with POWER7 processors are symmetric multiprocessing (SMP), rack-mounted servers. These modular-built systems use one to four enclosures; each enclosure is four EIA units (4U) tall and is housed in a 19-inch rack.

1.1 Overview of systems

Figure 1-1 shows a Power 770 with the maximum four enclosures, and the front and rear views of a single-enclosure Power 770.

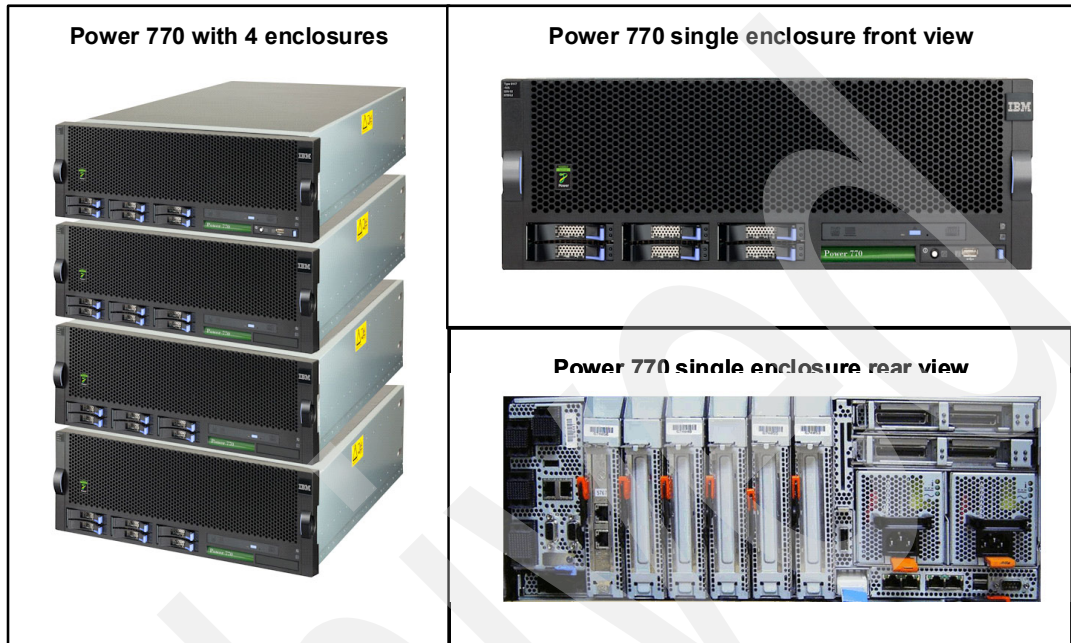


Figure 1-1 Four-enclosure Power 770, a single-enclosure Power 770 front and rear views.

Each of the four system enclosures contains one powerful POWER7 processor card feature, consisting of two single-chip module processors. Each of the POWER7 processors in the server has a 64-bit architecture, includes up to eight processor cores on a single-chip module (SCM), and contains up to 2 MB of L2 cache (256 KB per core) and up to 32 MB of L3 cache (4 MB per core).

The Power 770 model of server is available starting as low as four active cores; it increases in increments of one core at a time through built-in Capacity on Demand (CoD) functions.

The POWER7 DDR3 memory uses a new memory architecture to provide greater bandwidth and capacity. This enables operating at a higher data rate for large memory configurations.

Each new POWER7 processor can support up to eight DDR3 DIMMs running at speeds up to 1066 MHz. A full system can contain up to 1.0 TB of memory.

IBM Power 770 server

For the Power 770, each POWER7 processor SCM is available at frequencies of 3.1 GHz with eight cores and 3.5 GHz with six cores.

IBM Power 780 server

For the Power 780 server, each POWER7 SCM processor is available at frequencies of 3.86 GHz with eight cores and 4.14 GHz with four cores.

What makes the Power 780 truly unique is the ability to switch between its standard throughput optimized mode and its unique TurboCore mode. In TurboCore mode performance per core is boosted with access to both additional cache and additional clock speed.

Based on the user's configuration option, any Power 780 system can be booted in standard mode, enabling up to a maximum of 64 processor cores running at 3.86 GHz, or in TurboCore mode, enabling up to 32 processor cores running at 4.14 GHz and twice the cache per core.

1.2 Operating environment

The operating environment specifications for the servers are listed in Table 1-1:

Table 1-1 Operating environment for Power 770 and Power 780 (for one enclosure only)

Power 770 and Power 780 operating environment		
Description	Operating	Non-operating
Temperature	5 - 35 degrees C (41 to 95 degrees F)	5 - 45 degrees C (41 - 113 degrees F)
Relative humidity	20 - 80%	8 - 80%
Maximum dew point	29 degrees C (84 degrees F)	28 degrees C (82 degrees F)
Operating voltage	200 - 240 V ac	Not applicable
Operating frequency	50 - 60 +/- 3 Hz	Not applicable
Power consumption	1600 watts maximum per enclosure with 16 cores active	Not applicable
Power source loading	1.649 kVA maximum per enclosure with 16 cores active	Not applicable
Thermal output	5461 BTU/hr maximum per enclosure with 16 cores active	Not applicable
Maximum altitude	3048 m (10,000 ft)	Not applicable
Noise level One enclosure with 16 active cores	6.8 bels (operating/idle) 6.3 bels (operating/idle) with acoustic rack doors	
Noise level Four enclosures with 64 active cores	7.4 bels (operating/idle) 6.9 bels (operating/idle) with acoustic rack doors	

1.3 Physical package

Table 1-2 on page 4 lists the physical dimensions of an individual enclosure. Both servers are available only in a rack-mounted form factor. They are modular systems that can be constructed from one to four building-block enclosures. Each of these enclosures can take 4U (EIA units) of rack space. Thus, a two enclosure system requires 8U, three enclosures require 12U, and four enclosures require 16U.

Table 1-2 Physical dimensions of a Power 770 and Power 780 enclosure

Dimension	Power 770 (Model 9117-MMB) single enclosure	Power 780 (Model 9179-MHB) single enclosure
Width	483 mm (19.0 in.)	483 mm (19.0 in.)
Depth	863 mm (32.0 in.)	863 mm (32.0 in.)
Height	174 mm (6.85 in.), 4U (EIA units)	174 mm (6.85 in.), 4U (EIA units)
Weight	70.3 kg (155 lb)	70.3 kg (155 lb)

The front and rear views of the Power 770 and Power 780 are shown in Figure 1-2.

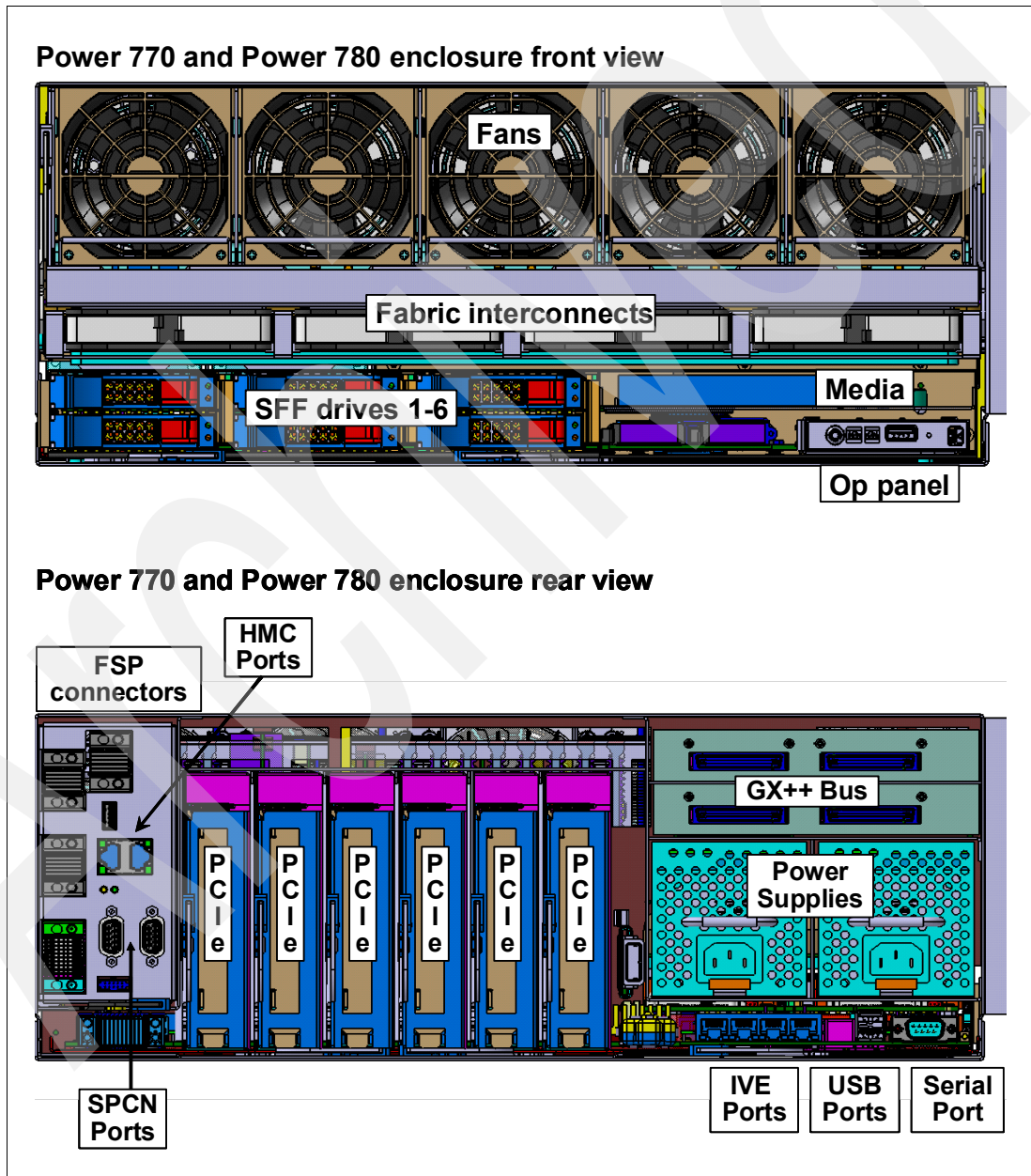


Figure 1-2 Front and rear views of the Power 770 and Power 780

1.4 System features

Each of the four system enclosures contains one powerful POWER7 processor card feature, consisting of two single-chip module processors. Each of the POWER7 processors in the server has a 64-bit architecture, includes up to eight processor cores on a single-chip module (SCM), and contains up to 2 MB of L2 cache (256 KB per core) and up to 32 MB of L3 cache (4 MB per core).

1.4.1 Power 770 system features

The following features are available on the Power 770:

- ▶ 4U 19-inch rack-mount system enclosure
- ▶ One to four system enclosures: 16U maximum system size
- ▶ One processor card feature per enclosure (includes the voltage regulator):
 - 0/12 way, 3.5 GHz processor card (#4980)
 - 0/16 way, 3.1 GHz processor card (#4981)
- ▶ POWER7 DDR3 Memory DIMMs (16 DIMM slots per processor card):
 - 0/32 GB (4 X 8 GB), 1066 MHz (#5600)
 - 0/64 GB (4 X 16 GB), 1066 MHz (#5601)
 - 0/128 GB (4 X 32 GB), 1066 MHz (#5602)
- ▶ Six hot-swappable, 2.5 inch, small form factor, SAS disk or SSD bays per enclosure
- ▶ One hot-plug, slim-line, SATA media bay per enclosure (optional)
- ▶ Redundant hot-swap AC power supplies in each enclosure
- ▶ Choice of integrated (HEA) I/O options; one per enclosure:
 - Quad 1 Gb Ethernet
 - Dual 10 Gb Optical + Dual 1 Gb Ethernet
 - Dual 10 Gb Copper + Dual 1 Gb Ethernet
- ▶ One serial port, three USB ports per enclosure (maximum nine per system)
- ▶ Two HMC ports per enclosure (maximum four per system)
- ▶ Eight I/O expansion slots per enclosure (maximum 32 per system)
 - Six PCIe 8x slots plus two GX++ slots per enclosure
- ▶ Dynamic LPAR support, Processor and Memory CUoD
- ▶ PowerVM (optional):
 - Micro-Partitioning™
 - Virtual I/O Server (VIOS)
 - Automated CPU and memory reconfiguration support for dedicated and shared processor logical partition groups (dynamic LPAR)
 - Support for manual provisioning of resources namely PowerVM Live Partition Migration (PowerVM Enterprise Edition)
- ▶ Optional PowerHA™ for AIX, IBM i and Linux

- ▶ 12X I/O drawer with PCI slots:
 - Up to 16 PCIe I/O drawers (#5802 or #5877)
 - Up to 32 PCI-X DDR I/O drawers (7314-G30 or #5796)
- ▶ Disk-only I/O drawers:
 - Up to 110 EXP12S SAS DASD/SSD I/O drawers on SAS PCI controllers (#5886)
 - Up to 60 EXP24 SCSI DASD Expansion drawers on SCSI PCI controllers (7031-D24)
- ▶ IBM Systems Director Active Energy Manager™

The Power 770 operator interface controls are located on the front panel of the primary I/O drawer consist of a power ON/OFF button with a POWER indicator, LCD display for diagnostic feedback, a RESET button, and a disturbance or system attention LED.

1.4.2 Power 780 system features

The following features are available on the Power 780:

- ▶ 4U 19-inch rack-mount system enclosure
- ▶ One to four system enclosures: 16U maximum system size
- ▶ One processor card feature per enclosure (includes the voltage regulator):
 - 0/16 way 3.86 GHz, or 0/8 way 4.14 GHz (TurboCore) processor card (#4982)
- ▶ POWER7 DDR3 Memory DIMMs (16 DIMM slots per processor card):
 - 0/32 GB (4 X 8 GB), 1066 MHz (#5600)
 - 0/64 GB (4 X 16 GB), 1066 MHz (#5601)
 - 0/128 GB (4 X 32 GB), 1066 MHz (#5602)
- ▶ Six hot-swappable, 2.5-inch, small form factor, SAS disk or SSD bays per enclosure
- ▶ One hot-plug, slim-line, SATA media bay per enclosure (optional)
- ▶ Redundant hot-swap AC power supplies in each enclosure
- ▶ Choice of integrated (HEA) I/O options (one per enclosure):
 - Quad 1 Gb Ethernet
 - Dual 10 Gb Optical and Dual 1 Gb Ethernet
 - Dual 10 Gb Copper and Dual 1 Gb Ethernet
- ▶ One serial port, three USB ports per enclosure (maximum nine per system)
- ▶ Two HMC ports per enclosure (maximum four per system)
- ▶ Eight I/O expansion slots per enclosure (maximum 32 per system):
 - Six PCIe 8x slots plus two GX++ slots per enclosure
- ▶ Dynamic LPAR support, Processor and Memory CUoD
- ▶ PowerVM (optional):
 - Micro-Partitioning
 - Virtual I/O Server (VIOS)
 - Automated CPU and memory reconfiguration support for dedicated and shared processor logical partition (LPAR) groups
 - Support for manual provisioning of resources partition migration (PowerVM Enterprise Edition)

- ▶ Optional PowerHA for AIX, IBM i, and Linux
- ▶ 12X I/O drawer with PCI slots:
 - Up to 16 PCIe I/O drawers (#5802 or #5877)
 - Up to 32 PCI-X DDR I/O drawers (7314-G30 or feature #5796)
- ▶ Disk-only I/O drawers:
 - Up to 110 EXP12S SAS DASD/SSD I/O drawers on SAS PCI controllers (#5886)
 - Up to 60 EXP24 SCSI DASD Expansion drawers on SCSI PCI controllers (7031-D24)
- ▶ IBM Systems Director Active Energy Manager

The Power 780 operator interface/controls located on the front panel of the primary I/O drawer consist of a power ON/OFF button with a POWER indicator, LCD display for diagnostic feedback, a RESET button, and a disturbance or system attention LED.

1.4.3 Minimum features

Each system has a minimum feature-set in order to be valid. The minimum system configuration for a Power 770 is shown in Table 1-3.

Table 1-3 Minimum features for Power 770 system

Power 770 minimum features	Additional notes
1x CEC enclosure (4U)	<ul style="list-style-type: none"> ▶ 1x System Enclosure with IBM Bezel (#5659) or OEM Bezel (#5669) ▶ 1x Service Processor (#5664) ▶ 1x DASD Backplane (#5652) ▶ 2x Power Cords (two selected by customer) ▶ 2x A/C Power Supply (#5632) ▶ 1x Operator Panel (#1853) ▶ 1x HEA Adapter (one of these): <ul style="list-style-type: none"> – Quad 4x 1 Gb (#1803) – Quad 2x 1 Gb and 2 x 10 Gb Optical (#1804) – Quad 2x 1 Gb and 2 x 10 Gb Copper (#1813)
1x primary operating system (one of these)	<ul style="list-style-type: none"> ▶ AIX (#2146) ▶ Linux (#2147) ▶ IBM i (#2145), IBM i 6.1.1 (#0566), or IBM i 7.1.0 (#0567)
1x Processor Card	<ul style="list-style-type: none"> ▶ 3.5 GHz, 12-Core POWER7 Processor Card, 0-core active (#4980) ▶ 3.1 GHz, 16-Core POWER7 Processor Card, 0-core active (#4981)
4x Processor Activations (quantity of four for one of these)	<ul style="list-style-type: none"> ▶ One Processor Activation for Processor Feature #4980 (#5459) ▶ One Processor Activation for Processor Feature #4981 (#5468)
1x DDR3 Memory DIMMs	0/32 GB (4 x 8 GB), 1066 MHz, (#5600 or larger)
16x Activations of 1 GB DDR3 - POWER7 Memory (#8212)	-
For AIX and Linux: 1x disk drive For IBM i: 2x disk drives	Formatted to match the system Primary O/S indicator selected, or if using a Fibre Channel attached SAN (indicated by #0837) a disk drive is not required.

Power 770 minimum features	Additional notes
1X Language Group (selected by the customer)	-
1x Removable Media Device (#5762)	Optionally orderable, a standalone system (not network-attached) would required this feature.
1x HMC	Required for every Power 770 (9117-MMB); however, a communal HMC is acceptable.
Note: A minimum number of four processor activations must be ordered per system. The minimum number of memory activations must enable at least 50% of the ordered memory.	

The minimum system configuration for a Power 780 system is shown in Table 1-4.

Table 1-4 Minimum features for Power 780 system

Power 780 minimum features	Additional notes
1x CEC enclosure (4U)	<ul style="list-style-type: none"> ▶ 1x System Enclosure with IBM Bezel (#5597) or OEM Bezel (#5598) ▶ 1X Service Processor (#5664) ▶ 1X DASD Backplane (#5652) ▶ 2X Power Cords (two selected by customer) ▶ 2X A/C Power Supply (#5632) ▶ 1X Operator Panel (#1853) ▶ 1X HEA Adapters (one of these): <ul style="list-style-type: none"> - Quad 4 x 1 Gb (#1803) - Quad 2 x 1 Gb and 2 x 10 Gb Optical (#1804) - Quad 2 x 1 Gb and 2 x 10 Gb Copper (#1813)
1x primary operating system (one of these)	<ul style="list-style-type: none"> ▶ AIX (#2146) ▶ Linux (#2147) ▶ IBM i (#2145), IBM i 6.1.1 (#0566), or IBM i 7.1.0 (#0567)
1x Processor Card (one of these)	3.86 GHz, 16-Core/4.14 GHz, 8-Core POWER7 processor card, 0-core active (#4982) or 3.1 GHz Proc Card, 0/16 Core POWER7, 16 DDR3 Memory Slots (#4981)
4x Processor Activations for Processor Feature #4982 (#5469)	-
1x DDR3 Memory DIMM	0/32 GB (4 x 8 GB), 1066 MHz, (#5600 or larger)
16x Activations of 1 GB DDR3 - POWER7 Memory (#8212)	-
For AIX and Linux: 1x disk drive For IBM i: 2x disk drives	Formatted to match the system Primary O/S indicator selected, or if using a Fibre Channel attached SAN (indicated by #0837) a disk drive is not required.
1X Language Group (selected by the customer)	-
1x Removable Media Device (#5762)	Optionally orderable, a standalone system (not network-attached) requires this feature.
1x HMC	Required for every Power 780 (9179-MHB); however, a communal HMC is acceptable.
Note: A minimum number of four processor activations must be ordered per system. The minimum number of memory activations must enable at least 50% of the ordered memory.	

1.4.4 Power supply features

Two system AC power supplies (#5632) are required for each CEC enclosure; the second power supply provides redundant power for enhanced system availability. To provide full redundancy, the two power supplies must be connected to separate power distribution units (PDUs).

A CEC enclosure will continue to function with one working power supply. A failed power supply can be hot-swapped but must remain in the system until the replacement power supply is available for exchange. The system requires one functional power supply in each CEC enclosure to remain operational.

Each Power 770 or Power 780 server with two or more CEC enclosures must have one Power Control Cable (#6006 or similar) to connect the service interface card in the first enclosure to the service interface card in the second enclosure.

1.4.5 Processor card features

Each of the four system enclosures contains one powerful POWER7 processor card feature, consisting of two single-chip module processors. Each of the POWER7 processors in the server has a 64-bit architecture, includes up to eight cores on a single-chip module, and contains 2 MB of L2 cache (256 KB per core) and 32 MB of L3 cache (4 MB per core)

Figure 1-3 on page 9 shows the top view of the Power 770 and Power 780 system. The two POWER7 SCMs and the system memory reside on a single processor card feature.

There are two types of Power 770 processor cards, offering the following features:

- ▶ Two 6-core POWER7 SCMs with 24 MB of L3 cache (12-cores per processor card, each core with 4 MB of L3 cache) at 3.5 GHz (#4980)
- ▶ Two 8-core POWER7 SCMs (16-cores per processor card, each core with 4 MB of L3 cache) at 3.1 GHz (#4981)

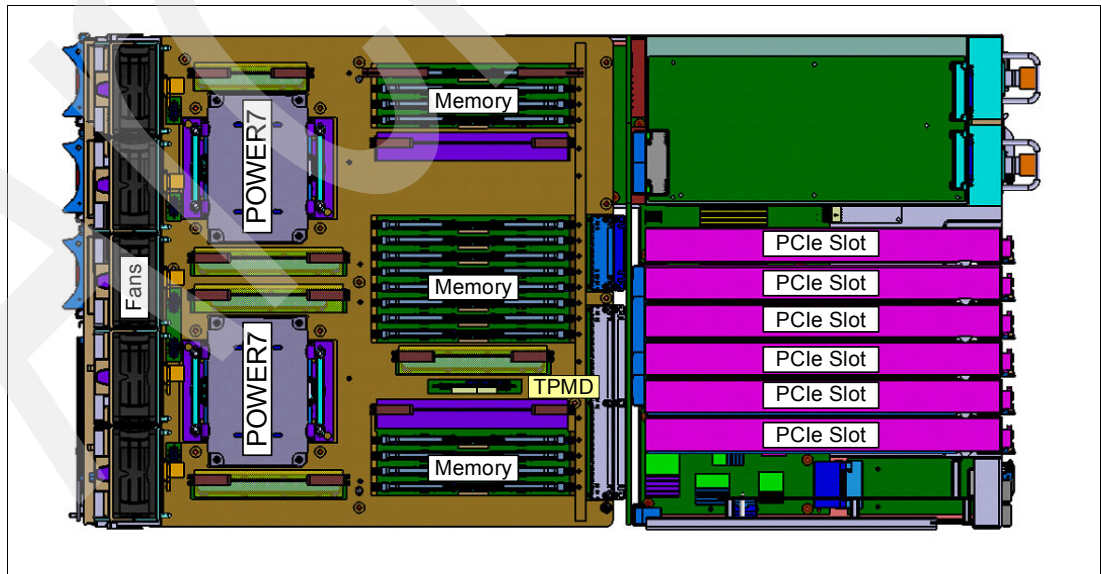


Figure 1-3 Top view of Power 770/Power 780 system

The Power 780 has only one type of processor card, but has two processing modes (MaxCore and TurboCore). The processor card houses the two POWER7 SCMs and the system memory. The Power 780 processor card offers the following feature:

- ▶ Two 8-core POWER7 SCMs with 32 MB of L3 cache (16-cores per processor card, each core with 4 MB of L3 cache) at 3.86 GHz (#4982)

Note: TurboCore mode is supported on the Power 780, but is not supported on the Power 770.

In standard or MaxCore mode, the Power 780 system uses all processor cores running at 3.86 GHz and has access to the full 32 MB of L3 cache. In TurboCore mode, only four of the eight processor cores are available but at a higher frequency (4.14 GHz) and these four cores have access to the full 32 MB of L3 cache. Thus, in Turbo-core mode there are fewer cores running at a higher frequency and a higher core-to-L3-cache ratio.

For a more detailed description of MaxCore and TurboCore modes, see 2.1.5, “Flexible POWER7 processor packaging and offerings” on page 33.

1.4.6 Summary of processor features

Table 1-5 summarizes the processor feature codes for the Power 770.

Table 1-5 Summary of processor features for the Power 770

Feature code	Description	OS support
4980	0/12 way, 3.5 GHz processor card 3.5 GHz processor card, 0/12 core POWER7, 16 DDR3 memory slots: twelve core 3.5 GHz POWER7 CUoD processor planar containing two six-core processors. There are 16 DDR3 DIMM slots on the processor planar (8 DIMM slots per processor), which may be used as capacity on demand (CoD) memory without activating the processors. The voltage regulators are included in this feature code.	AIX Linux IBM i
4981	0/16 way, 3.1 GHz processor card 3.1 GHz processor card, 0/16 core POWER7, 16 DDR3 memory slots: sixteen core 3.1 GHz POWER7 CUoD processor planar containing two eight-core processors. There are 16 DDR3 DIMM slots on the processor planar (8 DIMM slots per processor) which may be used as capacity on demand (CoD) memory without activating the processors. The voltage regulators are included in this feature code.	AIX Linux IBM i
5459	One processor activation for processor #4980: each occurrence of this feature permanently activates one processor on Processor Card #4980. One processor activation for processor feature #4980 with inactive processors	AIX Linux IBM i
5468	One processor activation for processor #4981: each occurrence of this feature will permanently activate one processor on Processor Card #4980. One processor activation for processor feature #4982 with inactive processors.	AIX Linux IBM i

Feature code	Description	OS support
7642	Processor CoD utility billing for #4980, 100 processor-minutes: provides payment for temporary use of processor #4980 with supported AIX or Linux operating systems. Each occurrence of this feature pays for 100 minutes of usage. The purchase of this feature occurs after the customer has 100 minutes of use on processors in the shared processor pool that are not permanently active.	AIX Linux
7643	Processor CoD utility billing for #4980, 100 processor-minutes: provides payment for temporary use of processor #4980 with supported IBM i operating system. Each occurrence of this feature pays for 100 minutes of usage. The purchase of this feature occurs after the customer has 100 minutes of use on processors in the shared processor pool that are not permanently active	IBM i
7646	Processor CoD utility billing for #4981, 100 processor-minutes: provides payment for temporary use of processor #4981 with supported AIX or Linux operating systems. Each occurrence of this feature pays for 100 minutes of usage. The purchase of this feature occurs after the customer has 100 minutes of use on processors in the shared processor pool that are not permanently active.	AIX Linux
7647	Processor CoD utility billing for #4981, 100 processor-minutes: provides payment for temporary use of processor #4981 with supported IBM i operating system. Each occurrence of this feature pays for 100 minutes of usage. The purchase of this feature occurs after the customer has 100 minutes of use on processors in the shared processor pool that are not permanently active	IBM i
7644	One processor-day on/off usage billing for #4980: 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 then provided to your sales channel. The sales channel will place an order for a quantity of On/ Off Processor Day Billing features and bill you. One #7644 should be ordered for each billable processor day of #4980 used by a supported AIX or Linux operating system.	AIX Linux
7645	One processor-day on/off billing for #4980: 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 then provided to your sales channel. The sales channel will place an order for a quantity of On/ Off Processor Day Billing features and bill you. One #7645 should be ordered for each billable processor day of #4980 used by a supported IBM i operating system.	IBM i
7648	One processor-day on/off billing for #4981: 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 then provided to your sales channel. The sales channel will place an order for a quantity of On/ Off Processor Day Billing features and bill you. One #7648 should be ordered for each billable processor day of #4981 used by a supported AIX or Linux operating system.	AIX Linux

Feature code	Description	OS support
7649	One processor-day on/off billing for #4981: 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 then provided to your sales channel. The sales channel will place an order for a quantity of On/ Off Processor Day Billing features and bill you. One #7649 should be ordered for each billable processor day of #4981 used by a supported IBM i operating system.	IBM i
7951	On/Off Processor Enablement: this feature can be ordered to enable your server for On/Off Capacity on Demand. After enabled, you can request processors on a temporary basis. You must sign an On/Off Capacity on Demand contract before you order this feature. Note: To renew this feature after the allowed 360 Processor Days have been used, this feature must be removed from the system configuration file and reordered by placing an MES order.	AIX Linux IBM i

Table 1-6 summarizes the processor feature codes for the Power 780.

Table 1-6 Summary of processor features for the Power 780

Feature code	Description	OS support
4982	3.86 GHz /4.14 GHz TurboCore processor card, 0/16 core POWER7, 16 DDR3 Memory Slots. Sixteen 3.86 GHz POWER7 processors (inactive) on 1 card or eight 4.14 GHz POWER7 processors (inactive) on 1 card when using TurboCore mode.	AIX Linux IBM i
5469	One processor activation for processor feature #4982 with inactive processors.	AIX Linux IBM i
7633	Processor CoD utility billing for #4982, 100 processor-minutes: provides payment for temporary use of processor #4982 with supported AIX or Linux operating systems. Each occurrence of this feature pays for 100 minutes of usage. The purchase of this feature occurs after the customer has 100 minutes of use on processors in the shared processor pool that are not permanently active.	AIX Linux
7634	Processor CoD utility billing for #4982, 100 processor-minutes: provides payment for temporary use of processor #4982 with supported IBM i operating system. Each occurrence of this feature pays for 100 minutes of usage. The purchase of this feature occurs after the customer has 100 minutes of use on processors in the shared processor pool that are not permanently active	IBM i
7635	1 Processor-day on/off billing for #4982: 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 then provided to your sales channel. The sales channel will place an order for a quantity of On/ Off Processor Day Billing features and bill you. One #7635 should be ordered for each billable processor day of #4982 used by a supported AIX or Linux operating system.	AIX Linux

Feature code	Description	OS support
7636	1 Processor-day on/off billing for #4982: 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 then provided to your sales channel. The sales channel places an order for a quantity of On/ Off Processor Day Billing features and bill you. One #7636 should be ordered for each billable processor day of #4982 used by a supported IBM i operating system.	IBM i
7951	<p>On/Off Processor Enablement: this feature can be ordered to enable your server for On/Off Capacity on Demand. After it is enabled, you can request processors on a temporary basis. You must sign an On/Off Capacity on Demand contract before you order this feature.</p> <p>Note: To renew this feature after the allowed 360 Processor Days have been used, this feature must be removed from the system configuration file and reordered by placing an MES order.</p>	AIX Linux IBM i
9982	<p>TurboCore mode specify code: ordering this feature instructs IBM to set up the new server in TurboCore mode during installation.</p> <p>Feature #9982 must be ordered with processor feature #4982 to have the TurboCore mode set-up on the system by the CE during initial installation. Although you may switch into and out of TurboCore mode, switching requires a system reboot. After the system is installed, feature #9982 is not required to switch into and out of TurboCore mode.</p>	AIX Linux IBM i

1.4.7 Memory features

In POWER7 systems, DDR3 memory is used throughout. The POWER7 DDR3 memory uses a new memory architecture to provide greater bandwidth and capacity. This enables operating at a higher data rate for large memory configurations. Each POWER7 processor can support up to eight DDR3 DIMMs running at speeds up to 1066 MHz.

Figure 1-4 on page 14 outlines the general connectivity of the POWER7 processor and DDR3 memory DIMMS. The eight memory channels (four per memory controller) can be clearly seen.

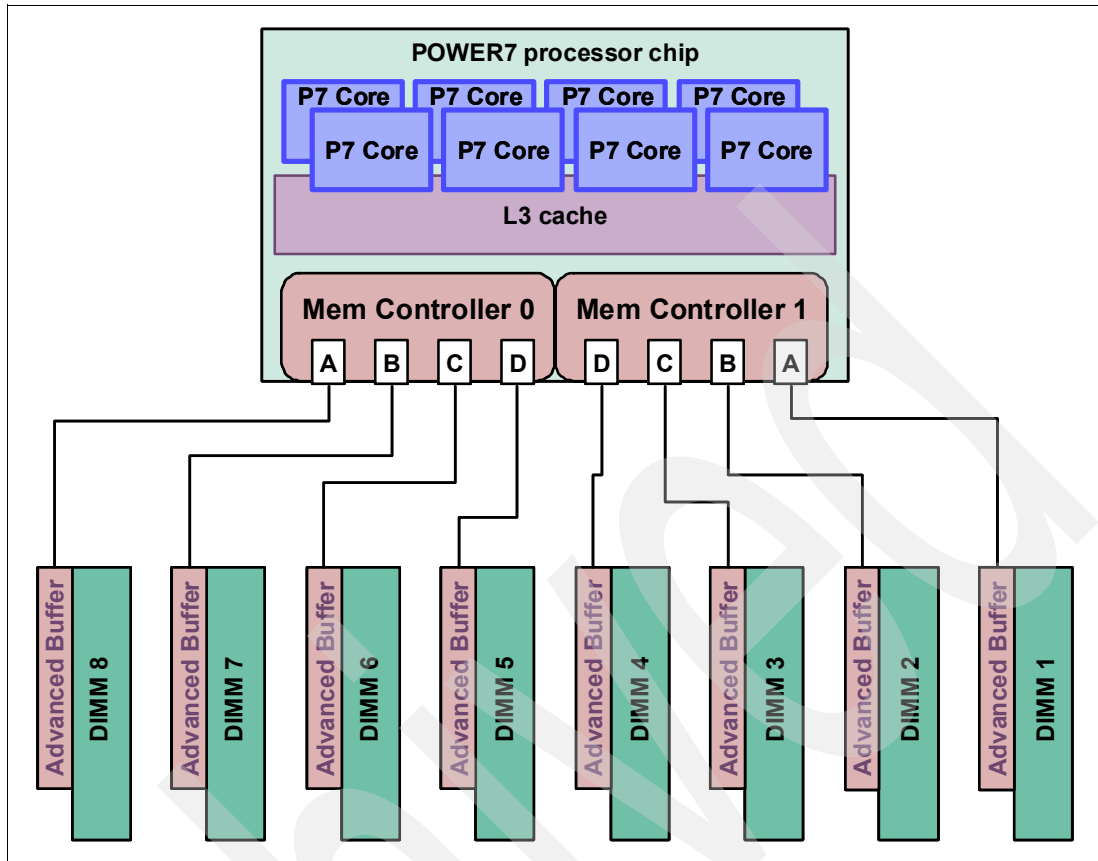


Figure 1-4 Outline of POWER7 memory connectivity to DDR3 DIMMs

On each processor card for the Power 770 and Power 780, there are two POWER7 SCMs, which allow a total of 16 DDR3 memory DIMM slots to be connected (eight DIMM slots per processor).

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

Each DIMM contains DDR3 x8 DRAMs in a configuration, with 10 DRAMs per rank, and plugs into a 276-pin DIMM slot connector.

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

The Power 770 and Power 780 have memory features in 32 GB, 64 GB, and 128 GB capacities. Table 1-7 on page 15 summarizes the capacities of the memory features and highlights other characteristics.

Table 1-7 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

None of the memory in these features is active. Feature number #8212 or #8213 must be purchased to activate the memory. Table 1-8 outlines the memory activation feature codes and corresponding memory capacity activations.

Table 1-8 CoD system memory activation features

Feature code	Activation capacity	Additional information	OS support
8212	1 GB	Activation of 1 GB of DDR3 POWER7 memory. Each occurrence of this feature permanently activates 1 GB of DDR3 - POWER7 memory	AIX Linux IBM i
8213	100 GB	Activation of 100 GB of DDR3 POWER7 memory. Each occurrence of this feature permanently activate 100 GB of DDR3 - POWER7 memory	AIX Linux IBM i
7954	N/A	On/Off Memory Enablement: This feature 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. Clients must sign an On/Off Capacity on Demand contract before this feature is ordered. To renew this feature after the allowed 999 GB Days have been used, this feature must be removed from the system configuration file and reordered by placing an MES order.	AIX Linux IBM i
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. One #7377 feature should be ordered 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.	AIX Linux IBM i
Note: All POWER7 memory features must be purchased with sufficient permanent memory activation features so that the system memory is at least 50% active			

Note: Memory CoD activations activate memory hardware only for the system serial number for which they are purchased. If memory hardware is moved to another system, the memory might not be functional in that system until arrangements are made to move the memory activations or purchase additional memory activations.

1.5 Disk and media features

Each system building block features two SAS DASD controller with six hot-swappable 2.5-inch Small Form Factor (SFF) disk bays and one hot-plug, slim-line media bay per enclosure. The SFF SAS disk drives and Solid State Drive (SSD) are supported internally. In a full configuration with four connected building blocks, the combined system supports up to 24 disk bays. SAS drives and SSD drives can share the same backplane.

Table 1-9 shows the disk drive feature codes that each bay can contain.

Table 1-9 Disk drive feature code description

Feature code	Description	OS support
1890	69 GB SFF SAS Solid State Drive	AIX, Linux
1909	69 GB SFF SAS Solid State Drive	IBM i
1884	69.7 GB 15K RPM SAS SFF Disk Drive	IBM i
1883	73.4 GB 15K RPM SAS SFF Disk Drive	AIX, Linux
1888	139 GB 15K RPM SFF SAS Disk Drive	IBM i
3677	139.5 GB 15k rpm SAS Disk Drive	IBM i
1886	146 GB 15K RPM SFF SAS Disk Drive	AIX, Linux
3647	146 GB 15K RPM SAS Disk Drive'	AIX, Linux
1882	146.8 GB 10K RPM SAS SFF Disk Drive	AIX, Linux
1775	177 GB SFF-1 SSD w/ eMLC	AIX, Linux
1787	177 GB SFF-1 SSD w/ eMLC	IBM i
1995	177 GB SSD Module with eMLC	AIX, Linux
1996	177 GB SSD Module with eMLC	IBM i
1911	283 GB 10K RPM SFF SAS Disk Drive	IBM i
3678	283.7 GB 15k rpm SAS Disk Drive	IBM i
1885	300 GB 10K RPM SFF SAS Disk Drive	AIX, Linux
3648	300 GB 15K RPM SAS Disk Drive	AIX, Linux
3658	428 GB 15K RPM SAS Disk Drive	IBM i
3649	450 GB 15K RPM SAS Disk Drive	AIX, Linux
1916	571 GB 10k RPM SAS SFF Disk Drive	IBM i
1790	600 GB 10K RPM SAS SFF Disk Drive	AIX, Linux

In a full configuration with four connected building blocks, the combined system supports up to four media devices with Media Enclosure and Backplane #5652.

The #5762 SATA Slimline DVD-RAM Drive is the only media device option.

1.6 I/O drawers

The system has eight I/O expansion slots per enclosure, including two dedicated GX++ slots. If more PCI slots are needed, such as to extend the number of LPARs, up to 32 PCI-DDR 12X Expansion Drawers (#5796), and up to 16 12X I/O Drawer PCIe (#5802 and #5877) can be attached.

1.6.1 PCI-DDR 12X Expansion Drawers (#5796)

The PCI-DDR 12X Expansion Drawer (#5796) is a 4U tall (EIA units) drawer and mounts in a 19-inch rack. Feature #5796 takes up half the width of the 4U (EIA units) rack space. Feature #5796 requires the use of a #7314 drawer mounting enclosure. The 4U vertical enclosure can hold up to two #5796 drawers mounted side by side in the enclosure. A maximum of four #5796 drawers can be placed on the same 12X loop.

The I/O drawer has the following attributes:

- ▶ A 4U (EIA units) rack-mount enclosure (#7314) holding one or two #5796 drawers
- ▶ Six PCI-X DDR slots: 64-bit, 3.3V, 266 MHz (blind-swap)
- ▶ Redundant hot-swappable power and cooling units

1.6.2 12X I/O Drawer PCIe (#5802 and #5877)

The #5802 and #5877 expansion units are 19-inch, rack-mountable, I/O expansion drawers that are designed to be attached to the system using 12X double data rate (DDR) cables. The expansion units can accommodate 10, generation 3 cassettes. These cassettes can be installed and removed without removing the drawer from the rack.

A maximum of two #5802 drawers can be placed on the same 12X loop. Feature #5877 is the same as #5802 except it does not support disk bays. Feature #5877 can be on the same loop as #5802. Feature #5877 cannot be upgraded to #5802.

The I/O drawer has the following attributes:

- ▶ Eighteen SAS hot-swap SFF disk bays (only #5802)
- ▶ Ten PCI Express (PCIe) based I/O adapter slots (blind-swap)
- ▶ Redundant hot-swappable power and cooling units

Note: Mixing #5802 or 5877 and #5796 on the same loop is not supported.

1.6.3 I/O drawers and usable PCI slot

The I/O drawer model types can be intermixed on a single server within the appropriate I/O loop. Depending on the system configuration, the maximum number of I/O drawers that is supported differs.

Table 1-10 summarizes the maximum number of I/O drawers supported and the total number of PCI slots available when expansion consists of a single drawer type.

Table 1-10 Maximum number of I/O drawers supported and total number of PCI slots

System drawers	Max #5796 drawers	Max #5802 and #5877 drawers	Total number of slots			
			#5796		#5802 and #5877	
			PCI-X	PCIe	PCI-X	PCIe
1 drawer	8	4	48	6	0	46
2 drawers	16	8	96	12	0	92
3 drawers	24	12	144	18	0	138
4 drawers	32	16	192	24	0	184

1.7 Comparison between models

The Power 770 offers configuration options where the POWER7 processor has 6-cores at 3.5 GHz, or 8-cores at 3.1 GHz. The POWER7 processor has 4 MB of on-chip L3 cache per core. For the 6-core version, a 24 MB of L3 cache is available; for the 8-core version, 32 MB of L3 cache is available.

The Power 780 system offers only one processor option: an 8-core POWER7 with 32 MB of L3 cache. However, the Power 780 can be booted to run in one of two modes:

- ▶ MaxCore mode: All eight processor cores are active at 3.86 GHz with 32 MB of L3 cache.
- ▶ TurboCore mode: Four of the eight processor cores are active at 4.14 GHz with 32 MB of L3 cache (2x the L3 cache per core).

Table 1-11 summarizes the processor core options and frequencies and matches them to the L3 cache sizes for the Power 770 and Power 780.

Table 1-11 Summary of processor core counts, core frequencies, and L3 cache sizes

System	Cores per POWER7 SCM	Frequency (GHz)	L3 cache ^a	Enclosure summation ^b
Power 770	6	3.5	24 MB	12-cores and 48 MB L3 cache
Power 770	8	3.1	32 MB	16-cores and 64 MB L3 cache
Power 780 in MaxCore mode ^c	8	3.86	32 MB	16-cores and 64 MB L3 cache
Power 780 in TurboCore mode ^d	4 active	4.14	32 MB	8-cores active and 64 MB L3 cache

a. the total L3 cache available on the POWER7 SCM, maintaining 4 MB per processor core

b. the total number of processor cores and L3 cache within a populated enclosure

c. MaxCore mode applies to Power 780 only. Each POWER7 SCM has 8 active cores and 32 MB L3 cache

d. TurboCore mode applies to Power 780 only. Each POWER SCM uses 4 of the 8 cores but at a higher frequency, and 32 MB L3 cache

1.8 Build to Order

You can perform a *Build to Order* (also called *a la carte*) configuration using the IBM Configurator for e-business (e-config), where you specify each configuration feature that you want on the system. You build on top of the base required features, such as the embedded Integrated Virtual Ethernet adapter.

This is the only configuration method for the IBM Power 770 and 780 servers.

1.9 IBM Editions

No IBM Edition offerings are available for the IBM Power 770 and 780 servers.

1.10 Model upgrade

You can upgrade the 9117-MMA with IBM POWER6 or POWER6+™ processors to the IBM Power 770 and 780 with POWER7 processors. For upgrades from POWER6 or POWER6+ processor-based systems, IBM will install new CEC enclosures to replace the enclosures you currently have. Your current CEC enclosures will be returned to IBM in exchange for the financial consideration identified under the applicable feature conversions for each upgrade.

Clients taking advantage of the model upgrade offer from a POWER6 or POWER6+ processor-based system are required to return all components of the serialized MT-model that were not ordered through feature codes. Any feature for which a feature conversion is used to obtain a new part must be returned to IBM also. Clients may keep and reuse any features from the CEC enclosures that were not involved in a feature conversion transaction.

1.10.1 Upgrade considerations

Feature conversions have been set up for:

- ▶ POWER6 and POWER6+ processors to POWER7 processors
- ▶ DDR2 memory DIMMS to DDR3 memory DIMMS
- ▶ Trim kits (a new trim kit is needed when upgrading to a two-, three-, or four-door system)
- ▶ Enterprise enablement

The following features that are present on the current system can be moved to the new system:

- ▶ PCIe adapters with cables
- ▶ Line cords, keyboards, and displays
- ▶ PowerVM (#7942 and #7995)
- ▶ I/O drawers (#5796, #5802, #5877, and #5886)
- ▶ Racks (#0551, #0553, and #0555):
 - Doors (#6068, #6069, #6248, #6249, #6858)
 - Trim kits (#6263 and #6272) for one-drawer configurations only or for racks holding only I/O and no 770 processor enclosures.
- ▶ SATA DVD-RAM (#5762)

The Power 770 and 780 can support the following drawers:

- ▶ #5802 and #5877 PCIe 12X I/O drawers
- ▶ #5796 and 7413-G30 PCI-X (12X) I/O Drawer
- ▶ #7031-D24 TotalStorage® EXP24 SCSI Disk Drawer
- ▶ #5886 EXP12S SAS Disk Drawer

The Power 770 and 780 support only the SAS DASD SFF hard disks internally. The older 3.5-inch DASD hard disks can be attached to Model MMB and MHB, but must be located in a I/O drawer such as #5886.

For POWER6 or POWER6+ processor-based systems that have the On/Off CoD function enabled, you must reorder the On/Off enablement features (#7951 and #7954) when placing the upgrade MES order for the new Power 770 or 780 system to keep the On/Off CoD function active. To initiate the model upgrade, the On/Off enablement features should be removed from the configuration file before the MES order is started. Any temporary use of processors or memory owed to IBM on the existing system must be paid before installing the new Power 770 model MMB or Power 780 model MHB.

Feature 8018 is available to support migration of the PowerVM feature #7942 during the initial order and build of the Upgrade MES MMB or MHB order.

Clients may add feature 8018 to their upgrade orders in a quantity not to exceed the quantity of feature #7942 obtained for the system being upgraded. The feature #7942 should be migrated to the new configuration report in a quantity that equals feature #8018. Additional #7942 features can be ordered during the upgrade.

1.11 Hardware Management Console models

The Hardware Management Console (HMC) is required for managing the IBM Power 770 and 780, and optional for the IBM Power 750 and 755. It provides a set of functions that are necessary to manage the system, including:

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

The IBM Power 770 and 780 are not supported by the Integrated Virtualization Manager (IVM).

Several HMC models are supported to manage POWER7 based systems. Licensed Machine Code Version 7 Revision 710 (#0962) is required to support POWER7 processor technology based servers, in addition to POWER5™, POWER5+™, POWER6, and POWER6+ processor technology-based servers. Two models (7042-C07 and 7042-CR5) are available for ordering at the time of writing, but you can also use one of the withdrawn models listed in Table 1-12.

Table 1-12 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	Available	IBM 7042 Model C07 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	Available	IBM 7042 Model CR5 Rack-Mounted Hardware Management Console

The base Licensed Machine Code Version 7 Revision 710 supports the IBM Power 750 and 755. Additionally, Service Pack 1 is needed to support IBM Power 770 and 780.

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

Upgrade the HMC memory to 4 GB if it will manage more than 254 partitions.

1.12 System racks

The Power 770 and its I/O Drawers are designed to mount in the 7014-T00, 7014-T42, 7014-B42, 7014-S25, #0551, #0553 or #0555 rack. The Power 780 and I/O drawers can be ordered only with the 7014-T00 and 7014-T42 racks. These are built to the 19-inch EIA standard. An existing 7014-T00, 7014-B42, 7014-S25, 7014-T42, #0551, #0553, or #0555 rack can be used for the Power 770 and 780 if sufficient space and power are available.

The 36U (1.8 meter) rack (#0551) and the 42U (2.0 meter) rack (#0553) are available for order on MES upgrade orders only. For initial system orders, the racks should be ordered as machine type 7014, Models T00, B42, S25, or T42.

If a system is to be installed in a rack or cabinet that is not IBM, it must meet requirements.

Note: The client is responsible for ensuring that the installation of the drawer in the preferred rack or cabinet results in a configuration that is stable, serviceable, safe, and compatible with the drawer requirements for power, cooling, cable management, weight, and rail security.

1.12.1 IBM 7014 Model T00 rack

The 1.8 Meter (71-in.) Model T00 is compatible with past and present IBM Power systems. The features of the T00 rack are as follows:

- ▶ Has 36U (EIA units) of usable space.
- ▶ Has optional removable side panels.
- ▶ Has optional highly perforated front door.

- ▶ Has optional side-to-side mounting hardware for joining multiple racks.
- ▶ Has standard business black or optional white color in OEM format.
- ▶ Has increased power distribution and weight capacity.
- ▶ Supports both AC and DC configurations.
- ▶ The rack height is increased to 1926 mm (75.8 in.) if a power distribution panel is fixed to the top of the rack.
- ▶ Up to four power distribution units (PDUs) can be mounted in the PDU bays (see Figure 1-5 on page 24), but others can fit inside the rack. See 1.12.7, “The AC power distribution unit and rack content” on page 23.
- ▶ Weights are:
 - T00 base empty rack: 244 kg (535 lb)
 - T00 full rack: 816 kg (1795 lb)

1.12.2 IBM 7014 Model T42 rack

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

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

Note: A special door (#6250) is available to make the rack appear as a high-end server (but in a 19-inch rack format instead of a 24-inch rack).

1.12.3 IBM 7014 Model S25 rack

The 1.3 Meter (49-inch) Model S25 rack has the following features:

- ▶ 25U (EIA units)
- ▶ Weights:
 - Base empty rack: 100.2 kg (221 lb)
 - Maximum load limit: 567.5 kg (1250 lb)

The S25 racks do not have vertical mounting space that accommodate feature number 7188 PDUs. All PDUs required for application in these racks must be installed horizontally in the rear of the rack. Each horizontally mounted PDU occupies 1U of space in the rack, and therefore reduces the space available for mounting servers and other components.

Note: The 780 cannot be ordered with a S25 or B25 rack.

1.12.4 Feature number 0555 rack

The 1.3 Meter Rack (#0555) is a 25U (EIA units) rack. The rack that is delivered as #0555 is the same rack delivered when you order the 7014-S25 rack. The included features might differ. The #0555 is supported, but no longer orderable.

1.12.5 Feature number 0551 rack

The 1.8 Meter Rack (#0551) is a 36U (EIA units) rack. The rack that is delivered as #0551 is the same rack delivered when you order the 7014-T00 rack. The included features might differ. Several features that are delivered as part of the 7014-T00 must be ordered separately with the #0551.

1.12.6 Feature number 0553 rack

The 2.0 Meter Rack (#0553) is a 42U (EIA units) rack. The rack that is delivered as #0553 is the same rack delivered when you order the 7014-T42 or B42 rack. The included features might differ. Several features that are delivered as part of the 7014-T42 or B42 must be ordered separately with the #0553.

1.12.7 The AC power distribution unit and rack content

For rack models T00 and T42, 12-outlet PDUs are available. These include PDUs Universal UTG0247 Connector (#9188 and #7188) and Intelligent PDU+ Universal UTG0247 Connector (#7109).

Four PDUs can be mounted vertically in the back of the T00 and T42 racks. See Figure 1-5 for the placement of the four vertically mounted PDUs. In the rear of the rack, two additional PDUs can be installed horizontally in the T00 rack and three in the T42 rack. The four vertical mounting locations will be filled first in the T00 and T42 racks. Mounting PDUs horizontally consumes 1 U per PDU and reduces the space available for other racked components. When mounting PDUs horizontally, use fillers in the EIA units occupied by these PDUs to facilitate proper air-flow and ventilation in the rack.

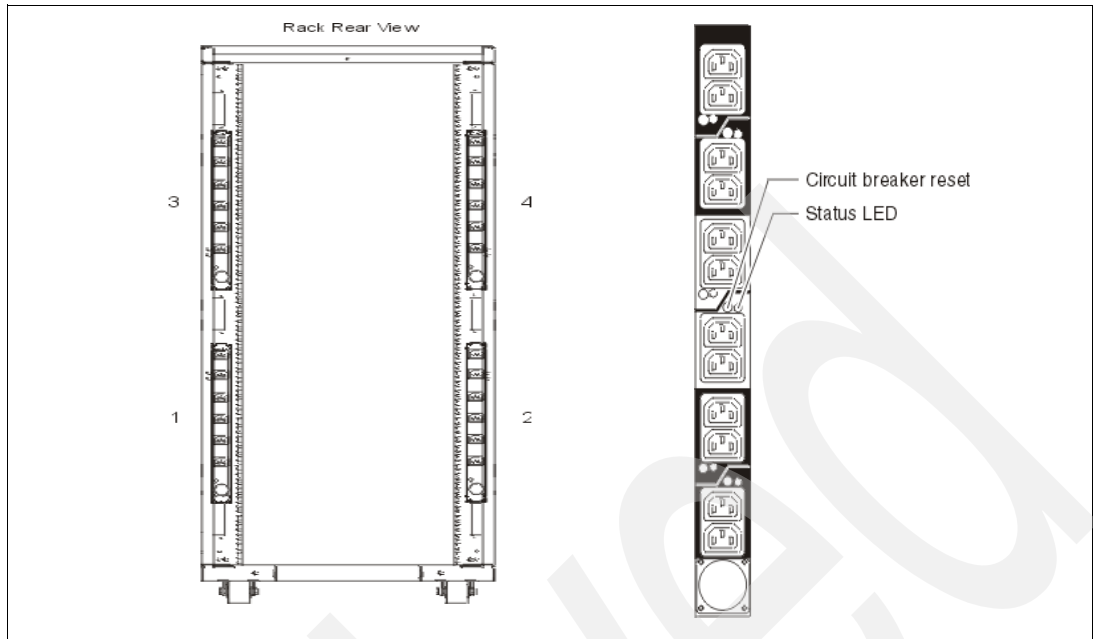


Figure 1-5 PDU placement and PDU view

For the Power 770 and 780 installed in IBM 7014 or #055x racks, the following PDU rules apply:

- ▶ For PDU #7188 and #7109 when using power cord #6654, #6655, #6656, #6657, or #6658: Each pair of PDUs can power up to three Power 770 and 780 CEC enclosures.
- ▶ For PDU #7188 and #7109 when using power cord #6489, #6491, #6492, or #6653: Each pair of PDUs can power up to seven Power 770 and 780 CEC enclosures.

For detailed power cord requirements and power cord feature codes, see the IBM Power Systems Hardware Information Center Web site:

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

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

The Base/Side Mount Universal PDU (#9188) and the optional, additional, Universal PDU (#7188) and the Intelligent PDU+ options (#7109) support a wide range of country requirements and electrical power specifications. The PDU receives power through a UTG0247 power line connector. Each PDU requires one PDU-to-wall power cord. Various power cord features are available for different countries and applications by varying the PDU-to-wall power cord, which must be ordered separately. Each power cord provides the unique design characteristics for the specific power requirements. To match new power requirements and save previous investments, these power cords can be requested with an initial order of the rack or with a later upgrade of the rack features.

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

The Universal PDUs are compatible with previous models.

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

Each system drawer to be mounted in the rack requires two power cords, which are not included in the base order. For maximum availability, be sure to connect power cords from the same system to two separate PDUs in the rack, and to connect each PDU to independent power sources.

1.12.8 Rack-mounting rules

The system consists of one to four CEC enclosures. Each enclosure occupies 4U of vertical rack space. The primary considerations that should be accounted for when mounting the system into a rack are:

- ▶ For configurations with two, three, or four drawers, all drawers must be installed together in the same rack, in a contiguous space of 8 U, 12 U, or 16 U within the rack. The uppermost enclosure in the system is the base enclosure. This enclosure will contain the active service processor and the operator panel. If a second CEC enclosure is part of the system, the backup service processor is contained in the second CEC enclosure.
- ▶ The 7014-T42, -B42, or #0553 rack is constructed with a small flange at the bottom of EIA location 37. When a system is installed near the top of a 7014-T42, -B42, or #0553 rack, no system drawer can be installed in EIA positions 34, 35, or 36. This approach is to avoid interference with the front bezel or with the front flex cable, depending on the system configuration. A two-drawer system cannot be installed above position 29. A three-drawer system cannot be installed above position 25. A four-drawer system cannot be installed above position 21. (The position number refers to the bottom of the lowest drawer.)
- ▶ When a system is installed in an 7014-T00, -T42, -B42, #0551, or #0553 rack that has no front door, a Thin Profile Front Trim Kit must be ordered for the rack. The required trim kit for the 7014-T00 or #0551 rack is #6263. The required trim kit for the 7014-T42, -B42, or #0553 rack is #6272. When upgrading from a 9117-MMA, trim kits #6263 or #6272 may be used for one drawer enclosures only.
- ▶ The design of the 770 and 780 is optimized for use in a 7014-T00, -T42, -B42, -S25, #0551, or #0553 rack. Both the front cover and the processor flex cables occupy space on the front left side of an IBM 7014, #0551 and #0553 rack that may not be available in typical non-IBM racks.
- ▶ Acoustic door features are available with the 7014-T00, 7014-B42, 7014-T42, #0551 and #0553 racks to meet the lower acoustic levels identified in the specification section of this document. The Acoustic Door feature can be ordered on new T00, B42, T42, #0551 and #0553 racks or ordered for the T00, B42, T42, #0551 and #0553 racks that you already own.

1.12.9 Useful rack additions

This section highlights several available solutions for IBM Power Systems rack-based systems.

IBM 7214 Model 1U2 SAS Storage Enclosure

The IBM System Storage® 7214 Tape and DVD Enclosure Express is designed to mount in one EIA unit of a standard IBM Power Systems 19-inch rack and can be configured with one or two tape drives, or either one or two Slim DVD-RAM or DVD-ROM drives in the right-side bay.

The two bays of the 7214 Express can accommodate the following tape or DVD drives for IBM Power servers:

- ▶ DAT72 36 GB Tape Drive: up to two drives
- ▶ DAT72 36 GB Tape Drive: up to two drives
- ▶ DAT160 80 GB Tape Drive: up to two drives
- ▶ Half-high LTO Ultrium 4 800 GB Tape Drive: up to two drives
- ▶ DVD-RAM Optical Drive: up to two drives
- ▶ DVD-ROM Optical Drive: up to two drives

Flat panel display options

The IBM 7316 Model TF3 is a rack-mountable flat panel console kit consisting of a 17-inch 337.9 mm x 270.3 mm flat panel color monitor, rack keyboard tray, IBM Travel Keyboard, support for IBM keyboard/video/mouse (KVM) switches, and language support. The IBM 7316-TF3 Flat Panel Console Kit offers:

- ▶ Slim, sleek, lightweight monitor design that occupies only 1U (1.75 inches) in a 19-inch standard rack
- ▶ A 17-inch, flat screen TFT monitor with truly accurate images and virtually no distortion
- ▶ Ability to mount the IBM Travel Keyboard in the 7316-TF3 rack keyboard tray
- ▶ Support for IBM keyboard/video/mouse (KVM) switches that provide control of as many as 128 servers, and support of both USB and PS/2 server-side keyboard and mouse connections

Architecture and technical overview

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

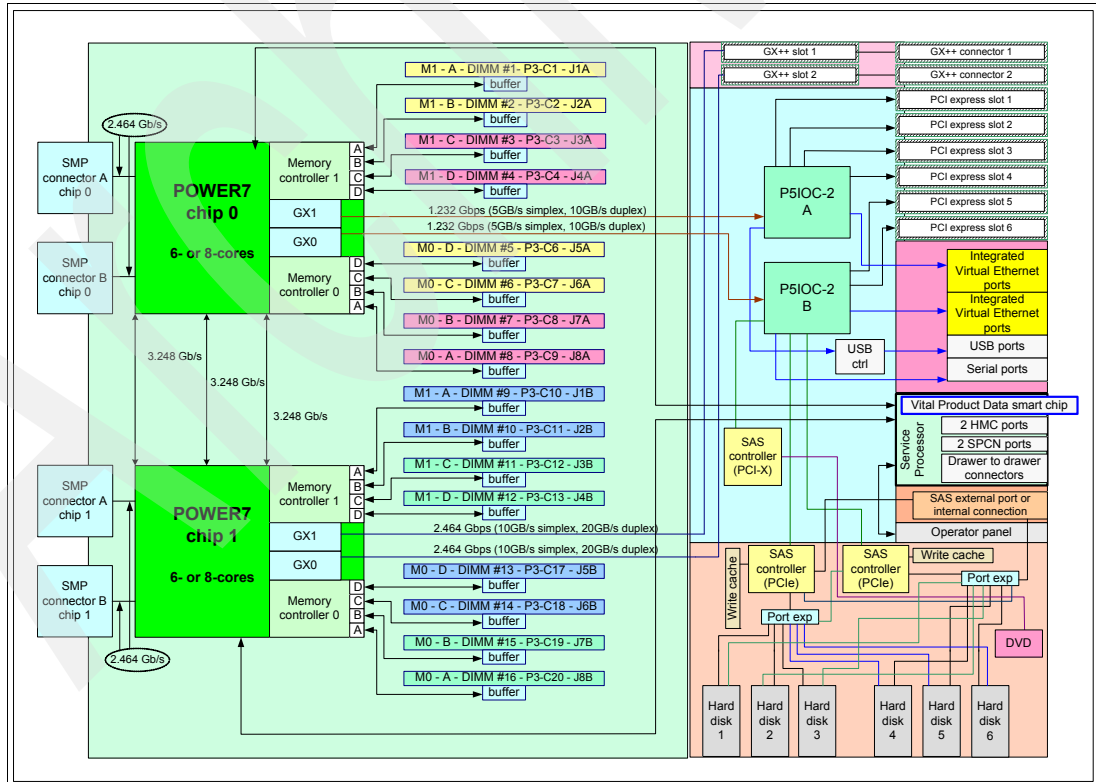


Figure 2-1 Power 770 or Power 780 logic data flow

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

You should always do performance sizing at the application-workload environment level and evaluate performance using real-world performance measurements using production workloads.

Archived

2.1 The IBM POWER7 processor

The IBM POWER7 processor represents a leap forward in technology achievement and associated computing capability. The multi core architecture of the POWER7 processor has been matched with innovation across a wide range of related technologies in order 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 in order 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 in order to achieve required levels of throughput and bandwidth. Areas of innovation for the POWER7 processor and POWER7 processor-based systems include (but are not limited to):

- ▶ 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 signaling
- ▶ 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 POWER6+ processor-based systems.

Figure 2-2 on page 30 shows the POWER7 processor die layout with the major areas identified; processor cores, L2 cache, L3 cache and chip interconnection, simultaneous multiprocessing (SMP) links, and memory controllers.

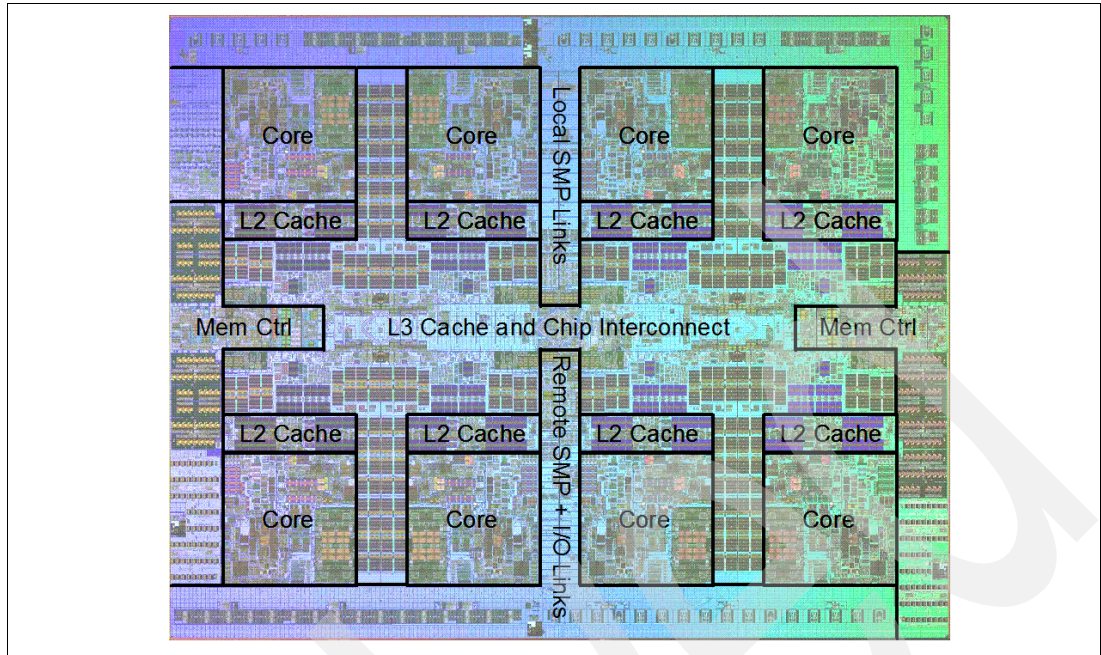


Figure 2-2 POWER7 processor die with key areas indicated

2.1.1 POWER7 processor overview

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

The POWER7 processor chip is 567 mm² and is built using 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 up to 32 MB of shared on-chip L3 cache.

For memory access, the POWER7 processor includes two DDR3 (double data rate 3) 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 very high coherency bandwidth and takes advantage of the IBM dual-scope broadcast coherence protocol.

Table 2-1 summarizes the technology characteristics of the POWER7 processor.

Table 2-1 Summary of POWER7 processor technology

Technology	POWER7 processor
Die size	567 mm ²
Fabrication technology	<ul style="list-style-type: none"> ▶ 45 nm lithography ▶ Copper interconnect ▶ Silicon-on-Insulator ▶ eDRAM
Components	1.2 billion components/transistors offering the equivalent function of 2.7 billion (for further details see 2.1.6, "On-chip L3 cache innovation and Intelligent Cache" on page 34)
Processor cores	8
Max execution threads core/chip	4/32
L2 cache core/chip	256 KB/2 MB
On-chip L3 cache 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

2.1.2 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 as follows:

- ▶ 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 caches that are tightly coupled to each POWER7 processor core are:

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

2.1.3 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 enables the POWER7 processor to maximize the throughput of the processor core by offering an increase in processor-core efficiency. SMT4 mode is the latest step in an evolution of multithreading technologies introduced by IBM. The diagram in Figure 2-3 shows the evolution of simultaneous multithreading in the industry.

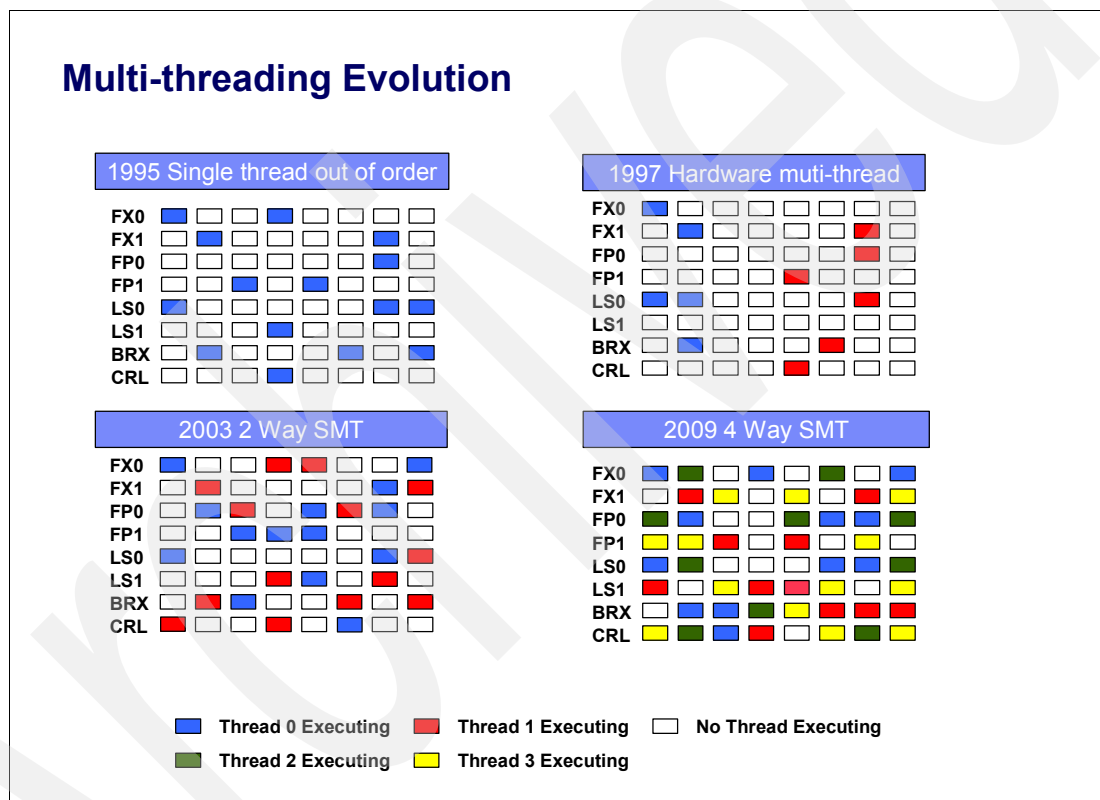


Figure 2-3 Evolution of simultaneous multi-threading

The various SMT modes offered by the POWER7 processor allow flexibility, enabling users to select the threading technology 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 very fast individual tasks can get the performance they need for maximum benefit.

2.1.4 Memory access

Each POWER7 processor chip has two DDR3 memory controllers each with four memory channels (enabling eight memory channels per POWER7 processor). 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 2-4 gives a simple overview of the POWER7 processor memory access structure.

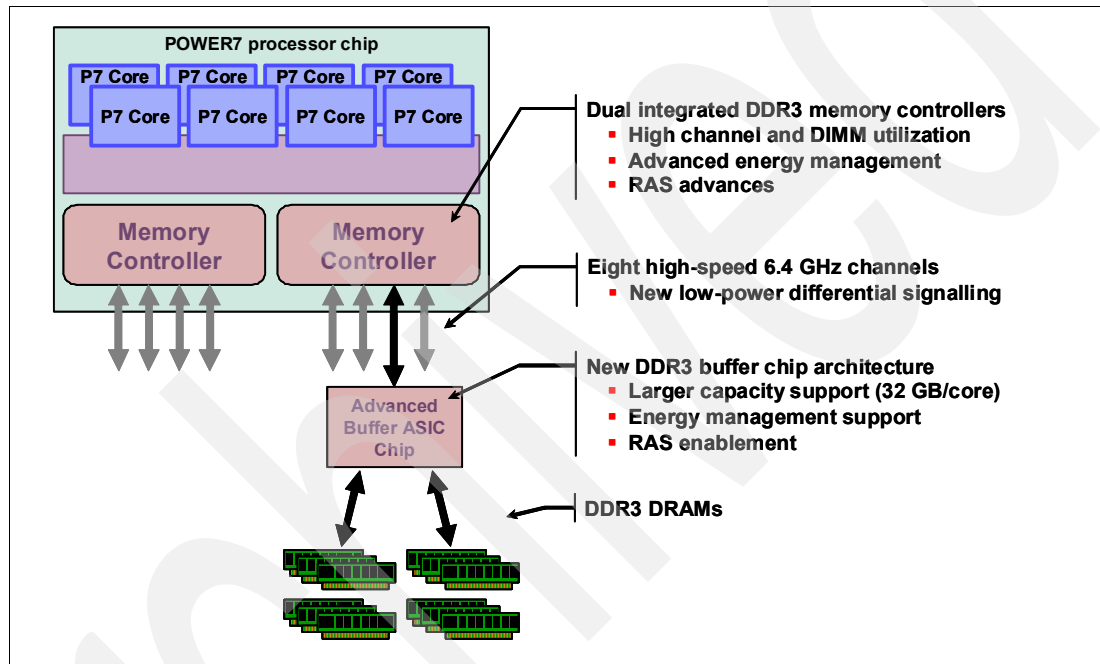


Figure 2-4 Overview of POWER7 memory access structure

2.1.5 Flexible POWER7 processor packaging and offerings

POWER7 processors have the unique ability to optimize to various workload types. For example, database workloads typically benefit from very 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. It uses four cores per POWER7 processor chip with access to the full 32 MB of L3 cache (8 MB per core) and at a faster processor core frequency, which might save on software costs for those applications that are licensed per core.

MaxCore mode

MaxCore mode is for workloads that benefit from a higher number of cores and threads handling multiple tasks simultaneously that 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, as well as the full 8-core version.

In most cases (MaxCore mode), the L3 cache associated with the implementation is dependant on the number of active cores. For a 6-core version, this 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.

Optimized for servers

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 2-5 outlines the physical packaging options that are supported with POWER7 processors.

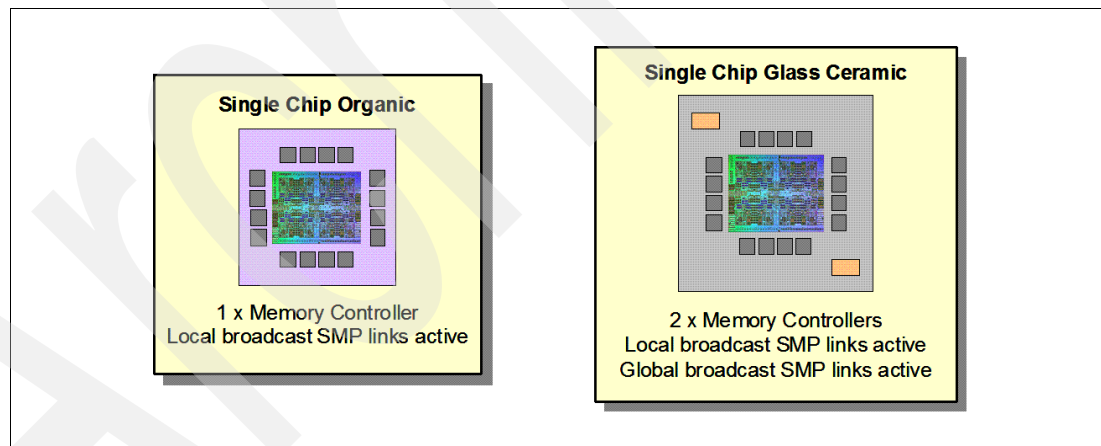


Figure 2-5 Outline of the POWER7 processor physical packaging

2.1.6 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 signaling 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 has 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 2-6 shows the FLR-L3 cache regions for each of the cores on the POWER7 processor die.

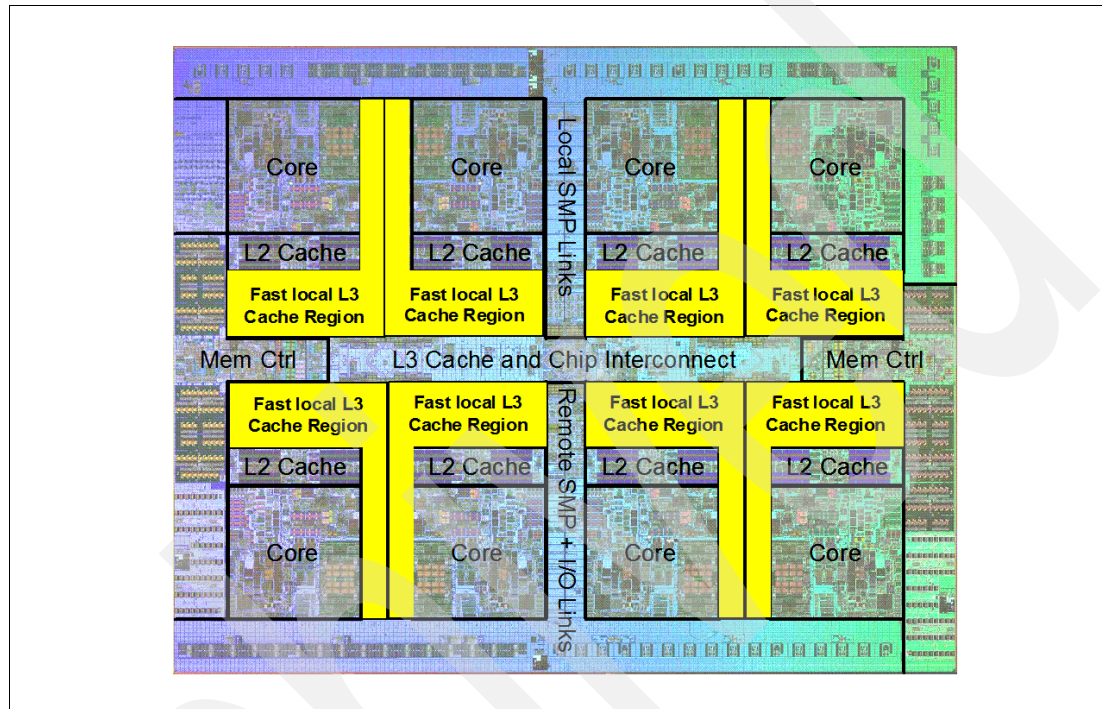


Figure 2-6 Fast local regions of L3 cache on the POWER7 processor

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 compared to L3 accesses on an external (on-ceramic) ASIC.
- ▶ Bandwidth improvement
 - A 2x bandwidth improvement 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 performance of eDRAM when implemented on-chip is similar to conventional SRAM but requires far less physical space. 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.

2.1.7 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 like EnergyScale work with IBM Systems Director Active Energy Manager to dynamically optimize processor speed based on thermal conditions and system utilization.

2.1.8 Comparison of the POWER7 and POWER6 processors

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

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

	POWER7	POWER6+	POWER6
Technology	45 nm	65 nm	65 nm
Die size	567 mm ²	341 mm ²	341 mm ²
Maximum cores	8	2	2
Maximum SMT threads per core	4 threads	2 threads	2 threads
Maximum frequency	4.14 GHz	5.0 GHz	4.7 GHz
L2 Cache	256 KB per core	4 MB per core	4 MB per core
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	32 MB off-chip eDRAM ASIC
Memory support	DDR3	DDR2	DDR2
I/O Bus	Two GX++	One GX++	One GX++
Enhanced Cache Mode (TurboCore)	Yes	No	No
Sleep & Nap Mode	Both	Nap only	Nap only

2.2 POWER7 processor cards

In the Power 770 and Power 780 systems, each enclosure houses a single POWER7 processor card, which hosts two populated POWER7 processor sockets and 16 DDR3 memory DIMM slots.

Figure 2-7 on page 37 shows the single POWER7 processor card for an enclosure, and highlights each of the POWER7 processor sockets and the DDR3 DIMM slots.

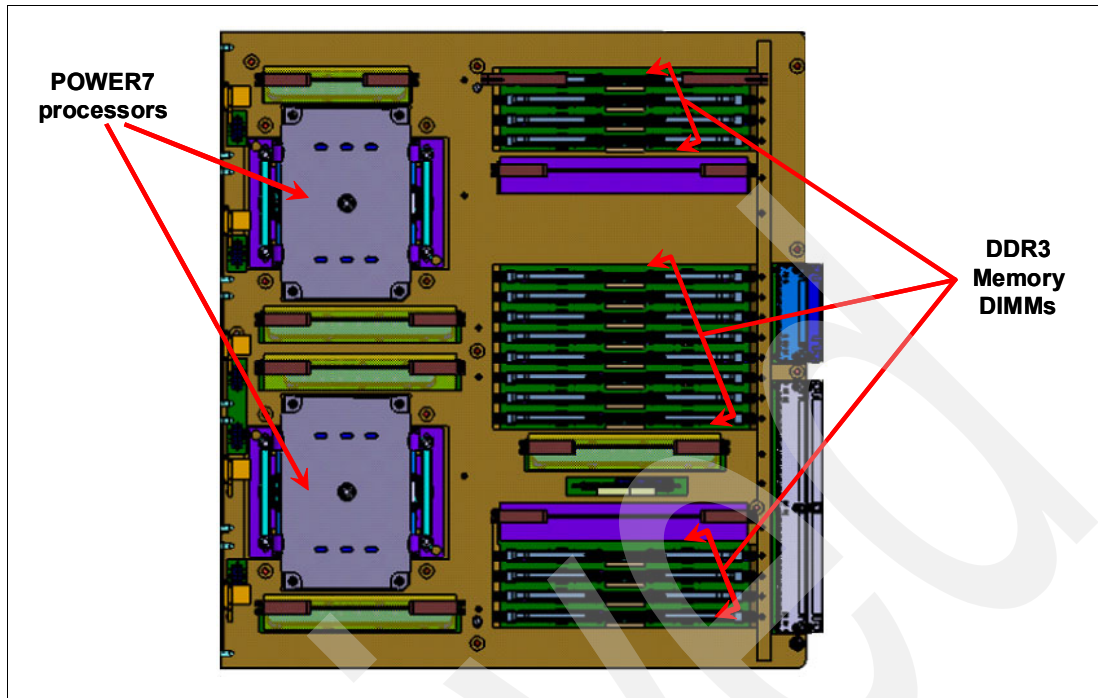


Figure 2-7 POWER7 processor card for an enclosure

Power 770 systems

Power 770 systems support POWER7 processors with various processor-core counts. Table 2-3 summarizes the POWER7 processor options for the Power 770 system.

Table 2-3 Summary of POWER7 processor options for the Power 770 server

Cores per POWER7 processor	Frequency	L3 cache size available per POWER7 processor
6	3.50 GHz	24 MB
8	3.10 GHz	32 MB

With two POWER7 processors in each enclosure, systems can be constructed as follows:

- ▶ Using 6-core POWER7 processors: 12-cores, 24-cores, 36-cores, or 48-cores
- ▶ Using 8-core POWER7 processors: 16-cores, 32-cores, 48-cores, or 64-cores

Power 780 systems

Power 780 systems support POWER7 processors with 8-cores. However, the system can be booted on one of two modes; MaxCore mode or TurboCore mode.

In MaxCore mode, all eight cores of each POWER7 processor is active, runs at 3.86 GHz, and full access to the 32 MB of L3 cache. In TurboCore mode the system uses just four of the POWER7 processor cores, runs at the higher frequency of 4.14 GHz, and has access to the full 32 MB of L3 cache.

Table 2-4 on page 38 summarizes the POWER7 processor and mode options for the Power 780 system.

Table 2-4 Summary of POWER7 processor options and modes for the Power 780 server

Active cores per POWER7 processor	System mode	Frequency	L3 cache size available per POWER7 processor
8	MaxCore	3.86 GHz	32 MB
4	TurboCore	4.14 GHz	32 MB

With two POWER7 processors in each enclosure, systems can be constructed with:

- ▶ MaxCore mode: 16-cores, 32-cores, 48-cores, or 64-cores
- ▶ TurboCore mode: 8-cores, 16-cores, 24-cores, or 32-cores at a higher frequency and larger L3 cache/core ratio

2.3 Memory subsystem

For Power 770 and Power 780 systems, each enclosure houses one POWER7 processor card with two POWER7 single-chip modules (SCMs). Each POWER7 processor has two on-chip DDR3 memory controllers which can interface with a total of 16 DDR3 DIMM cards.

The DIMM cards for the Power 770 and Power 780 are 96 mm tall and placed in one of the 16 DIMM slots on the processor card.

2.3.1 Fully buffered DIMM

Fully buffered DIMM technology is used to increase reliability, speed and density of memory subsystems. Conventionally, data lines from the memory controllers have to be connected to the data lines in every DRAM module. As memory width and access speed increases, the signal decays at the interface of the bus and the device. This effect traditionally degrades either the memory access times or memory density. Fully buffered DIMMs overcome this effect by implementing an advanced buffer between the memory controllers and the DRAMs with two independent signaling interfaces. This technique decouples the DRAMs from the bus and memory controller interfaces, allowing efficient signaling between the buffer and the DRAM.

2.3.2 Memory placement rules

The minimum DDR3 memory capacity for the Power 770 and Power 780 systems is 16 GB of activated memory (32 GB installed).

Note: DDR2 memory (used in POWER6 processor-based systems) is not supported in POWER7 processor-based systems.

Figure 2-8 on page 39 shows the physical memory DIMM topology for the POWER7 processor card.

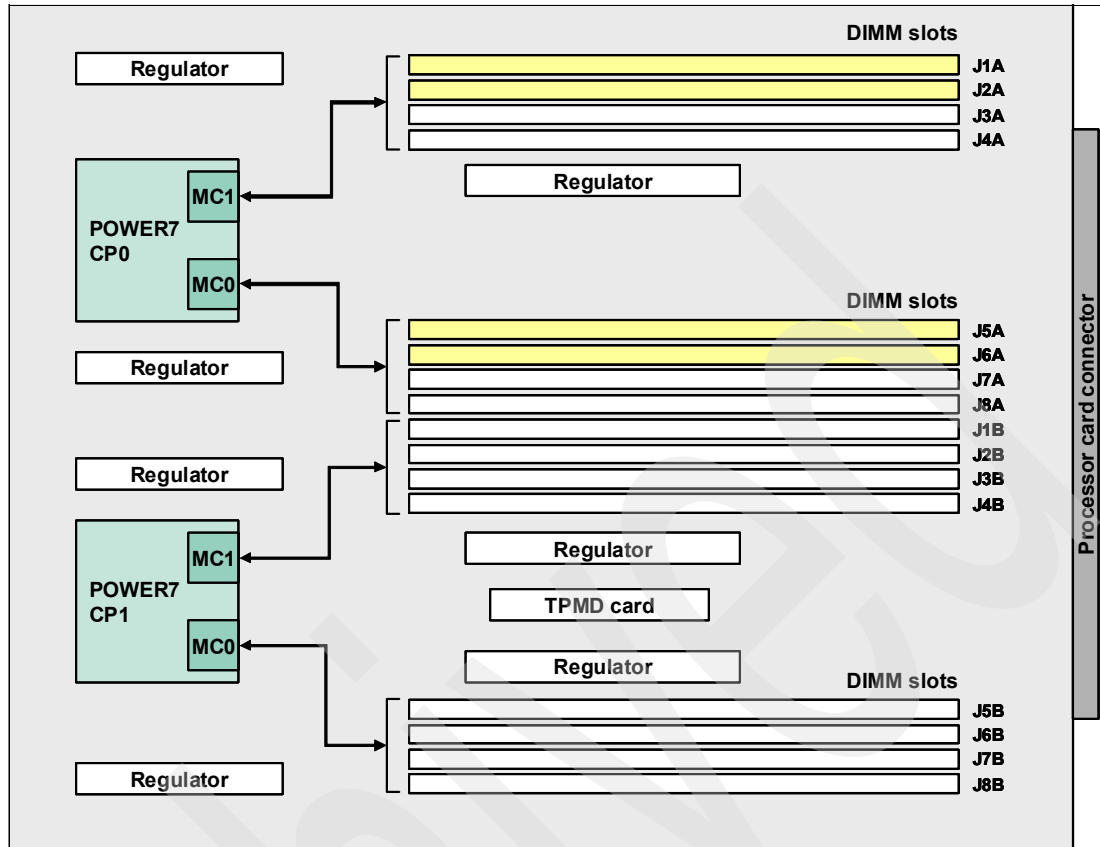


Figure 2-8 Physical memory DIMM topology for the POWER7 processor card

There are 16 buffered DIMM slots: DIMM slots J1A to J8A are connected to the memory controllers on POWER7 processor 0, and DIMM slots J1B to J8B are connected to the memory controllers on POWER7 processor 1.

The memory-plugging rules are as follows:

- ▶ DIMMs must be installed 4x DIMMs at a time, referred to as a *DIMM-quad*.
- ▶ DIMM-quads must be homogeneous (DRAMs of the same capacity).
- ▶ A DIMM-quad is the minimum installable unit.
- ▶ No mixing of memory speeds can occur within a memory controller. For example, 32 GB DIMMs running at 800 MHz cannot be mixed with 8 GB or 16 GB DIMMs running at 1066 MHz.
- ▶ For maximum memory performance, the total memory capacity on each memory controller should be equivalent.
- ▶ The DIMM-quad placement rules for a single enclosure are as follows (see Figure 2-8 for the physical memory topology):
 - Quad 1: J1A, J2A, J5A, J6A (mandatory minimum for each enclosure)
 - Quad 2: J1B, J2B, J5B, J6B
 - Quad 3: J3A, J4A, J7A, J8A
 - Quad 4: J3B, J4B, J7B, J8B

Table 2-5 shows the optimal placement of each DIMM-quad within a single enclosure system. Each enclosure *must have* at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A, as shown with the highlighted color.

Table 2-5 Optimum DIMM-quad placement for a 1x enclosure system

Enclosure 0															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q1	Q1	Q3	Q3	Q1	Q1	Q3	Q3	Q2	Q2	Q4	Q4	Q2	Q2	Q4	Q4
Mandatory: Each enclosure must have at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A. Note: For maximum memory performance, the total memory capacity on each memory controller should be equivalent.															

When populating a multi-enclosure system with DIMM-quads, each enclosure *must have* at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A. After the mandatory requirements and memory-plugging rules are followed, there is an optimal approach to populating the systems.

Table 2-6 shows the optimal placement of each DIMM-quad within a dual-enclosure system. Each enclosure *must have* at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A.

Table 2-6 Optimum DIMM-quad placement for a 2x enclosure system

Enclosure 0															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q1	Q1	Q5	Q5	Q1	Q1	Q5	Q5	Q3	Q3	Q7	Q7	Q3	Q3	Q7	Q7
Enclosure 1															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q2	Q2	Q6	Q6	Q2	Q2	Q6	Q6	Q4	Q4	Q8	Q8	Q4	Q4	Q8	Q8
Mandatory: Each enclosure must have at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A. Note: For maximum memory performance, the total memory capacity on each memory controller should be equivalent.															

Table 2-7 shows the optimal placement of each DIMM-quad within a three-enclosure system. Each enclosure *must have* at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A.

Table 2-7 Optimum DIMM-quad placement for a 3x enclosure system

Enclosure 0															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q1	Q1	Q7	Q7	Q1	Q1	Q7	Q7	Q4	Q4	Q10	Q10	Q4	Q4	Q10	Q10
Enclosure 1															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q2	Q2	Q8	Q8	Q2	Q2	Q8	Q8	Q5	Q5	Q11	Q11	Q5	Q5	Q11	Q11
Enclosure 2															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q3	Q3	Q9	Q9	Q3	Q3	Q9	Q9	Q6	Q6	Q12	Q12	Q6	Q6	Q12	Q12
Mandatory: Each enclosure must have at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A.															
Note: For maximum memory performance, the total memory capacity on each memory controller should be equivalent.															

Table 2-8 shows the optimal placement of each DIMM-quad within a four-enclosure system. Each enclosure must have at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A.

Table 2-8 Optimum DIMM-quad placement for a 4xenclosure system

Enclosure 0															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q1	Q1	Q9	Q9	Q1	Q1	Q9	Q9	Q5	Q5	Q13	Q13	Q5	Q5	Q13	Q13
Enclosure 1															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q2	Q2	Q10	Q10	Q2	Q2	Q10	Q10	Q6	Q6	Q14	Q14	Q6	Q6	Q14	Q14
Enclosure 2															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q3	Q3	Q11	Q11	Q3	Q3	Q11	Q11	Q7	Q7	Q15	Q15	Q7	Q7	Q15	Q15
Enclosure 3															
POWER7 Processor 0								POWER7 Processor 1							
Memory Controller 1				Memory Controller 0				Memory Controller 1				Memory Controller 0			
J1A	J2A	J3A	J4A	J5A	J6A	J7A	J8A	J1B	J2B	J3B	J4B	J5B	J6B	J7B	J8B
Q4	Q4	Q12	Q12	Q4	Q4	Q12	Q12	Q8	Q8	Q16	Q16	Q8	Q8	Q16	Q16
Mandatory: Each enclosure must have at least one DIMM-quad installed in slots J1A, J2A, J5A, and J6A. Note: For maximum memory performance, the total memory capacity on each memory controller should be equivalent.															

2.3.3 Memory throughput

POWER7 has exceptional cache, memory, and interconnect bandwidths. Table 2-9 shows the bandwidth estimate for the Power 770 system running at 3.1 GHz.

Table 2-9 Power 770 memory bandwidth estimates for POWER7 cores running at 3.1 GHz

Memory	Bandwidth
L1 (data) cache	148.8 Gbps
L2 cache	148.8 Gbps
L3 cache	99.2 Gbps
System memory: 4x enclosures:	136.45 Gbps 1091.58 Gbps
Inter-node buses (four enclosures)	158.02 Gbps
Intra-node buses (four enclosures)	415.74 Gbps
Internal GX bus 1 and 2: 4x enclosures:	19.71 Gbps 78.85 Gbps
External GX buss 1 and 2: 4x enclosures:	39.42 Gbps 157.70 Gbps
Total I/O bandwidth (4 enclosures)	236.54 Gbps

With an increase in frequency, the Power 780 running at 3.86 GHz generates higher cache bandwidth, as shown in Table 2-10.

Table 2-10 Power 780 memory bandwidth estimates for POWER7 cores running at 3.86 GHz

Memory	Bandwidth
L1 (data) cache	185.28 Gbps
L2 cache	185.28 Gbps
L3 cache	123.52 Gbps
System memory: 4x enclosures:	136.45 Gbps 1091.58 Gbps
Inter-node buses (4 enclosures)	158.02 Gbps
Intra-node buses (4 enclosures)	415.74 Gbps
Internal GX bus 1 and 2: 4x enclosures:	19.71 Gbps 78.85 Gbps
External GX buss 1 and 2: 4x enclosures:	39.42 Gbps 157.70 Gbps
Total I/O bandwidth (4 enclosures)	236.54 Gbps

2.4 Capacity on Demand

Several types of Capacity on Demand (CoD) are optionally available on the Power 770 and 780 servers 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.

2.4.1 Capacity Upgrade on Demand (CUoD)

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

2.4.2 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 collected monthly. Processors and memory may be activated and turned off an unlimited number of times, when additional processing resources are needed.

This offering provides 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, an enablement feature (MES only) must be ordered and the required contracts must be in place.

If a Power 770 or Power 780 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 on the Power 770 and 780 can be seen in 1.4.6, “Summary of processor features” on page 10 and 1.4.7, “Memory features” on page 13.

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, 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 lets you request up to 999 memory days of temporary capacity. If you have reached 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 requests

When On/Off CoD temporary capacity is needed, simply use 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, visit:

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

2.4.3 Utility Capacity on Demand (Utility CoD)

Utility CoD automatically 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 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, and 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 Web site and payment is based on reported usage. Utility CoD requires PowerVM Standard Edition or PowerVM Enterprise Edition to be active.

If a Power 770 or Power 780 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.

For more information regarding registration, enablement, and use of Utility CoD, visit:

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

2.4.4 Trial Capacity On Demand (Trial CoD)

A *standard request* for Trial CoD requires you to complete a form including contact information and vital product data (VPD) from your Power 770 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 770 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, visit:

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

2.4.5 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* at:

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

2.5 Drawer interconnection cables

IBM Power 770 or 780 systems can be configured with more than one system enclosure. The connection between the processor cards in the separate system enclosures requires a set of processor drawer interconnect cables. Each system enclosure must be connected to each other through a flat flexible SMP cable. These cables are connected on the front of the drawers.

Furthermore, service processor cables are needed to connect the components in each system enclosure to the active service processor for system functions monitoring. These cables connect at the rear of each enclosure and are required for two-, three-, and four-drawer configurations.

The SMP and FSP cable features described in Table 2-11 are required to connect the processors together when system configuration is made of two-, three-, or four-drawer system enclosures.

Table 2-11 Required flex cables feature codes

Enclosure	SMP cables	FSP cables
Two-drawer	3711, 3712	3671
Three-drawer	3712, 3713	3671, 3672
Four-Drawer	3712, 3713, 3714	3671, 3672, 3673

The cables are designed to support hot-addition of system enclosure up to the maximum scalability. When adding a new drawer, existing cables remain in place and new cables are added. The only exception is for cable #3711, which is replaced when growing from a 2-drawer to 3-drawer configuration.

The cables are also designed to allow the concurrent maintenance of the Power 770 or Power 780 in case the IBM service representative needs to extract a system enclosure from the

rack. The design of the flexible cables allows each system enclosure to be disconnected without any impact on the other drawers.

To allow such concurrent maintenance operation, plugging the SMP Flex cables in the order of their numbering is extremely important. Each cable is numbered, as shown in Figure 2-9

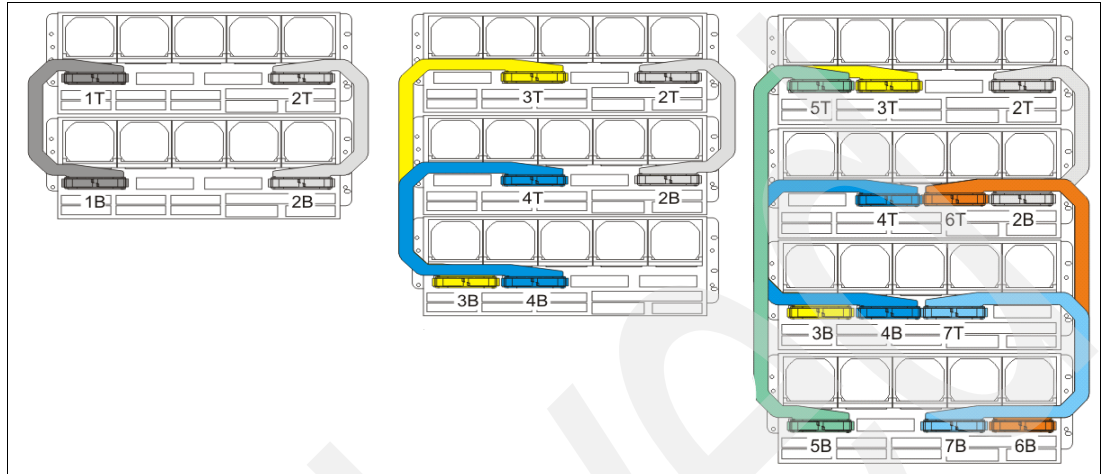


Figure 2-9 SMP Cables installation order

Similarly, the Flexible Service Processor (FSP) flex cables must be installed in the correct order (see Figure 2-10 on page 48), as follows:

1. Install a second node Flex Cable from node 1 to node 2.
2. Add a third node Flex Cable from node 1 and node 2 to node 3.
3. Add a fourth node Flex Cable from node 1 and node 2 to node 4.

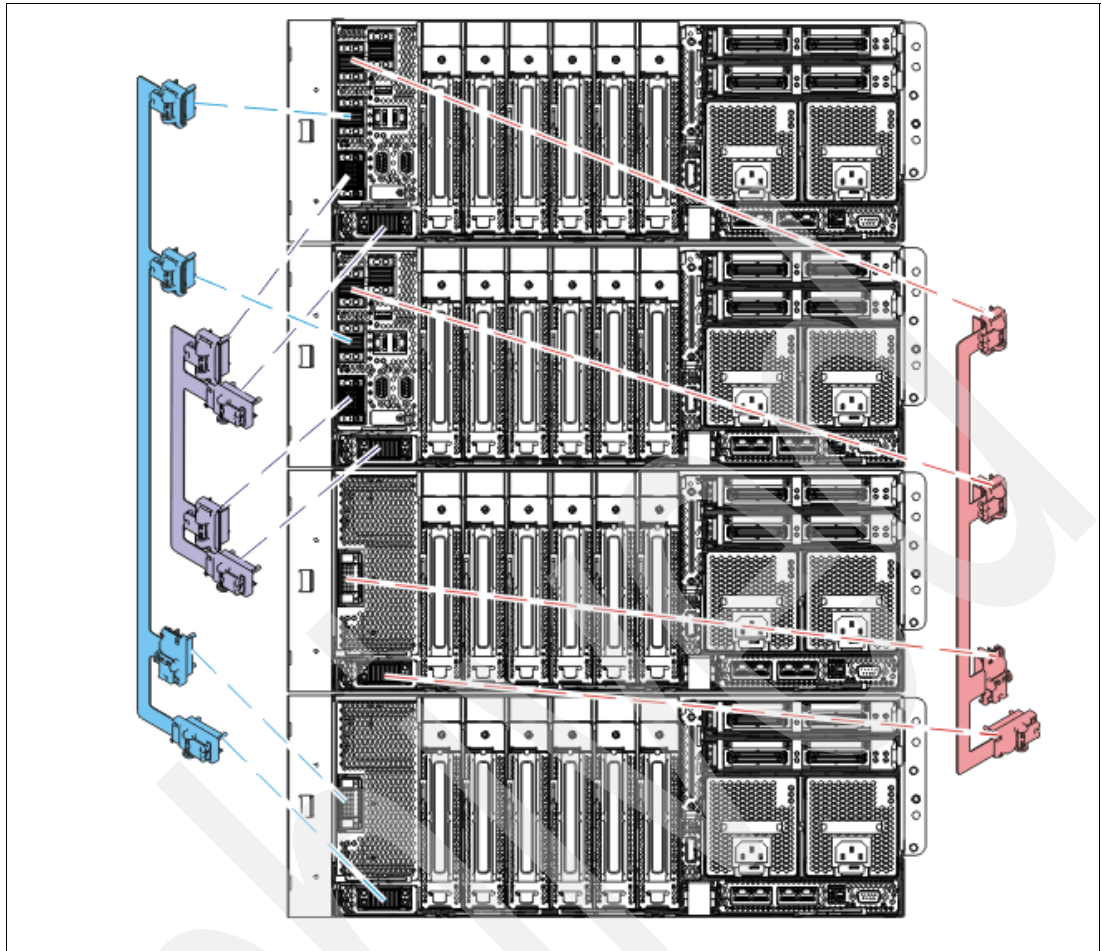


Figure 2-10 FSP flex cables

The design of the Power 770 is optimized for use in an IBM 7014-T00 or 7014-T42 rack. Both the front cover and the external processor fabric cables occupy space on the front left and right side of an IBM 7014 rack; racks that are not from IBM might not offer the same room. When a Power 770 or Power 780 is configured with two or more system enclosures in a 7014-T42 or 7014-B42 rack, the CEC enclosures must be located in EIA 36 or below to allow space for the flex cables.

The total width of the server, with cables installed, is 21 inches, as shown in Figure 2-11.

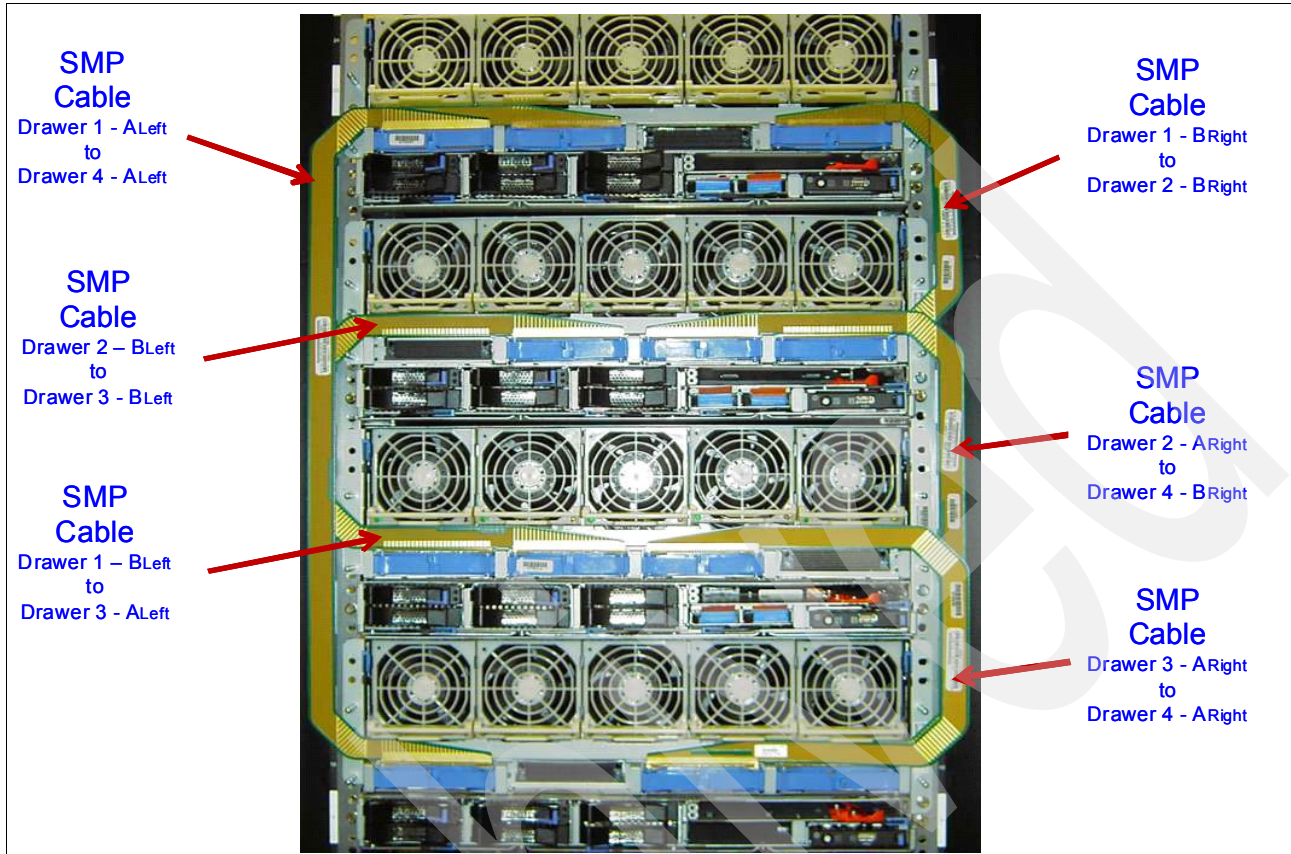


Figure 2-11 Front view of the rack with SMP cables overlapping the rack rails

In the rear of the rack, the FSP cables require only some room in the left side of the racks, as Figure 2-12 shows.

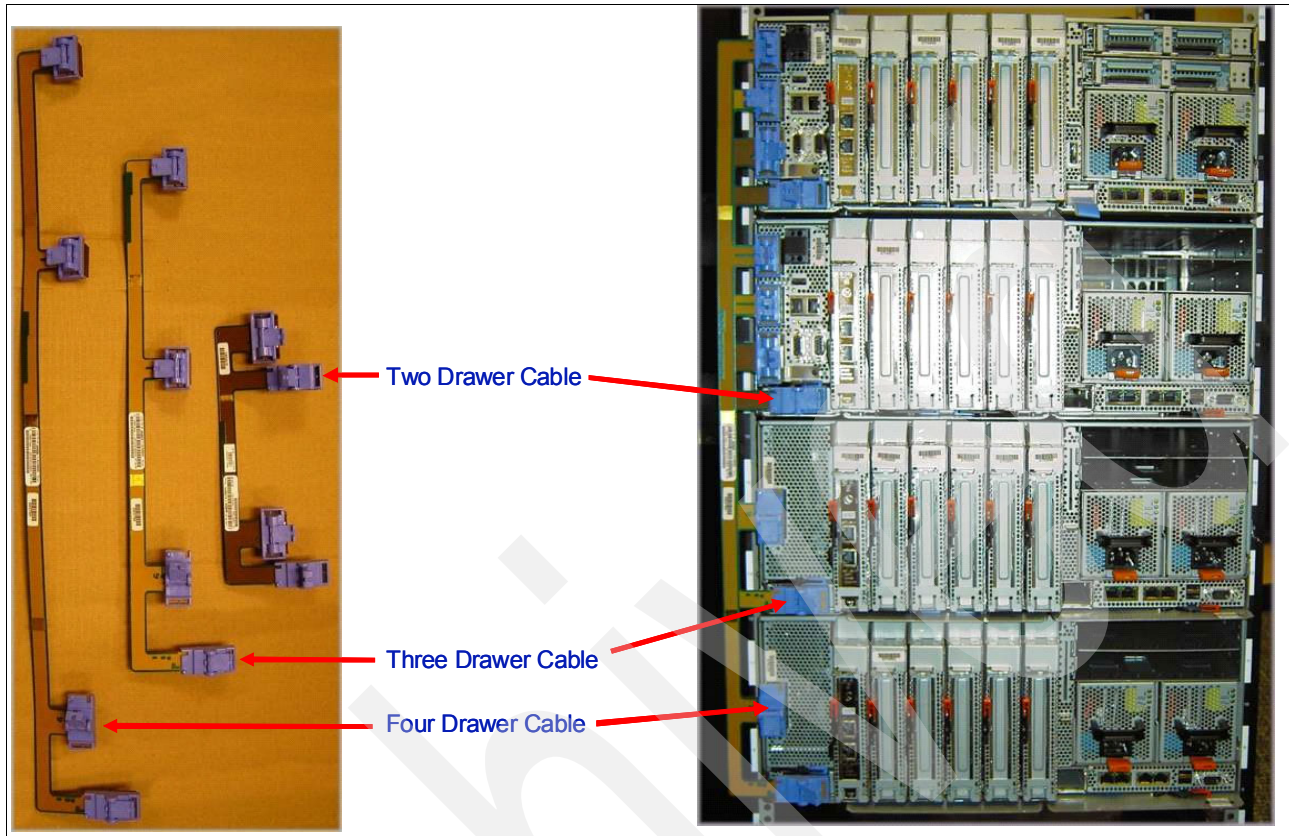


Figure 2-12 Rear view of rack with detail of FSP flex cables

2.6 System bus

This section provides additional information related to the internal buses.

2.6.1 I/O buses and GX++ card

Each system drawer contains one processor card. Each processor card houses two single chip multi-core POWER7 processors.

For I/O connectivity, the POWER7 processor provides two GX++ buses. With two processors per system drawer, a total of four GX++ buses are available for I/O connectivity and expansion. The two GX++ buses off the first processor are routed to the I/O backplane, and drive two of the GX+ multifunctional host bridge chips which provide the following major interfaces:

- ▶ One GX+ pass-through bus: This port is unused.
- ▶ Two 64-bit PCI-X 2.0 buses, one 64-bit PCI-X 1.0 bus, and one 32-bit PCI-X 1.0 bus.
- ▶ Four 8x PCI Express links.
- ▶ Two 10 Gbps Ethernet ports: Each port is individually configurable to function as two 1 Gbps port.

The two remaining GX++ buses from the other processor feed two GX++ Adapter slots. Optional GX++ 12X DDR Adapter, Dual-port (#1808) which is installed in GX++ Adapter slot enables the attachment of a 12X loop which runs at either SDR or DDR speed depending upon the 12X I/O drawers attached. These GX++ Adapter slots are hot-pluggable and do not share the space with any of the PCIe slots.

Table 2-12 shows the configuration of I/O bandwidth of 3.1 GHz processors.

Table 2-12 I/O bandwidth

I/O	Bandwidth
Internal GX Bus 1&2	19.712 GBps
External GX Bus 1&2	39.424 GBps (per node)
Total IO (4 Enclosures)	236.544 GBps (per node)

2.6.2 FSP bus

The Flexible Service Processor (FSP) flex cable, which is located at the rear of the system, is used for FSP communication between the system drawers. FSP card (#5664) is installed in system drawer 1 and system drawer 2, and FSP/Clock Pass-Through card (#5665) is installed in system drawer 3 and system drawer 4 for connecting FSP flex cable. The FSP cable is changed to point to point cabling like processor drawer interconnect cable. When a system drawer is added, another FSP flex cable is added. A detailed cable configuration is discussed in 2.5, “Drawer interconnection cables” on page 46.

2.7 Internal I/O subsystem

The internal I/O subsystem resides on the I/O planar, which supports six PCIe slots. All PCIe slots are hot-pluggable and enabled with enhanced error handling (EEH). In the unlikely event of a problem, EEH-enabled adapters respond to a special data packet that is generated from the affected PCIe or PCI-X slot hardware by calling system firmware, which examines the affected bus, allows the device driver to reset it, and continues without a system reboot.

Table 2-13 lists slot configuration of the Power 770 and 780.

Table 2-13 Slot configuration of the Power 770 and 780

Slot number	Description	Location code	PCI host bridge (PHB)	Max. card size
Slot 1	PCIe x8	P1-C1	P5IOC2 A PCIe PHB0	Full length
Slot 2	PCIe x8	P1-C2	P5IOC2 A PCIe PHB1	Full length
Slot 3	PCIe x8	P1-C3	P5IOC2 A PCIe PHB2	Full length
Slot 4	PCIe x8	P1-C4	P5IOC2 A PCIe PHB3	Full length
Slot 5	PCIe x8	P1-C5	P5IOC2 B PCIe PHB0	Full length
Slot 6	PCIe x8	P1-C6	P5IOC2 B PHB1	Full length
Slot 7	GX++	P1-C2	-	-
Slot 8	GX++	P1-C3	-	-

2.7.1 Blind-swap cassettes

The Power 770 and 780 uses new fourth-generation blind-swap cassettes to manage the installation and removal of adapters. This new mechanism requires an interposer card that allows the PCIe adapters to plug in horizontally into the system, allows more airflow through the cassette, and provides more room under the PCIe cards to accommodate the GX+ multifunctional host bridge chip heat sink height. Cassettes can be installed and removed without removing the drawer from the rack.

2.7.2 System ports

Each system enclosure is equipped with an Integrated Virtual Ethernet adapter (IVE) that has one serial port, called a system port. Since Power 770 / 780 is a HMC managed system, this serial port is always OS-controlled and it is available in any system configuration. It supports any serial device that has an OS device driver. The FSP virtual console will be on the HMC. Also there is no existing uninterruptible power supply connection support.

2.8 Integrated Virtual Ethernet adapter

The POWER7 processor-based servers extend the virtualization technologies introduced in POWER5 by offering the Integrated Virtual Ethernet adapter (IVE). IVE, also named Host Ethernet Adapter (HEA), enables an easy way to manage the sharing of the integrated high-speed Ethernet adapter ports. Three Integrated Virtual Ethernet adapter card options are available, which are listed in 2.8.1, “IVE subsystem and feature codes” on page 53.

IVE includes special hardware features to provide logical Ethernet ports that can communicate to logical partitions (LPAR) reducing the use of POWER Hypervisor™. Its design provides a direct connection for multiple LPARs, allowing LPARs to access external networks through the IVE without having to go through an Ethernet bridge on another logical partition, such as a the Virtual I/O Server. Therefore, this eliminates the need to move packets (using Virtual Ethernet Adapters) between partitions and then through a Shared Ethernet Adapter (SEA) to a physical Ethernet port. LPARs can share IVE ports with improved performance.

Figure 2-13 shows the difference between IVE and SEA implementations.

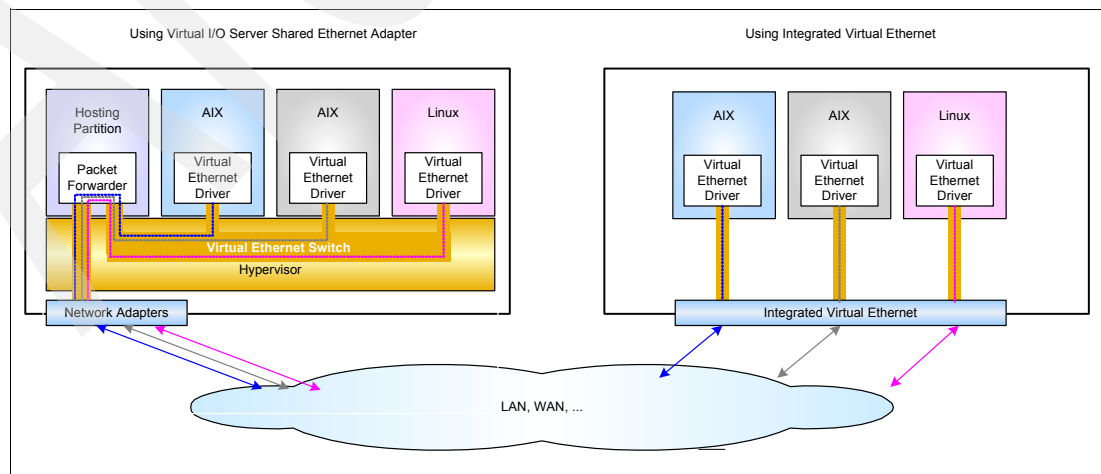


Figure 2-13 Integrated Virtual Ethernet compared to Virtual I/O Server Shared Ethernet Adapter

IVE design meets general market requirements for better performance and better virtualization for Ethernet. It offers:

- ▶ External network connectivity for LPARs using dedicated ports without the need of a Virtual I/O Server
- ▶ Industry standard hardware acceleration, loaded with flexible configuration possibilities
- ▶ The speed and performance of the GX+ bus
- ▶ Great improvement of latency for short packets that are ideal for messaging applications, such as distributed databases that require low latency communication for synchronization and short transactions

For more information about IVE features, see *Integrated Virtual Ethernet Adapter Technical Overview and Introduction*, REDP-4340.

2.8.1 IVE subsystem and feature codes

Figure 2-14 shows a high level-logical diagram of the IVE available in the server IBM Power 770.

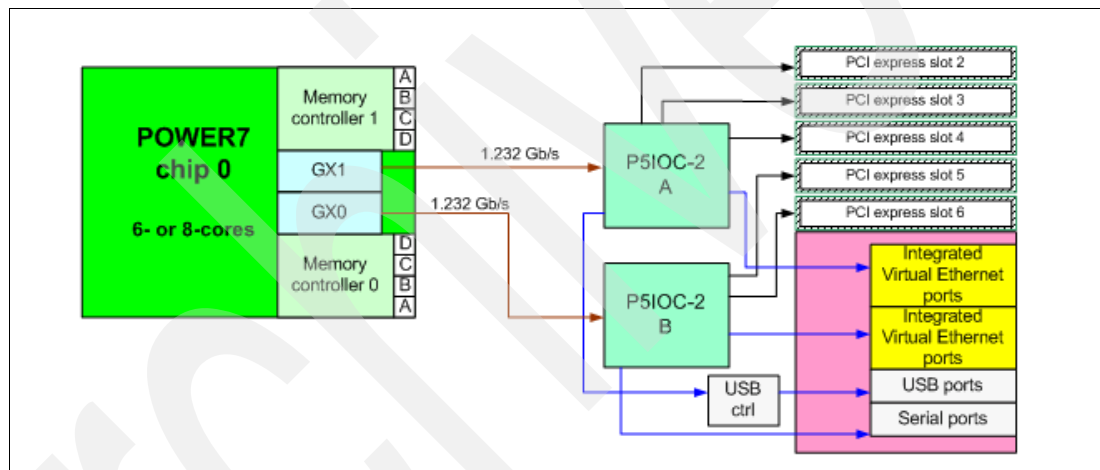


Figure 2-14 IVE system placement

A key design goal of the IVE is the capability to integrate one of the following options:

- ▶ #1803: Four 1 Gbps Ethernet ports
- ▶ #1804: Two 10 Gbps SFP+ optical (SR only) Ethernet ports and two 1 Gbps copper Ethernet ports
- ▶ #1813: Two 10 Gbps SFP+ copper twinax ports and two 1 Gbps Ethernet ports

The two P5IOC2 chips implemented inside any IBM POWER 770 drawer have IVE support. P5IOC2-A chip controls only two 1 Gbps Ethernet physical ports; P5IOC2-B controls either two 1 Gbps or two 10 Gbps Ethernet physical ports. IVE card assembly also provides the USB ports and the serial ports.

Each IVE feature code can address up to 64 logical Ethernet ports to support up to 64 LPARs. IVE physical port provides up to 16 logical Ethernet ports; a maximum of 16 MAC addresses are assigned to any physical port. If the IVE card is replaced, the new IVE card provides a new set of MAC addresses.

IVE does not have flash memory for its open firmware but it is stored in the service processor flash and then passed to POWER Hypervisor control. Therefore, flash code update is done by the POWER Hypervisor.

2.9 PCI adapters

Peripheral Component Interconnect Express (PCIe) uses a serial interface and allows for point-to-point interconnections between devices (using directly wired interface between these connection points). A single PCIe serial link is a dual-simplex connection that uses two pairs of wires, one pair for transmit and one pair for receive, and can transmit only one bit per cycle. It can transmit at the extremely high speed of 2.5 Gbps, which equates to a burst mode of 320 MBps on a single connection. These two pairs of wires are called a *lane*. A PCIe link can consist of multiple lanes. In such configurations, the connection is labeled as x1, x2, x8, x12, x16, or x32, where the number is effectively the number of lanes.

IBM offers only PCIe adapter options for the Power 770/780 system enclosure. If you want to use a PCI-extended (PCI-X) adapter, attach PCI-X DDR 12X I/O Drawer (#5796) by using a GX++ adapter loop. PCIe adapters use a different type of slot than PCI and PCI-X adapters. If you attempt to force an adapter into the wrong type of slot, you might damage the adapter or the slot. All adapters support Extended Error Handling (EEH).

IBM i IOPs are not supported, which means:

- ▶ Older PCI adapters that require an IOP are affected.
- ▶ Older I/O devices are affected, such as certain tape libraries or optical drive libraries, or any HVD SCSI device.
- ▶ Twinax displays or printers cannot be attached except through an OEM protocol converter.
- ▶ SDLC-attached devices using a LAN or WAN adapter are not supported.

SNA applications can still run when encapsulated inside TCP/IP, but the physical device attachment cannot be SNA, which means the earlier Fibre Channel and SCSI controllers that depended on an IOP being present are not supported.

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

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

If you are installing a new feature, ensure that you have the software required to support the new feature and determine whether there are any existing PTF prerequisites to install. See the IBM Prerequisite Web site for information:

https://www-912.ibm.com/e_dir/eServerPreReq.nsf

Tables of adapter features

The following sections discuss the supported adapters under various classifications (LAN, SCSI, SAS, Fibre Channel (FC), and so forth) and provide tables of orderable feature numbers. The tables indicate that the feature is supported by the AIX (A), IBM i (i), and Linux (L) operating systems.

2.9.1 LAN adapters

To connect a Power 770 and 780 local area network (LAN), you can use Integrated Virtual Ethernet. Other LAN adapters are supported in the system enclosure PCIe slots or in I/O enclosures that are attached to the system using a 12X technology loop. Table 2-14 lists the additional LAN adapters that are available.

Table 2-14 Available LAN adapters

Feature code	Adapter description	Slot	Size	OS support
5706	IBM 2-Port 10/100/1000 Base-TX Ethernet PCI-X Adapter	PCI-X	Short	A, i, L
5717	4-Port 10/100/1000 Base-TX PCI Express Adapter	PCIe	Short	A, L
5732	10 Gigabit Ethernet-CX4 PCI Express Adapter	PCIe	Short	A, L
5740	4-Port 10/100/1000 Base-TX PCI-X Adapter	PCI-X	Short	A, L
5767	2-Port 10/100/1000 Base-TX Ethernet PCI Express Adapter	PCIe	Short	A, i, L
5768	2-Port Gigabit Ethernet-SX PCI Express Adapter	PCIe	Short	A, i, L
5769	10 Gigabit Ethernet-SR PCI Express Adapter	PCIe	Short	A, L
5772	10 Gigabit Ethernet-LR PCI Express Adapter	PCIe	Short	A, i, L
5899	4-Port Gigabit Ethernet PCI Express Adapter	PCIe	Short	

2.9.2 Graphics accelerators

The Power 770 and 780 support up to eight graphics adapters. Table 2-15 lists the available graphic accelerators. They can be configured to operate in either 8-bit or 24-bit color modes. These adapters support both analog and digital monitors, and do not support hot-pluggable.

Table 2-15 Available graphics accelerators

Feature code	Adapter description	Slot	Size	OS support
2849 ^a	POWER GXT135P Graphics Accelerator with Digital Support	PCI-X	Short	A, L
5748	POWER GXT145 PCI Express Graphics Accelerator	PCIe	Short	A, L

a. Supported, but is no longer orderable.

2.9.3 SCSI and SAS adapters

To connect to external SCSI or SAS devices, the adapters that are listed in Table 2-16 are available to be configured.

Table 2-16 Available SCSI and SAS adapters

Feature code	Adapter description	Slot	Size	OS support
1912 ^a	PCI-X DDR Dual Channel Ultra320 SCSI Adapter	PCI-X	Short	A, i, L
2055	PCIe RAID & SSD SAS Adapter 3 Gb w/ Blind Swap Cassette	PCIe	Short	A, i, L
5646	Blind Swap Type III Cassette- PCIe, Short Slot	PCIe	Short	na
5647	Blind Swap Type III Cassette- PCI-X or PCIe, Standard Slot	PCI-X or PCIe	Short	na
5736	PCI-X DDR Dual Channel Ultra320 SCSI Adapter	PCI-X	Short	A, i, L
5901	PCIe Dual-x4 SAS Adapter	PCIe	Short	A, i, L
5903 ^a	PCIe 380MB Cache Dual - x4 3 Gb SAS RAID Adapter	PCIe	Short	A, i, L
5908	PCI-X DDR 1.5 GB Cache SAS RAID Adapter (BSC)	PCI-X	Long	A, i, L
5912	PCI-X DDR Dual - x4 SAS Adapter	PCI-X	Short	A, i, L
7863	PCI Blind Swap Cassette Kit, Double Wide Adapters, Type III	PCI	Short	na

a. Supported, but is no longer orderable.

Table 2-17 compares Parallel SCSI to SAS.

Table 2-17 Comparing Parallel SCSI to SAS

Items to compare	Parallel SCSI	SAS
Architecture	Parallel, all devices connected to shared bus	Serial, point-to-point, discrete signal paths
Performance	320 MBps (Ultra320 SCSI), performance degrades as devices are added to shared bus	3 Gbps, scalable to 12 Gbps, performance maintained as more devices are added
Scalability	15 drives	Over 16,000 drives
Compatibility	Incompatible with all other drive interfaces	Compatible with Serial ATA (SATA)
Max. Cable Length	12 meters total (must sum lengths of all cables used on bus)	8 meters per discrete connection, total domain cabling hundreds of meters
Cable Form Factor	Multitude of conductors adds bulk, cost	Compact connectors and cabling save space, cost
Hot Pluggability	No	Yes
Device Identification	Manually set, user must ensure no ID number conflicts on bus	Worldwide unique ID set at time of manufacture
Termination	Manually set, user must ensure proper installation and functionality of terminators	Discrete signal paths enable device to include termination by default

2.9.4 iSCSI

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

Table 2-18 lists the orderable iSCSI adapters.

Table 2-18 Available iSCSI adapters

Feature code	Adapter description	Slot	Size	OS support
5713	1 Gigabit iSCSI TOE PCI-X on Copper Media Adapter	PCI-X	Short	A, i, L

2.9.5 Fibre Channel adapter

The Power 770/780 servers support direct or SAN connection to devices that use Fibre Channel adapters. Table 2-19 summarizes the available Fibre Channel adapters.

All of these adapters except #5735 have LC connectors. If you are attaching a device or switch with an SC type fibre connector, an LC-SC 50 Micron Fiber Converter Cable (#2456) or an LC-SC 62.5 Micron Fiber Converter Cable (#2459) is required.

Table 2-19 Available Fibre Channel adapters

Feature code	Adapter description	Slot	Size	OS support
5735 ^a	8 Gigabit PCI Express Dual Port Fibre Channel Adapter	PCIe	Short	A, i, L
5749	4 Gbps Fibre Channel (2-Port)	PCI-X	Short	i
5759	4 Gb Dual-Port Fibre Channel PCI-X 2.0 DDR Adapter	PCI-X	Short	A, L
5774	4 Gigabit PCI Express Dual Port Fibre Channel Adapter	PCIe	Short	A, i, L

a. N_Port ID Virtualization (NPIV) capability is supported through VIOS.

2.9.6 Fibre Channel over Ethernet (FCoE)

There is a new protocol emerging. This new protocol, Fibre Channel over Ethernet (FCoE), which is being developed within T11 as part of the Fibre Channel Backbone 5 (FC-BB-5) project, is not meant to displace or replace FC. FCoE is an enhancement that expands FC into the Ethernet by combining two leading-edge technologies (FC and the Ethernet). This evolution with FCoE makes network consolidation a reality by the combination of Fibre Channel and Ethernet. This network consolidation will still maintain the resiliency, efficiency, and seamlessness of the existing FC-based data center.

Figure 2-15 on page 58 shows comparison between existing FC and network connection and FCoE connection.

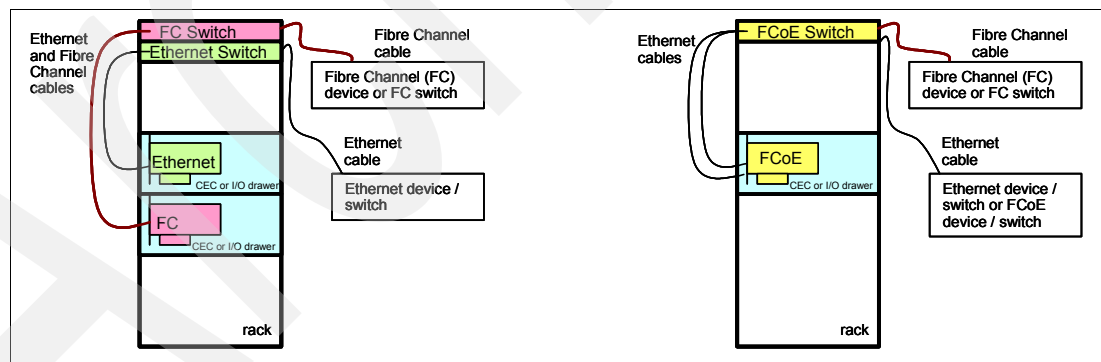


Figure 2-15 comparison between existing FC and network connection and FCoE connection

For more information about FCoE, read *An Introduction to Fibre Channel over Ethernet, and Fibre Channel over Convergence Enhanced Ethernet*, REDP-4493

IBM offer 10 Gb FCoE PCIe Dual Port Adapter (#5708) that is a high performance, 10 Gb, dual port, PCIe Converged Network Adapter (CNA) utilizing SR optics. Each port can provide NIC (Network Interface Card) traffic and Fibre Channel functions simultaneously. It is supported on AIX and Linux for FC and Ethernet.

2.9.7 InfiniBand Host Channel adapter

The InfiniBand Architecture (IBA) is an industry-standard architecture for server I/O and inter-server communication. It was developed by the InfiniBand Trade Association (IBTA) to provide the levels of reliability, availability, performance, and scalability necessary for present and future server systems with levels significantly better than can be achieved by using bus-oriented I/O structures.

InfiniBand (IB) is an open set of interconnect standards and specifications. The main IB specification has been published by the InfiniBand Trade Association and is available at:

<http://www.infinibandta.org/>

InfiniBand is based on a switched fabric architecture of serial point-to-point links, where these IB links can be connected to either host channel adapters (HCAs), used primarily in servers, or to target channel adapters (TCAs), used primarily in storage subsystems.

The InfiniBand physical connection consists of multiple byte lanes. Each individual byte lane is a four-wire, 2.5, 5.0, or 10.0 Gbps bidirectional connection. Combinations of link width and byte-lane speed allow for overall link speeds from 2.5 Gbps to 120 Gbps. The architecture defines a layered hardware protocol as well as a software layer to manage initialization and the communication between devices. Each link can support multiple transport services for reliability and multiple prioritized virtual communication channels.

For more information about Infiniband, read *HPC Clusters Using InfiniBand on IBM Power Systems Servers*, SG24-7767.

IBM offers the GX++ 12X DDR Adapter (#1808) that plugs into the system backplane (GX++ slot). There are two GX++ slots in each CEC enclosure. By attaching a 12X to 4X converter cable (#1828), an IB switch can be attached.

2.9.8 Asynchronous adapter

Asynchronous PCI adapters provide connection of asynchronous EIA-232 or RS-422 devices. If you have a cluster configuration or high-availability configuration and plan to connect the IBM Power Systems using a serial connection, you may use one of the features listed in Table 2-20.

Table 2-20 Available Asynchronous adapter

Feature code	Adapter description	Slot	Size	OS support
2728	4-port USB PCIe Adapter	PCIe	Short	A, L
5785	4-Port Asynchronous EIA-232 PCIe Adapter	PCIe	Short	A, L

2.10 Internal storage

Introduced with POWER6 processor-based servers, Serial Attached SCSI (SAS) drives the Power 770 internal disk subsystem. SAS 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; the other differential signal transmits in the opposite direction. Data can be transmitted in both directions simultaneously.

There are two PCIe integrated SAS controllers under the P5IOC2-B chip and also the PCI-X SAS controller that is directly connected to the DVD media bay (see Figure 2-16 on page 60).

Power 770 supports various internal storage configurations:

- ▶ Dual split backplane mode
- ▶ Triple split backplane mode
- ▶ Dual storage IOA configuration using internal disk drives
- ▶ Dual storage IOA configuration using internal disk drives and external enclosure

Each SAS port expander is connected to all the six disk drive bays, but depending on the combination of feature codes configured for the Power 770 enclosure, each SAS port expander can drive two or three disk drive bays. Power 770 and Power 780, with more than one enclosure, support enclosures with different internal storage configurations.

An optional 175 MB cache RAID – Dual IOA enablement card (#5662) is used to enable write cache on the two embedded SAS RAID adapters of the disk or media backplane by providing the necessary rechargeable batteries for memory backup. It also enables the two embedded SAS RAID adapters to work as dual storage IOAs, that is, high availability (HA) RAID mode. This feature plugs in to the disk or media backplane and enables a 175 MB write cache on each of the two embedded RAID adapters by providing two rechargeable batteries with associated charger circuitry. The write cache can provide additional I/O performance for attached disk or solid-state drives, particularly for RAID 5 and RAID 6. The write cache contents are mirrored for redundancy between the two RAID adapters resulting in an effective write cache size of 175 MB. The batteries provide power to maintain both copies of write-cache information in the event power is lost.

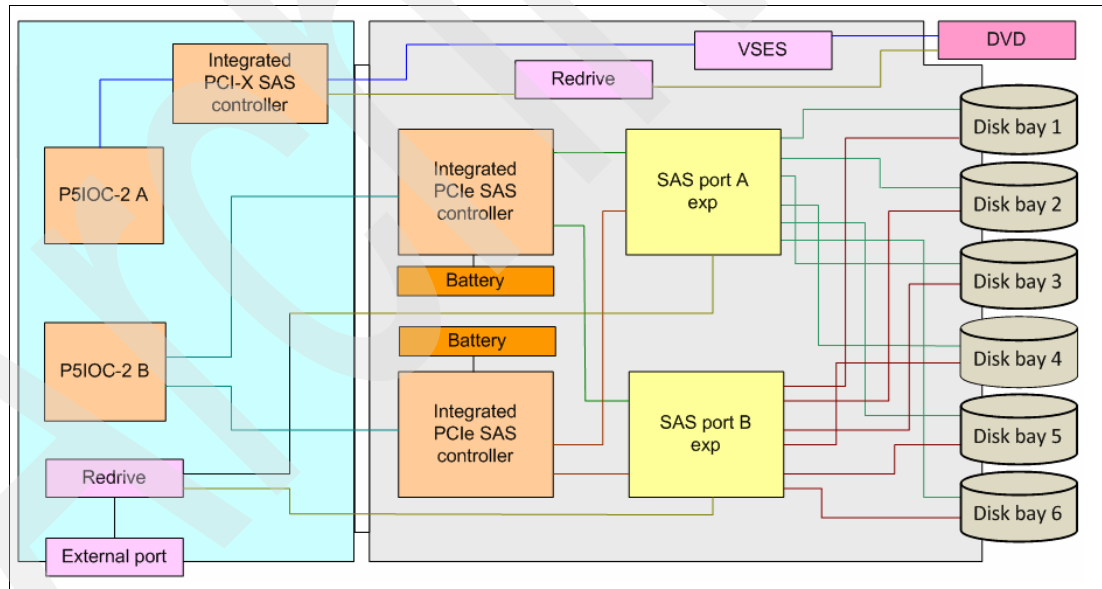


Figure 2-16 Internal SAS topology overview

Table 2-21 summarizes the internal storage combination and the feature codes required for any combination.

Table 2-21 SAS configurations summary

SAS subsystem configuration	#5662	External SAS components	SAS port cables	SAS cables	Notes
Two-way split backplane	No	None	None	N/A	IBM i does not support this combination. Connecting to an external disk enclosure is not supported.
Three-way split backplane	No	Dual x4 SAS adapter (#5901)	Internal SAS port (#1815) SAS cable for three-way split backplane	AI cable (#3679) - Adapter to internal drive (1 meter)	IBM i does not support this combination. An I/O adapter can be located in another enclosure of the system.
Dual storage IOA with internal disk	Yes	None	None	N/A	Internal SAS port cable (#1815) cannot be used with this or HA RAID configuration.
Dual storage IOA with internal disk and external disk enclosure	Yes	Requires an external disk enclosure (#5886)	Internal SAS port (#1819) SAS cable assembly for connecting to an external SAS drive enclosure	#3686 or #3687	#3686 is a 1-meter cable. #3687 is a 3-meter cable.

2.10.1 Dual split backplane mode

Dual split backplane mode offers two packs of three disks and is the standard configuration. If desired, one of the packs can be connected to an external SAS PCIe or PCI-X adapter if #1819 is selected. Figure 2-17 on page 62 shows how the six disk bays are shared with the dual split backplane mode. Although solid-state drives (SSDs) are supported with a dual split backplane configuration, mixing SSDs and hard disk drives (HDDs) in the same split domain is not possible. Also, mirroring SSDs with HDDs is not possible.

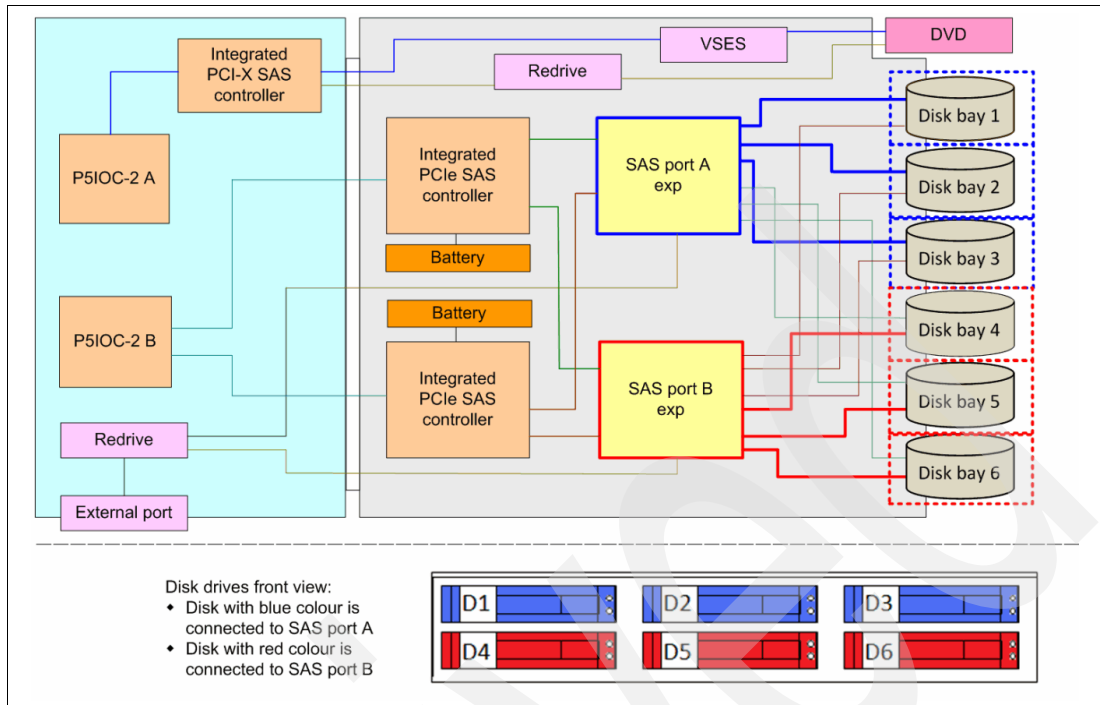


Figure 2-17 Dual split backplane overview

2.10.2 Triple split backplane

Triple split backplane mode offers three packs of two disk drives each. This mode requires #1815 internal SAS cable, the SAS cable #3679, and a SAS controller, such as #5901. Figure 2-18 on page 63 shows how the six disk bays are shared with the triple split backplane mode. The PCI adapter that drives two of the six disks can be located in the same Power 770 (or Power 780) enclosure as the disk drives or adapter, even in a different system enclosure or external I/O drawer.

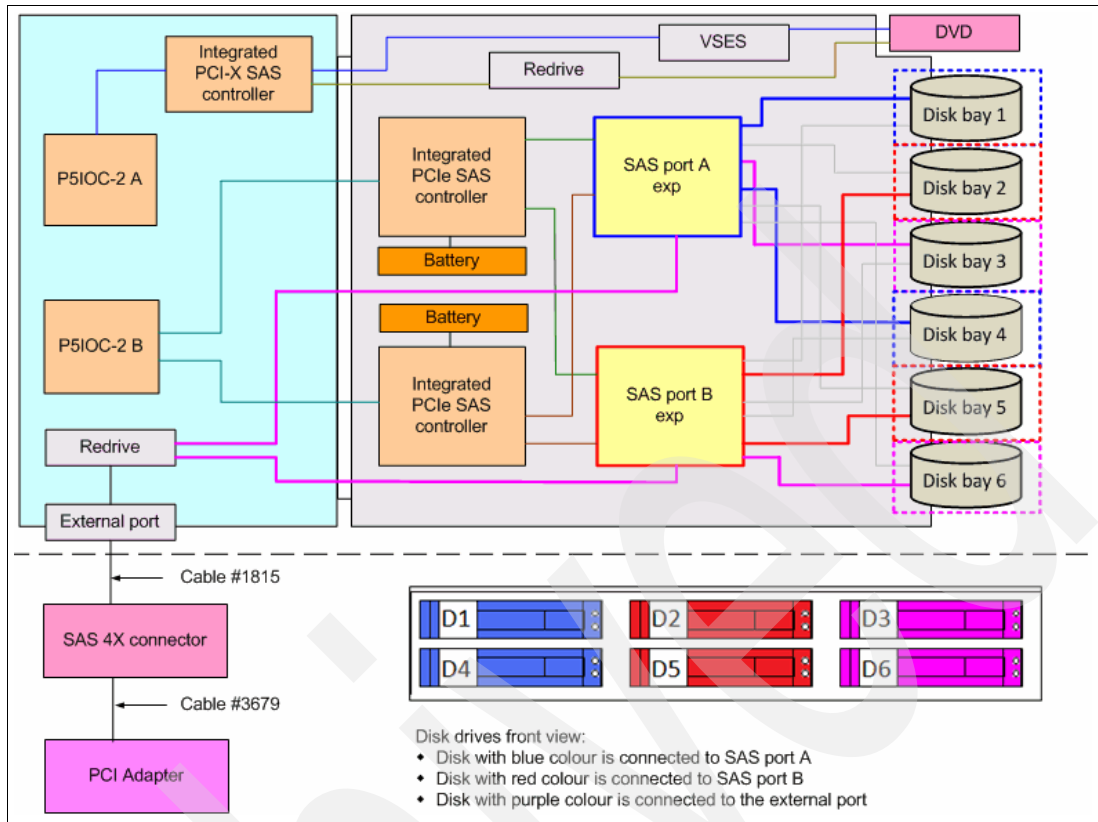


Figure 2-18 Triple split backplane overview

Although SSDs are supported with a triple split backplane configuration, mixing SSDs and Hard Disk Drives (HDDs) in the same split domain is not possible. Also, mirroring SSDs with HDDs is not possible.

2.10.3 Dual storage IOA configurations

The Dual storage IOA configurations are available with internal or internal with external disk drives from another I/O drawer. SSDs are not supported with this mode.

If #1819 is selected for an enclosure, selecting SAS cable #3686 or #3687 to support RAID internal and external drives is necessary. See Figure 2-19 on page 64. If #1819 is not selected for the enclosure, the RAID supports only enclosure internal disks.

This configuration increases availability using dual storage IOA or high availability (HA) to connect multiple adapters to a common set of internal disk drives. It also increases the performance of RAID arrays. The following rules apply to this configuration:

- ▶ This configuration uses the 175 MB Cache RAID – Dual IOA enablement card.
- ▶ Using the dual IOA enablement card, the two embedded adapters can connect to each other and all six disk drives. as well as, the 12 disk drives in an external disk drive enclosure if one is used.
- ▶ The disk drives are required to be in RAID arrays.
- ▶ There are no separate SAS cables required to connect the two embedded SAS RAID adapters to each other. The connection is contained within the backplane.
- ▶ RAID 0, 10, 5, and 6 support up to six drives.

- ▶ Solid-state drives (SSD) and hard disk drives (HDD) can be used, but can never be mixed in the same disk enclosure
- ▶ To connect to the external storage, you need to connect to the #5886 disk drive enclosure.

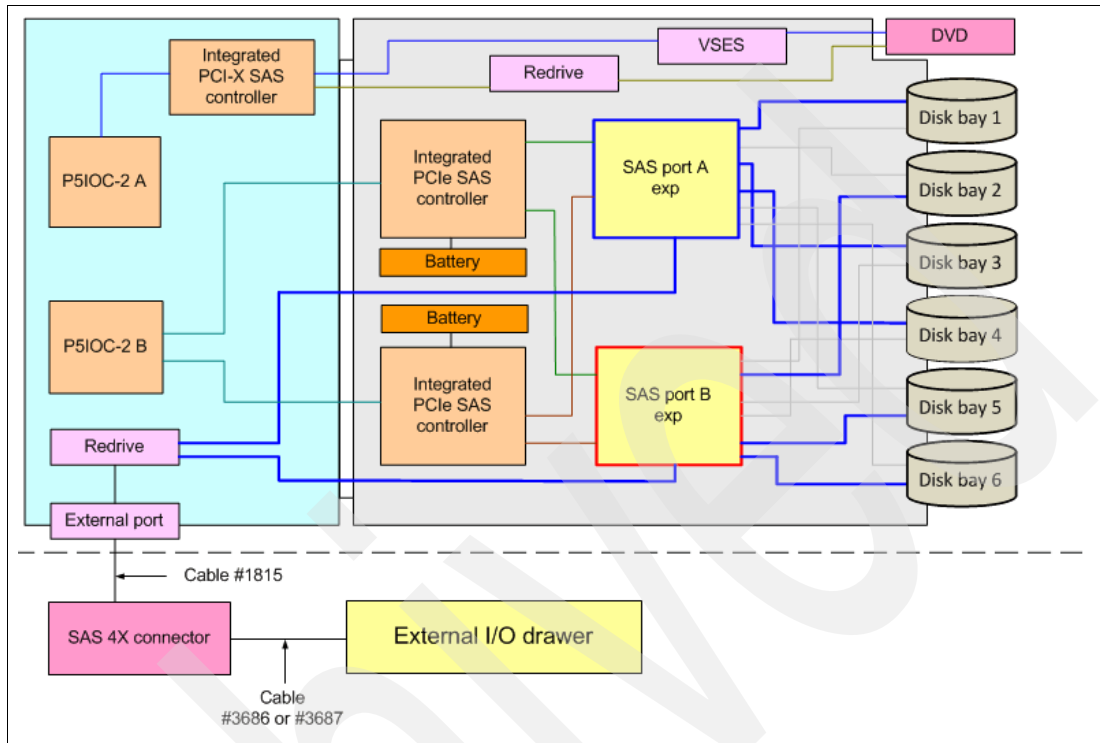


Figure 2-19 RAID mode (external disk drives option)

2.10.4 DVD

The DVD media bay is directly connected to the PCI-X controller on the I/O backplane and has a specific chip (VSES) for controlling the DVD LED and power. The VSES appears as a separate device to the device driver and operating systems (see Figure 2-16 on page 60).

2.11 External I/O subsystems

This section describes the external I/O subsystems, which include the I/O drawers, the PCI-DDR 12X Expansion Drawer (#5796), 12X I/O Drawer PCIe, SFF disk (#5802), 12X I/O Drawer PCIe, No Disk (#5877), and EXP 12S Expansion Drawer (#5886).

Table 2-22 provides an overview of all the supported I/O drawers.

Table 2-22 I/O drawer capabilities

Drawer feature code	DASD	PCI slots	Requirements for a 770/780
5796	-	6 x PCI-X	GX++ adapter card #1808
5802	18 x SAS disk drive bays	10 x PCIe	GX++ adapter card #1808
5877	-	10 x PCIe	GX++ adapter card #1808

Drawer feature code	DASD	PCI slots	Requirements for a 770/780
5886	12 x SAS disk drive bays	-	Any supported SAS adapter

The two GX++ buses from the second processor card feed two GX++ Adapter slots. An optional GX++ 12X DDR Adapter, Dual-port (#1808) which is installed in GX++ Adapter slot enables the attachment of a 12X loop, which runs at either SDR or DDR speed depending on the 12X I/O drawers that are attached.

2.11.1 PCI-DDR 12X Expansion drawer (#5796)

The PCI-DDR 12X Expansion Drawer (#5796) is a 4U (EIA units) drawer and mounts in a 19-inch rack. Feature #5796 is 224 mm (8.8 in.) wide and takes up half the width of the 4U (EIA units) rack space. The 4U enclosure can hold up to two #5796 drawers mounted side-by-side in the enclosure. The drawer is 800 mm (31.5 in.) deep and can weigh up to 20 kg (44 lb).

The PCI-DDR 12X Expansion Drawer has six 64-bit, 3.3 V, PCI-X DDR slots, running at 266 MHz and that use blind-swap cassettes and support hot-plugging of adapter cards. The drawer includes redundant hot-plug power and cooling.

Two interface adapters are available for use in the #5796 drawer: Dual-Port 12X Channel Attach Adapter Long Run (#6457) or Dual-Port 12X Channel Attach Adapter Short Run (#6446). The adapter selection is based on how close the host system or the next I/O drawer in the loop is physically located. Feature #5796 attaches to a host system CEC enclosure with a 12X adapter in a GX++ slot through SDR or DDR cables (or both SDR and DDR cables). A maximum of four #5796 drawers can be placed on the same 12X loop. Mixing #5802/5877 and #5796 on the same loop is not supported.

A minimum configuration of two 12X cables (either SDR or DDR), two AC power cables, and two SPCN cables is required to ensure proper redundancy.

The Figure 2-20 on page 66 shows the back view of the expansion unit.

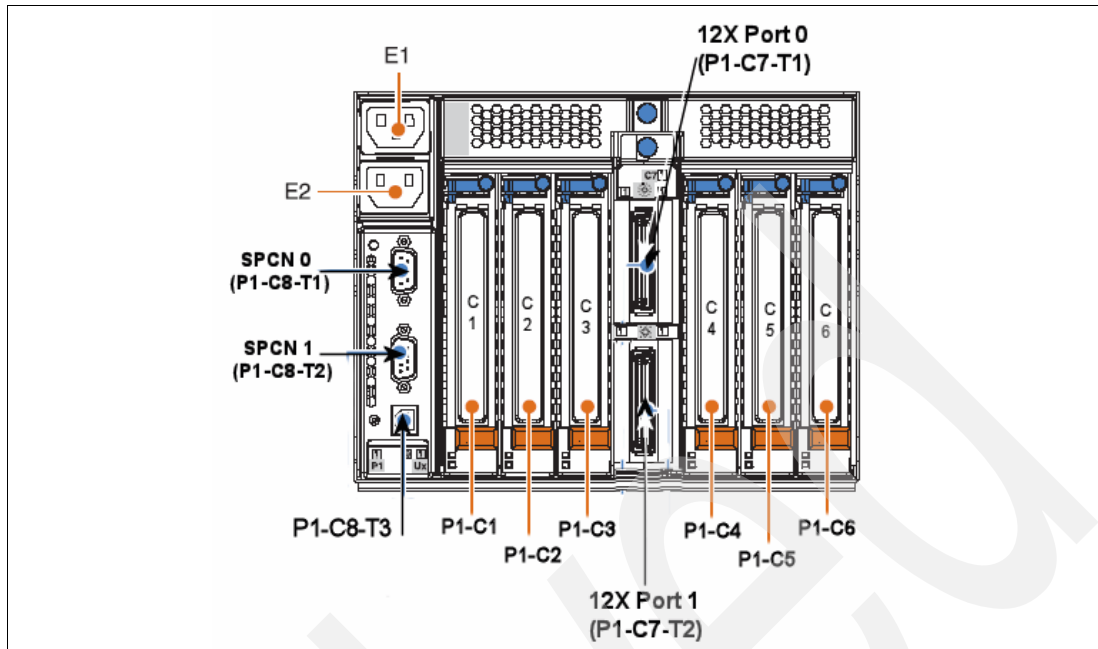


Figure 2-20 PCI-X DDR 12X Expansion Drawer rear side

2.11.2 12X I/O Drawer PCIe (#5802 and #5877)

The 12X I/O Drawer PCIe is a 19-inch I/O and storage drawer. It provides a 4U-tall (EIA units) drawer containing ten PCIe based I/O adapter slots and eighteen SAS hot-swap Small Form Factor disk bays, which can be used for either disk drives or SSD. The adapter slots use blind-swap cassettes and support hot-plugging of adapter cards.

A maximum of two #5802 drawers can be placed on the same 12X loop. Feature #5877 is the same as #5802 except it does not support any disk bays. Feature #5877 can be on the same loop as #5802. Feature #5877 can not be upgraded to #5802.

The physical dimensions of the drawer measure 444.5 mm (17.5 in.) wide by 177.8 mm (7.0 in.) high by 711.2 mm (28.0 in.) deep for use in a 19-inch rack.

A minimum configurations of two 12X DDR cables, two AC power cables, and two SPCN cables is required to ensure proper redundancy. The drawer attaches to the host CEC enclosure with a 12X adapter in a GX++ slot through 12X DDR cables that are available in various cable lengths: 0.6 (#1861), 1.5 (#1862), 3.0 (#1865), or 8 meters (#1864). The 12X SDR cables are not supported.

Figure 2-21 on page 67 shows the front view of the 12X I/O Drawer PCIe (#5802).

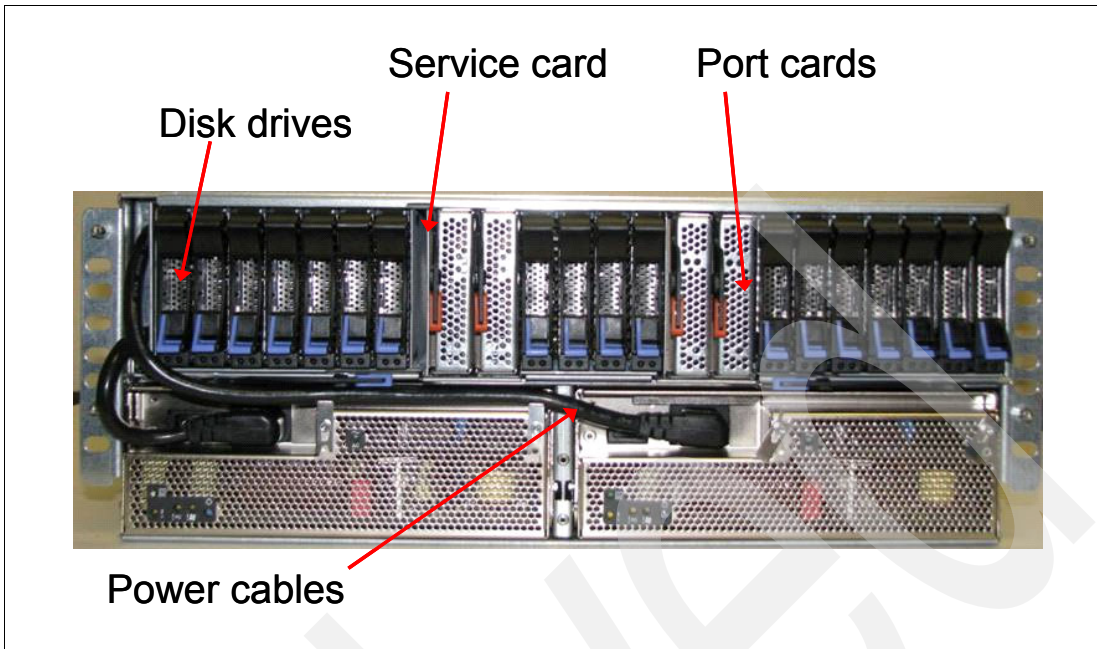


Figure 2-21 Front view of the 12X I/O Drawer PCIe

Figure 2-22 shows the rear view of the 12X I/O Drawer PCIe (#5802).

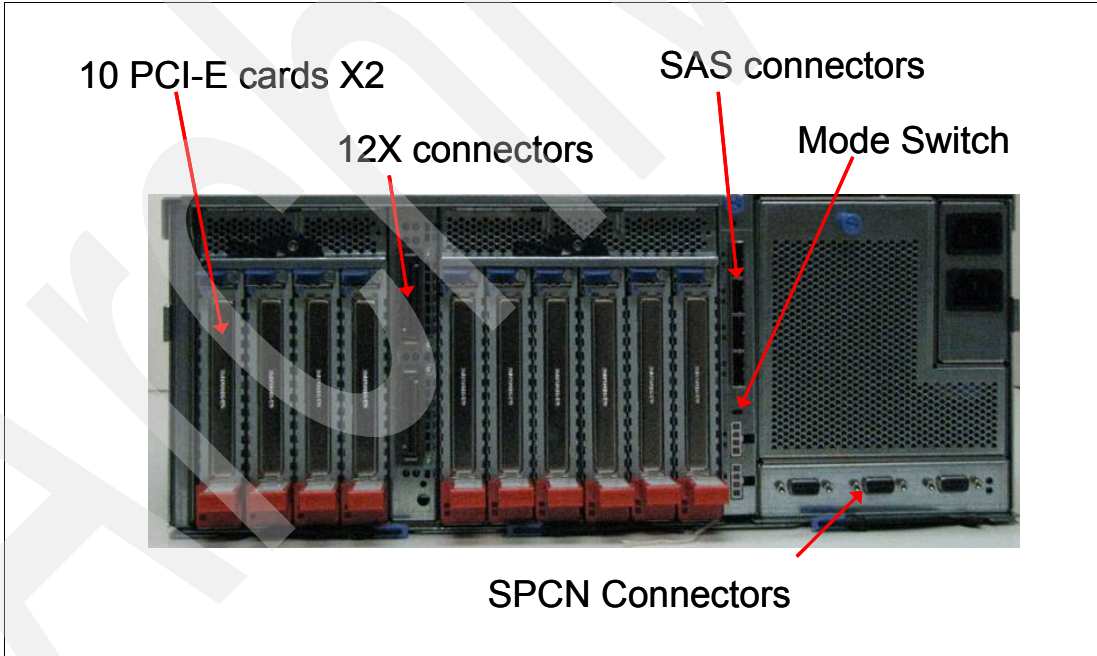


Figure 2-22 rear view of the 12X I/O Drawer PCIe

2.11.3 Dividing SFF drive bays in 12X I/O drawer PCIe

Disk drive bays in 12X I/O drawer PCIe can be configured as one, two, or four set. This allows for partitioning of disk bays. Disk bay partitioning configuration can be done via physical mode switch on the I/O drawer.

Note: Mode change using physical mode switch requires power-off/on of drawer.

Figure 2-22 on page 67 indicates the Mode Switch in the rear view of the #5802 I/O Drawer.

Each disk bay set can be attached to its own controller or adapter. #5802 PCIe 12X I/O Drawer has four SAS connections to drive bays. It connects to PCIe SAS adapters or controllers on the host system.

Figure 2-23 shows the configuration rule of disk bay partitioning in #5802 PCIe 12X I/O Drawer. There is no specific feature code for mode switch setting.

Note: IBM System Planning Tool supports disk bay partitioning. Also, the IBM configuration tool accepts this configuration from IBM System Planning Tool and passes it through IBM manufacturing using Customer Specified Placement (CSP) option.

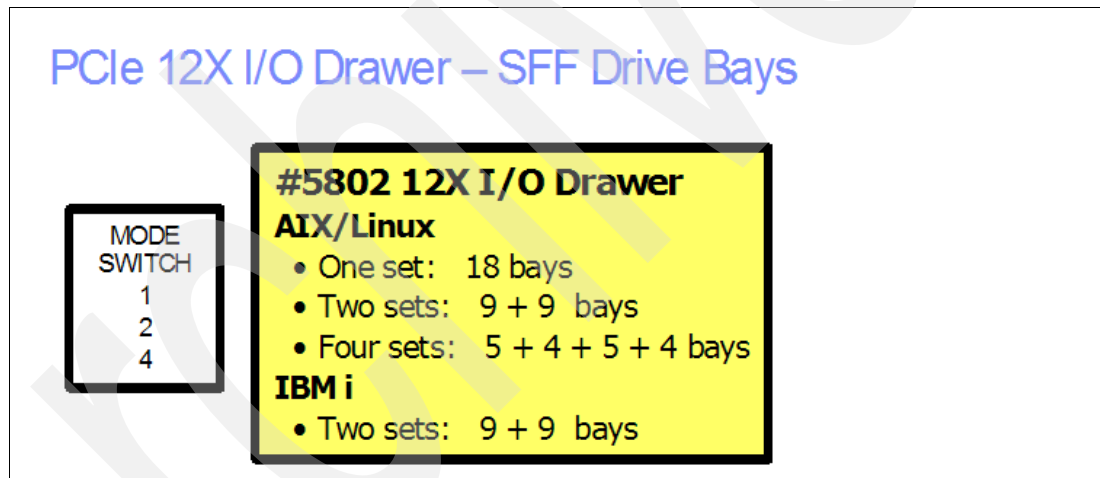


Figure 2-23 Disk Bay Partitioning in #5802 PCIe 12X I/O drawer

The SAS ports as associated with the mode selector switch map to the disk bays have the mappings shown in Table 2-23.

Table 2-23 SAS connection mappings

Location code	Mappings	Number of bays
P4-T1	P3-D1 to P3-D5	5 bays
P4-T2	P3-D6 to P3-D9	4 bays
P4-T3	P3-D10 to P3-D14	5 bays
P4-T3	P3-D15 to P3-D18	4 bays

The location codes for the front and rear views of the #5802 I/O drawer are provided in Figure 2-24 on page 69 and Figure 2-25 on page 69.

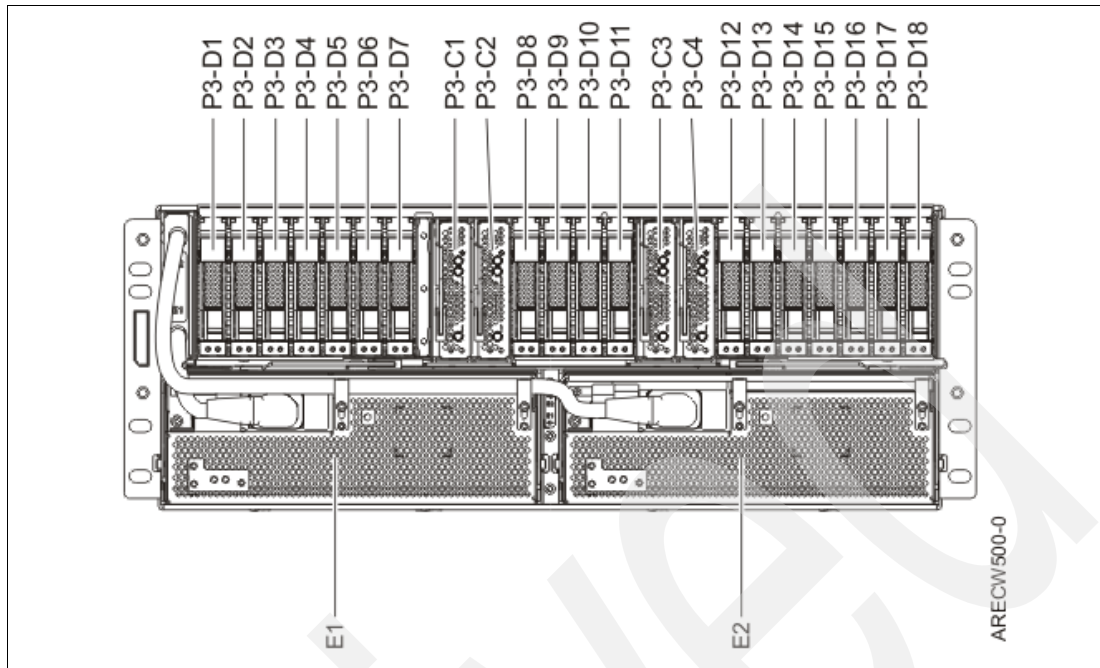


Figure 2-24 5802 I/O drawer front view location codes

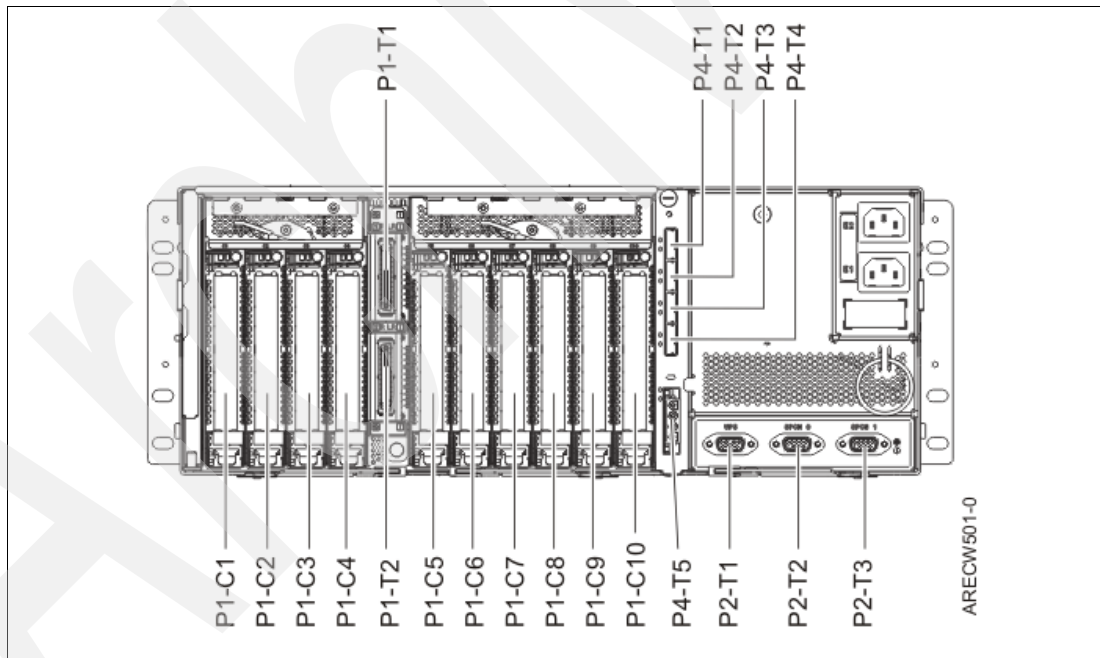


Figure 2-25 5802 I/O drawer rear view location codes

Configuring the #5802 disk drive subsystem

The #5802 SAS disk drive enclosure can hold up to 18 disk drives. The disks in this enclosure can be organized in several configurations depending on the operating system used, the type of SAS adapter card, and the position of the mode switch.

Each disk bay set can be attached to its own controller or adapter. Feature #5802 PCIe 12X I/O Drawer has four SAS connections to drive bays. It connects to PCIe SAS adapters or controllers on the host systems.

For detailed information about how to configure, see the IBM Power Systems Hardware Information Center at:

<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp>

2.11.4 12X I/O Drawer PCIe and PCI-DDR 12X Expansion Drawer 12X cabling

I/O Drawers are connected to the adapters in the CEC enclosure with data transfer cables: 12X DDR cables for the #5802 and #5877 I/O drawers, and 12X SDR and/or DDR cables for the #5796 I/O drawers. The first 12X I/O Drawer that is attached in any I/O drawer loop requires two data transfer cables. Each additional drawer, up to the maximum allowed in the loop, requires one additional data transfer cable. Note the following information:

- ▶ A 12X I/O loop starts at a CEC bus adapter port 0 and attaches to port 0 of an I/O drawer.
- ▶ I/O drawer attaches from port 1 of the current unit to port 0 of the next I/O drawer.
- ▶ Port 1 of the last I/O drawer on the 12X I/O loop connects to port 1 of the same CEC bus adapter to complete the loop.

Figure 2-26 shows typical 12X I/O loop port connections.

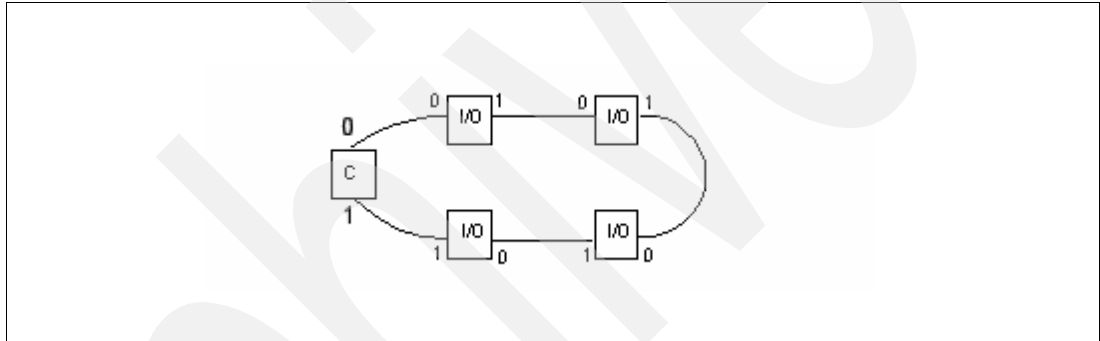


Figure 2-26 Typical 12X I/O loop port connections

Table 2-24 shows various 12X cables to satisfy the various length requirements:

Table 2-24 12X connection cables

Feature code	Description
1861	0.6 Meter 12X DDR Cable
1862	1.5 Meter 12X DDR Cable
1865	3.0 Meter 12X DDR Cable
1864	8.0 Meter 12X DDR Cable

General rule for 12X IO Drawer configuration

Use multiple GX++ buses, as many as possible. Figure 2-27 shows an example of a 12X IO Drawer configuration.

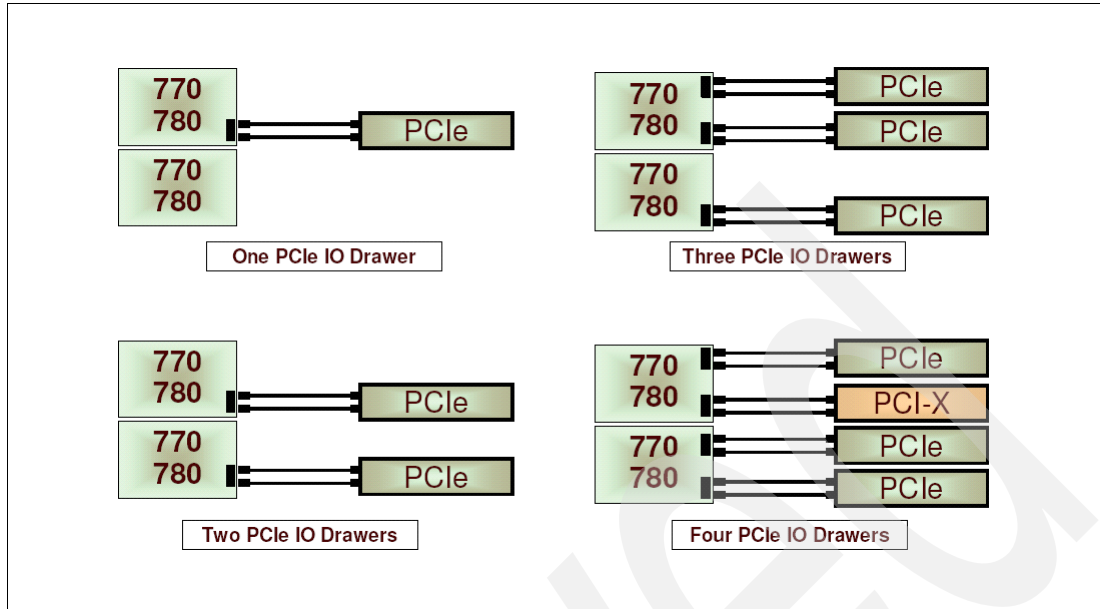


Figure 2-27 12X IO Drawer configuration

Supported 12X cable length for PCI-DDR 12X Expansion Drawer

Each #5796 drawer requires one Dual Port PCI DDR 12X Channel Adapter, either Short Run (#6446) or Long Run (#6457). The choice of adapters is dependent on the distance to the next 12X Channel connection in the loop, either to another I/O drawer or the system unit. Table 2-25 identifies the supported cable lengths for each 12X channel adapter. I/O drawers containing the Short Range adapter can be mixed in a single loop with I/O drawers containing the Long Range adapter. In Table 2-25, a “Yes” indicates that the 12X cable identified in that column can be used to connect the drawer configuration identified to the left. A “No” means it cannot be used.

Table 2-25 Supported 12X cable length

Connection type	12X cable options			
	0.6 M	1.5 M	3.0 M	8.0 M
5796 to 5796 with 6446 in both drawers	Yes	Yes	No	No
5796 with 6446 adapter to 5796 with 6457 adapter	Yes	Yes	Yes	No
5796 to 5796 with 6457 adapter in both drawers	Yes	Yes	Yes	Yes
5796 with 6446 adapter to system unit	No	Yes	Yes	No
5796 with 6457 adapter to system unit	No	Yes	Yes	Yes

2.11.5 12X I/O Drawer PCIe and PCI-DDR 12X Expansion Drawer SPCN cabling

System Power Control Network (SPCN) is used to control and monitor the status of power and cooling within the I/O drawer.

SPCN cables connect all AC-powered expansion units as shown in the example diagram, Figure 2-28. Note the following information:

- ▶ Start at SPCN 0 (T1) of the first (top) CEC unit to J15 (T1) of the first expansion unit.
- ▶ Cable all units from J16 (T2) of the previous unit to J15 (T1) of the next unit.
- ▶ From J16 (T2) of the final expansion unit, connect to the second CEC, SPCN 1 (T2).
- ▶ To complete the cabling loop, connect SPCN 1 (T2) of the topmost (first) CEC to the SPCN 0 (T1) of the next (second) CEC.
- ▶ Ensure a complete loop exists from the topmost CEC, through all attached expansions and back to the next lower (second) CEC drawer.

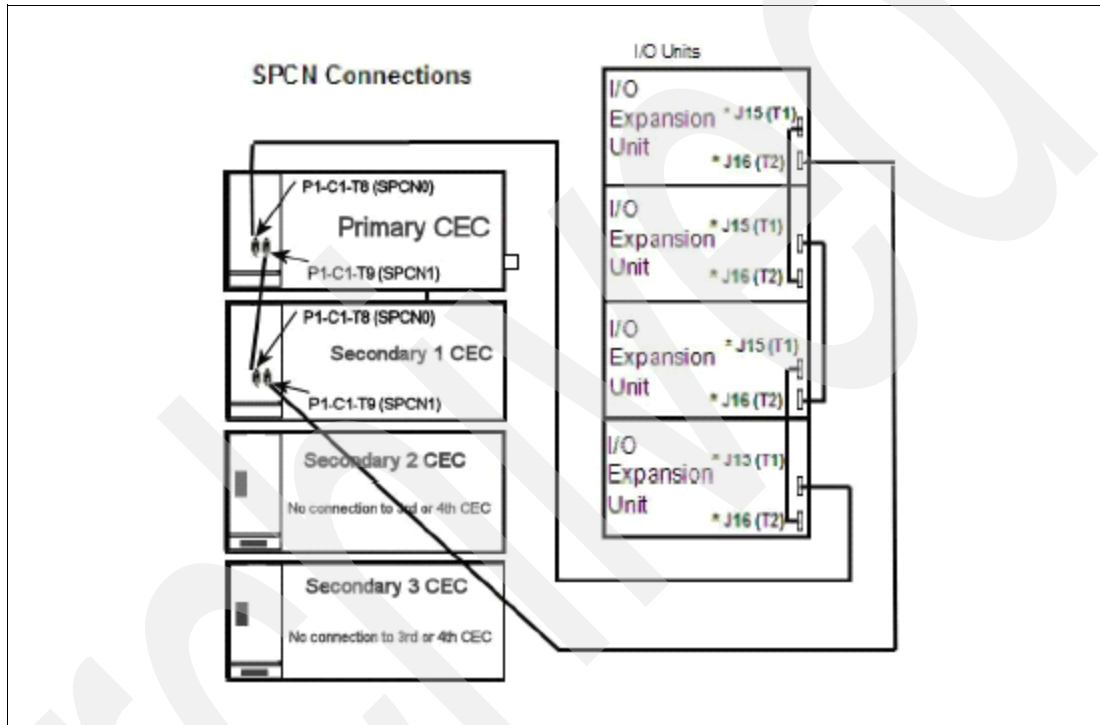


Figure 2-28 SPCN cabling examples

Note: Only the first two CECs of a multi-CEC system are included in SPCN cabling with I/O expansion units; CECs three and four do not connect.

Table 2-26 shows the SPCN cables to satisfy various length requirements:

Table 2-26 SPCN cables

Feature code	Description
6006	SPCN cable drawer-to-drawer, 3 m
6007	SPCN cable rack-to-rack, 15 m

2.12 External disk subsystems

This section describes the external disk subsystems, which include the EXP 12S Expansion Drawer and supported IBM System Storage family of products.

2.12.1 EXP 12S Expansion Drawer

The EXP 12S Expansion Drawer (#5886) is a 2U (EIA units) drawer and mounts in a 19-inch rack. The drawer can hold either SAS disk drives or SSD. The EXP 12S Expansion Drawer has twelve (12) 3.5-inch SAS disk bays with redundant data paths to each bay. The drawer supports redundant hot-plug power and cooling and redundant hot-swap SAS expanders (Enclosure Services Manager-ESM). Each ESM has an independent SCSI Enclosure Services (SES) diagnostic processor.

The SAS disk drives or SSD contained in the EXP12S are controlled by one or two PCIe or PCI-X SAS adapters connected to the EXP12S through SAS cables. The SAS cable varies, depending on the adapter being used, the operating system being used, and the protection desired. Note the following information:

- ▶ The large cache PCI-X DDR 1.5 GB Cache SAS RAID Adapter (#5904) and PCI-X DDR 1.5 GB Cache SAS RAID Adapter (BSC) (#5908) uses a SAS Y cable when a single port is running the EXP12S. A SAS X cable is used when a pair of adapters is used for controller redundancy.
- ▶ The medium cache PCIe 380 MB Cache Dual - x4 3 Gb SAS RAID Adapter (#5903) is always paired and use a SAS X cable to attach the feature #5886 I/O drawer.
- ▶ The zero cache PCI-X DDR Dual - x4 SAS Adapter (#5912) and PCIe Dual-x4 SAS Adapter (#5901) use a SAS Y cable when a single port is running the EXP12S. A SAS X cable is used for AIX or Linux environments when a pair of adapters is used for controller redundancy.

In all of these configurations, all 12 SAS bays are controlled by a single controller or a single pair of controllers.

A second EXP12S drawer can be attached to another drawer by using two SAS EE cables, providing 24 SAS bays instead of 12 bays for the same SAS controller port. This configuration is called cascading. In this configuration, all 24 SAS bays are controlled by a single controller or a single pair of controllers.

There is a maximum of up to 110 EXP12S SAS DASD/SSD I/O drawers on SAS PCI controllers.

The #5886 can be directly attached to the SAS port on the rear of the Power 770 and 780, providing a very low-cost disk storage solution.

Adding the optional 175 MB Cache RAID - Dual IOA Enablement Card (#5662) to the Power 770 and 780 causes the pair of embedded controllers in that processor enclosure to be configured as dual controllers, accessing all six SAS bays. Using the internal SAS Cable Assembly for SAS Port (#1819) connected to the rear port, the pair of embedded controllers is now running eighteen (18) SAS bays (six SFF bays in the system unit and twelve 3.5-inch bays in the drawer). The disk drawer is attached to the SAS port with a SAS YI cable. In this 18-bay configuration, all drives must be HDD.

A second unit can not be cascaded to a #5886 attached in this way.

2.12.2 EXP 24S Expansion Drawer

The EXP 24S Expansion Drawer (#5887) is a 2U (EIA units) drawer and mounts in a 19-inch rack. The drawer holds SAS disk drives. The EXP 24S Expansion Drawer has twenty-four (24) 2.5-inch SAS disk bays with redundant data paths to each bay. The drawer supports redundant hot-plug power and cooling and redundant hot-swap SAS expanders (Enclosure Services Manager-ESM). Each ESM has an independent SCSI Enclosure Services (SES) diagnostic processor.

The SAS disk drives contained in the EXP24S are controlled by one or two PCIe or PCI-X SAS adapters connected to the EXP24S through SAS cables. The SAS cable varies, depending on the adapter being used, the operating system being used, and the protection desired. Note the following information:

The large cache PCI-X DDR 1.5 GB Cache SAS RAID Adapter (#5904) and PCI-X DDR 1.5 GB Cache SAS RAID Adapter (BSC) (#5908) uses a SAS Y cable when a single port is running the EXP24S. A SAS X cable is used when a pair of adapters is used for controller redundancy.

The medium cache PCIe 380 MB Cache Dual - x4 3 Gb SAS RAID Adapter (# 5903) is always paired and use a SAS X cable to attach the feature #5887 I/O drawer.

The zero cache PCI-X DDR Dual - x4 SAS Adapter (#5912) and PCIe Dual-x4 SAS Adapter (#5901) use a SAS Y cable when a single port is running the EXP24S. A SAS X cable is used for AIX or Linux environments when a pair of adapters is used for controller redundancy.

In all of these configurations, all 24 SAS bays are controlled by a single controller or a single pair of controllers.

A second EXP24S drawer can be attached to another drawer by using two SAS EE cables, providing 48 SAS bays instead of 24 bays for the same SAS controller port. This configuration is called cascading. In this configuration, all 48 SAS bays are controlled by a single controller or a single pair of controllers.

There is a maximum of up to 110 EXP12S SAS DASD/SSD I/O drawers on SAS PCI controllers.

The #5887 can be directly attached to the SAS port on the rear of the Power 770 and 780, providing a very low-cost disk storage solution.

Adding the optional 175 MB Cache RAID - Dual IOA Enablement Card (#5662) to the Power 770 and 780 causes the pair of embedded controllers in that processor enclosure to be configured as dual controllers, accessing all six SAS bays. Using the internal SAS Cable Assembly for SAS Port (#1819) connected to the rear port, the pair of embedded controllers is now running thirty (30) SAS bays (six SFF bays in the system unit and twenty-four 2.5-inch bays in the drawer). The disk drawer is attached to the SAS port with a SAS YI cable. In this 30-bay configuration, all drives must be HDD.

A second unit can not be cascaded to a #5887 attached in this way.

2.12.3 IBM System Storage

The IBM System Storage Disk Systems products and offerings provide compelling storage solutions with superior value for all levels of business.

IBM System Storage N series

IBM N series unified system storage solutions can provide customers with the latest technology to help them improve performance, virtualization manageability, and system efficiency at a reduced total cost of ownership. Several enhancements have been incorporated to the N series product line, to complement and reinvigorate this portfolio of solutions:

- ▶ The new SnapManager for Hyper-V provides extensive management for backup, restore and replication for Microsoft Hyper-V environments.
- ▶ New N series Software Packs allow customers to get the benefits of a broad set of N series solutions at a noticeably reduced cost.
- ▶ An essential component to this launch is Fibre Channel over Ethernet access and 10 Gb Ethernet, to help integrate Fibre Channel and Ethernet flow into a unified network, and leverage current Fibre Channel installations at the time

For more information, visit the following address:

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

IBM System Storage DS3000 family

The IBM System Storage DS3000 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 450 GB SAS drives, increased data protection features like RAID 6, and more FlashCopy® images per volume, provide a reliable virtualization platform with the support of Microsoft Windows Server 2008 with HyperV.

For more information, visit the following address:

<http://www.ibm.com/systems/storage/disk/ds3000/index.html>

IBM System Storage DS5020 Express

Optimized data management requires storage solutions with high data availability, strong storage management capabilities and powerful performance features. IBM offers the IBM System Storage DS5020 Express, designed to provide lower total cost of ownership, high performance, robust functionality, and unparalleled ease of use. As part of the IBM DS series, the DS5020 Express offers:

- ▶ High-performance 8 Gbps capable Fibre Channel connections
- ▶ Optional 1 Gbps iSCSI interface
- ▶ Up to 112 TB of physical storage capacity with 112 1 TB SATA disk drives
- ▶ Powerful system management, data management, and data protection features

For more information, visit the following address:

<http://www.ibm.com/systems/storage/disk/ds5020/index.html>

IBM System Storage DS5000

DS5000 enhancements help reduce cost by reducing power per performance by introducing SSD drives. Also with the new EXP5060 expansion unit supporting 60 1 TB SATA drives in a 4U package, customers can see up to a one-third 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, using the 8 Gbps FC host ports. With DS5000, you get consistent performance

from a smarter design, that simplifies your infrastructure, improves your TCO, and reduces your cost.

For more information, visit the following address:

<http://www.ibm.com/systems/storage/disk/ds5000>

IBM XIV Storage System

IBM is introducing a mid-sized 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 capacity requirements can take advantage of the latest technology from IBM for their most demanding applications with as little as 27 TB of usable capacity and incremental upgrades.

For more information, visit the following address:

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

IBM System Storage DS8700

The IBM System Storage DS8700 is the most advanced model in the IBM DS8000® lineup and introduces new dual IBM POWER6-based controllers that usher in a new level of performance for the company's flagship enterprise disk platform. With overall performance improving up to over 2.5x, the new DS8700 is designed to support the most demanding business applications with its superior data throughput, unparalleled resiliency features and five-nines availability. In today's dynamic, global business environment, where organizations like yours need information to be reliably available around the clock and with minimal delay, can you really afford not to run your business on the DS8000 series? Moreover, with its tremendous scalability, flexible tiered storage options, broad server support, and support for advanced IBM deduplication technology, the DS8000 can help simplify the storage environment by consolidating multiple storage systems onto a single system, while providing the availability and performance you have come to trust for your most important business applications.

For more information, visit the following address:

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

2.13 Hardware Management Console

The Hardware Management Console (HMC) is a dedicated workstation that provides a graphical user interface (GUI) for configuring, operating, and performing basic system tasks for the POWER7 processor-based systems (and the POWER5, POWER5+, POWER6 and POWER6+ processor-based systems) that function in either non-partitioned, partitioned, or clustered environments. In addition, the HMC is used to configure and manage partitions. One HMC is capable of controlling multiple POWER5, POWER5+, POWER6 and POWER6+ and POWER7 processor-based systems.

At the time of writing, one HMC supports up to 1000 LPARs using the HMC machine code Version 7 Release 710. It can also support up to 48 Power 750, or 780 systems. Updates of the machine code and HMC functions, and hardware prerequisites, are at this address:

<https://www14.software.ibm.com/webapp/set2/sas/f/hmc/home.htm>

2.13.1 HMC Functional overview

The HMC provides three groups of functions: server, virtualization, and HMC management.

Server management

The first group contains all functions related to the management of the physical servers under the control of the HMC:

- ▶ System password
- ▶ Status Bar
- ▶ Power On/Off
- ▶ Capacity on Demand
- ▶ Error management
 - System indicators
 - Error and 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 functions related to virtualization features such as the partitions configuration or dynamic reconfiguration of resources:

- ▶ System Plans
- ▶ System Profiles
- ▶ Partitions (create, activate, shutdown)
- ▶ Profiles
- ▶ Partition Mobility
- ▶ DLPAR (processors, memory, I/O, etc.)
- ▶ Custom Groups

HMC Console 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 and reconnect
- ▶ Network Security
 - Remote operation enable and disable
 - User definable SSL certificates

- ▶ Console logging
- ▶ HMC Redundancy
- ▶ Scheduled Operations
- ▶ Back-up & Restore
- ▶ Updates, Upgrades
- ▶ Customizable Message of the day

The versions V7R710 of the HMC code adds the following functions to these families:

- ▶ Server Management
 - Support for POWER7 750/755 Server
 - Support for POWER7 770/780 Server (V7R710 SP1)
- ▶ Virtualization Management
 - Remove limit of 128 Active Memory Sharing Partitions
 - Increased limit of partitions managed by an HMC to 1024,
 - Active Memory Expansion
- ▶ Console Management
 - Increase Capacity on Demand Billing Capacity
 - Ongoing HMC Performance Improvements

The HMC provides both a graphical and command-line interface (CLI) for all management tasks. Remote connection to the HMC using a Web browser (as of HMC Version 7, previous versions required a special client program, called WebSM) or SSH are possible. The CLI is also available by using the Secure Shell (SSH) connection to the HMC. It can be used by an external management system or a partition to remotely perform HMC operations.

2.13.2 HMC connectivity to the POWER7 processor based systems

POWER5, POWER5+, POWER6, POWER6+, and POWER7 processor-technology based servers that are managed by an HMC require Ethernet connectivity between the HMC and the server Service Processor. In addition, if dynamic LPAR, Live Partition Mobility or PowerVM Active Memory Sharing operations are required on the managed partitions, Ethernet connectivity is needed between these partitions and the HMC. A minimum of two Ethernet ports are needed on the HMC to provide such connectivity. The rack-mounted 7042-CR5 HMC default configuration provides four Ethernet ports. The deskside 7042-C07 HMC standard configuration offers only one Ethernet port; be sure to order an optional PCI-e adapter to provide additional Ethernet ports.

For any logical partition in a server, a possibility is to use a Shared Ethernet Adapter set in Virtual I/O Server or Logical Ports of the Integrated Virtual Ethernet card, for a unique or fewer connections from the HMC to partitions. Therefore, a partition does not require its own physical adapter to communicate to an HMC.

A good practice is to connect the HMC to the first HMC port on the server, which is labeled as HMC Port 1, although other network configurations are possible. You can attach a second HMC to HMC Port 2 of the server for redundancy (or vice versa). Figure 2-29 on page 79 shows a simple network configuration to enable the connection from HMC to server and to enable Dynamic LPAR operations. For more details about HMC and the possible network connections, see *Hardware Management Console V7 Handbook*, SG24-7491.

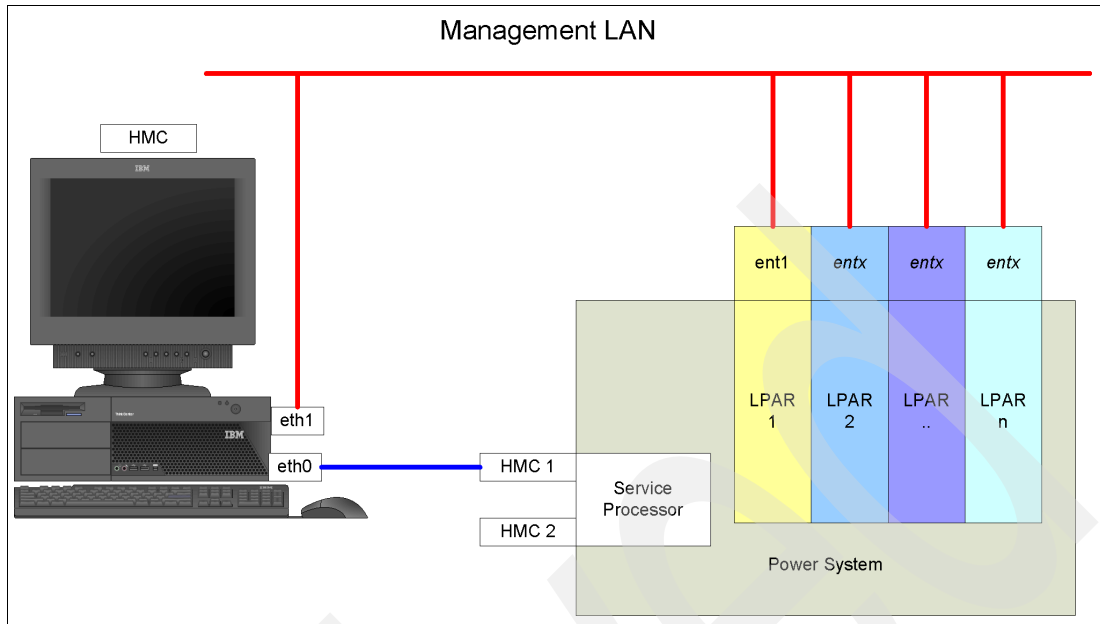


Figure 2-29 HMC to service processor and LPARs network connection

The default mechanism for allocation of the IP addresses for the service processor HMC ports is dynamic. The HMC can be configured as a DHCP server, providing the IP address at the time the managed server is powered on. In this case, the FSP are allocated IP address from a set of address ranges predefined in the HMC software. These predefined ranges are identical for version 710 of the HMC code and for previous versions.

If the service processor of the managed server does not receive a DHCP reply before time-out, predefined IP addresses will be set up on both ports. Static IP address allocation is also an option. You can also configure the IP address of the service processor ports with a static IP address by using the Advanced System Management Interface (ASMI) menus.

Note: 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 as connections. 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 options from a client Web browser, using the HTTP server integrated into the service processor internal operating system.
- ▶ When not configured otherwise (DHCP or from a previous ASMI setting), both Ethernet ports of the first FSP have predefined IP addresses:
 - Service processor Eth0 or HMC1 port is configured as 169.254.2.147 with netmask 255.255.255.0
 - Service processor Eth1 or HMC2 port is configured as 169.254.3.147 with netmask 255.255.255.0

For the second FSP of IBM Power 770 and 780, these default addresses are:

- Service processor Eth0 or HMC1 port is configured as 169.254.2.146 with netmask 255.255.255.0
- Service processor Eth1 or HMC2 port is configured as 169.254.3.146 with netmask 255.255.255.0

For more information about the service processor, see “Service processor” on page 148.

2.13.3 High availability using the HMC

The HMC is an important hardware component. When in operation, POWER7 processor-based servers and their hosted partitions can continue to operate when no HMC is available. However, in such conditions, certain operations cannot be performed such as a DLPAR reconfiguration, a partition migration using PowerVM Live Partition Mobility, or the creation of a new partition. You might therefore 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 redundant HMC function is desired, the servers can be attached to two separate HMCs to address availability requirements. Both HMCs must have the same level of 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 2-30 on page 81 shows one possible highly available HMC configuration managing two servers. These servers have only one CEC and therefore only one FSP. Each HMC is connected to one FSP port of all managed servers.

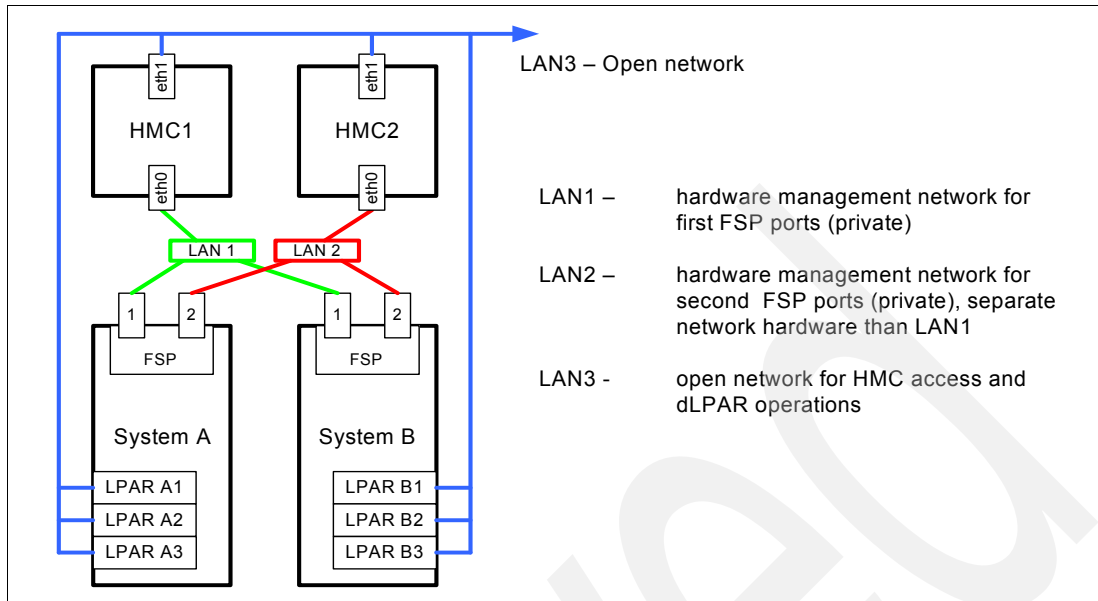


Figure 2-30 Highly available HMC and network architecture

Note that only hardware management networks (LAN1 and LAN2) are highly available (as shown in the figure) for simplicity. However, management network (LAN3) can be made highly available by using a similar concept and adding more Ethernet adapters to LPARs and HMCs

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

Redundant service processor connectivity

For POWER7 770 and 780 models with two or more CECs, two redundant service processors are installed in CEC enclosures 1 and 2. Redundant service processor function requires that each HMC be attached to both one Ethernet port in enclosure 1 and one Ethernet port in CEC enclosure 2.

Figure 2-31 on page 82 shows a redundant HMC and redundant service processor connectivity configuration.

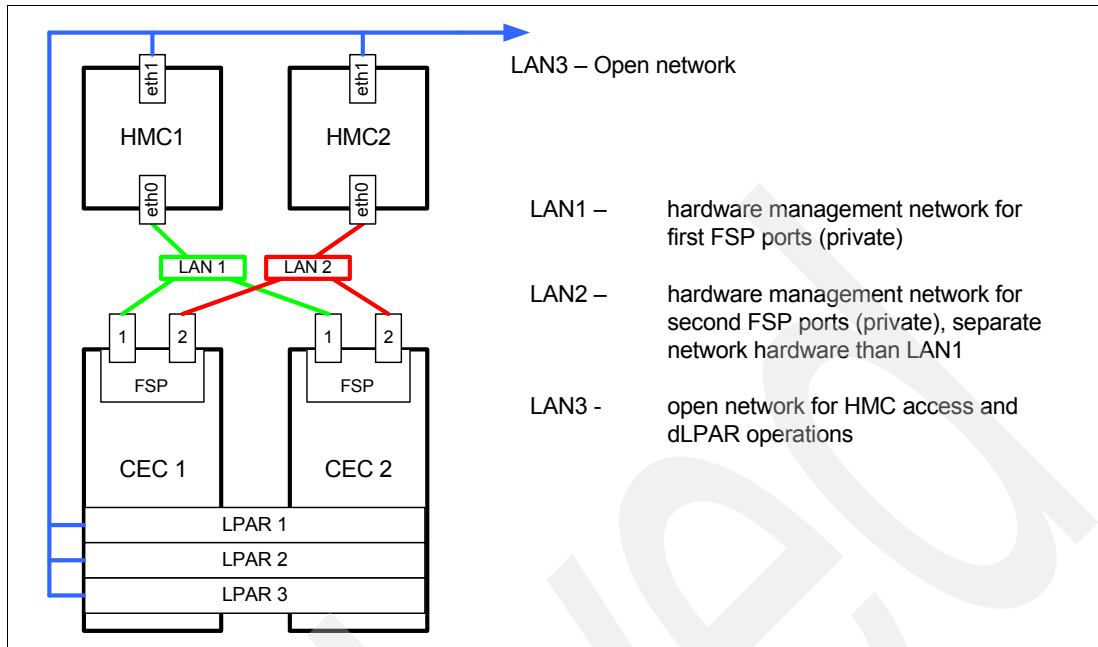


Figure 2-31 Redundant HMC connection and redundant service processor configuration

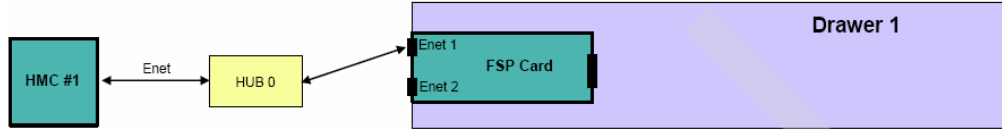
In a configuration with multiple systems or HMCs, the customer is required to provide switches or hubs to connect each HMC to the server FSP Ethernet ports in each system:

- ▶ One HMC should connect to the port labeled as HMC Port 1 on the first two CEC drawers of each system
- ▶ A second HMC should be attached to HMC Port 2 on the first two CEC drawers of each system.

This solution provides redundancy both for the HMCs and the service processors.

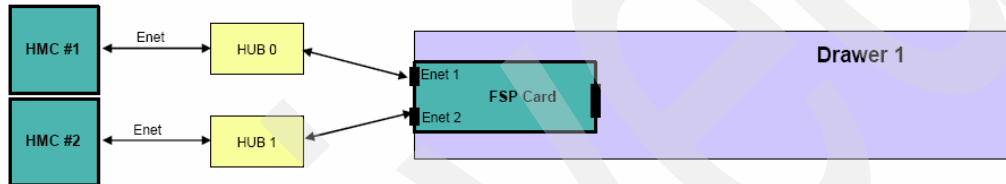
Figure 2-32 on page 83 describes the four possible Ethernet connectivity options between HMCs and FSPs.

Configuration #1 – Single Drawer and One HMCs



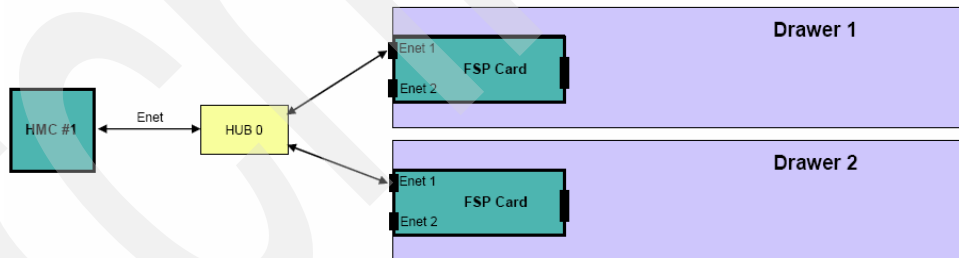
NOTE: HUB is optional.
Customer can have a direct connection to the FSP card

Configuration #2 – Single Drawer and Two HMCs



NOTE: HUBs are optional.

Configuration #3 – Multi-drawer with One HMCs



Configuration #4 – Multi-drawer with Two HMCs

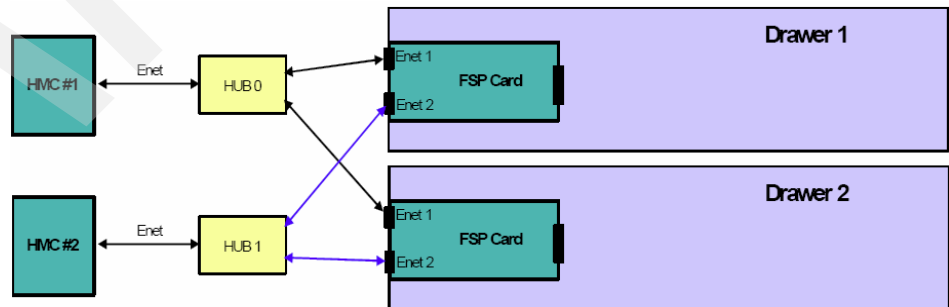


Figure 2-32 Summary of HMC to FSP configuration option depending on number of CEC

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

2.13.4 HMC code level

The HMC code must be at the following levels:

- ▶ V7R710 to support the POWER7 750 and 755 systems
- ▶ V7R710 SP1 to support the POWER7 770 and 780 systems

In a dual HMC configuration, both must be at the same version and release of the HMC.

Tips:

- ▶ When upgrading the code of a dual HMC configuration, a good practice is to disconnect one HMC to avoid having both HMCs connected to the same server but running different levels of code. If no profiles or partition changes take place during the upgrade, both HMCs can stay connected. If the HMCs are at different levels and a profile change is made from the HMC at level 7.10 for example, the format of the data stored in the server could be changed, causing the HMC at a previous level (for example 3.50) to possibly go into a recovery state, because it does not understand the new data format.
- ▶ Compatibility rules exist between the various software that is executing within a POWER7 processor-based server environment: HMC, VIO, system firmware, or partition operating systems. To check which combinations are supported, and to identify required upgrades, you may use the Fix Level Recommendation Tool Web page:

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

Two rules are related to HMC code level when you use 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.

2.14 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.

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

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

2.14.1 Virtual I/O Server

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>

2.14.2 IBM AIX operating system

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

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 Web site:

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

The Fix Central Web site also provides information about how to obtain the fixes shipping on CD-ROM.

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 Web site:

<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 minimum level of AIX Version 5.3 to support the Power 710, 720, 730, and 740 is:

- ▶ 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 minimum level of AIX Version 5.3 to support the Power 750, 755, 770, and 780 is:

- ▶ 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 minimum level of AIX Version 5.3 to support the Power 795 is:

- ▶ 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 using AIX Version 5.3 will be executing in POWER6 or POWER6+ compatibility mode. This means that 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 minimum level of AIX Version 6.1 to support the Power 710, 720, 730, 740, and 795 is:

- ▶ 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 minimum level of AIX Version 6.1 to support the Power 750 and 755 is:

- ▶ 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 minimum level of AIX Version 6.1 to support the Power 770 and 780 is:

- ▶ 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 using AIX 6.1 with TL6 can run in POWER6, POWER6+ or POWER7 mode. It is best to run the partition in POWER7 mode to allow exploitation of new hardware capabilities such as SMT4 and Active Memory Expansion (AME).

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

IBM AIX Version 7.1

AIX Version 7.1 comes with full support for the Power 710, 720, 730, 740, 750, 755, 770, 780, and 795, exploiting all the hardware features from the POWER7 processor, as well as 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.

Note: Partition sizes greater than 128-cores (up to 256-cores) will require a software key to enable. Purchase will require lab services pre-analysis as a prerequisite to shipment. The software key requires feature #1256 to be installed.

2.14.3 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 Web site:

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

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

- ▶ 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 wanting to configure Linux partitions in virtualized Power Systems have to be aware of these 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.

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

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

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

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

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

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

Supported virtualization features are listed in 3.4.9, "Operating System support for PowerVM".

2.15 Compiler technology

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 will have advanced a more than 20 year investment in the XL compilers for POWER series and PowerPC® series architectures.

XL C, XL C/C++ and XL Fortran features introduced to exploit the latest POWER7 processor include 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, and architecture and tune compiler options to optimize and tune your applications.

COBOL for AIX and PL/I for AIX support application development on the latest POWER7 processor.

IBM Rational® Development Studio for IBM i 7.1 provides programming languages for creating modern business applications. This includes the ILE RPG, ILE COBOL, C, and C++ compilers as well as 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 has also released a new product called Rational Open Access: RPG Edition. This opens up the ILE RPG file I/O processing, enabling partners, tool providers, and users to write custom I/O handlers that can access other devices like databases, services, and Web user interfaces.

IBM Rational Developer for Power Systems Software provides a rich set of integrated development tools that support the XL C/C++ for AIX compiler, the XL C for AIX compiler and the COBOL for AIX compiler. Rational Developer for Power Systems Software offers capabilities of file management, searching, editing, analysis, build, and debug, all integrated into an Eclipse workbench. XL C/C++, XL C and COBOL for AIX developers can boost productivity by moving from older, text-based, command line development tools to a rich set of integrated development tools.

2.16 Energy management

The Power 770 and 780 is adding support for EnergyScale Power and Thermal Management. The IBM Systems Director Active Energy Manager exploits EnergyScale technology, enabling advanced energy management features to dramatically and dynamically conserve power and further improve energy efficiency. Intelligent Energy optimization capabilities enable the POWER7 processor to operate at a higher frequency for increased performance and performance per watt or dramatically reduce frequency to save energy.

2.16.1 IBM EnergyScale technology

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

This section describes IBM EnergyScale design features, and hardware and software requirements.

IBM EnergyScale consists of:

- ▶ A built-in Thermal Power Management Device (TPMD) or TPMD card
- ▶ Power executive software: IBM Systems Director Active Energy Manager, an IBM Systems Directors plug-in

IBM EnergyScale functions include:

- ▶ Energy trending

EnergyScale provides continuous collection of real-time server energy consumption. This enables administrators to predict power consumption within their infrastructure and to react to business and processing needs. For example, administrators may 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 require 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 50% down from nominal. Power Saver Mode is not supported during boot or reboot operations although it is a persistent condition that will be sustained after booting, 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. You configure this setting from IBM Director Active Energy Manager. 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 usage. When a system is idle, the system firmware lowers the frequency and

voltage to Power Energy Saver Mode values. When fully utilized, the maximum frequency can 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 at least 100% of nominal. Dynamic Power Saver Mode is mutually exclusive with Power Saver mode: only one of these modes may be enabled at a given time.

- ▶ **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 should never be reached but frees up margined power in the data center. The margined power is the amount of extra power that is allocated to a server during its installation in a data center. It is based on the server environmental specifications that usually are never reached 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**

The two power ranges into which the power cap may be set are: Power Capping (described previously) and Soft Power Capping. 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, Soft Power Capping is the mechanism to use.

- ▶ **Processor Core Nap**

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 falls within a partition dispatch (context switch) so that the POWER Hypervisor can use it 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. Nap mode allows the hardware to clock-off most of the 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. Unlicensed cores are kept in core Nap until they are licensed and return to core Nap when they are unlicensed again.

- ▶ **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 can vary based on power consumption, ambient temperature, and altitude available. System altitude may be set in IBM Director Active Energy Manager. If no altitude is set, the system assumes 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 longer.

- ▶ EnergyScale for I/O

IBM POWER processor-based systems automatically power off 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 POWER processor-based servers, and the expansion units that they support.

2.16.2 Thermal power management device (TPMD) card

The TPMD card is part of the energy management of performance and thermal proposal, which dynamically optimizes the processor performance depending on processor power and system workload.

The IBM POWER7 chip is a significant improvement in power and performance over the IBM POWER6 chip. POWER7 has more internal hardware and power and thermal management functions to interact with:

- ▶ More hardware eight (8) cores versus two (2) cores, four (4) threads versus two (2) threads per core, asynchronous processor core chiplet
- ▶ Advanced Idle Power Management functions at chiplet level
- ▶ Advanced Dynamic Power Management functions (DPM) in all units in hardware (processor cores, processor core chiplet, chip-level nest unit level, and chip level)
- ▶ Advanced Actuators/Control
- ▶ Advanced Accelerators

Thus, the new TPMD card has a more powerful microcontroller, more A/D channels and more busses to handle the increase workload, link traffic, and new power and thermal functions.

Archived

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. Key capabilities include:

- ▶ 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 processor 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 105).
- ▶ Saves and restores all processor state information during a logical processor context switch.
- ▶ Controls hardware I/O interrupt management facilities for logical partitions.
- ▶ Provides virtual LAN channels between logical partitions that help to reduce the need for physical Ethernet adapters for inter-partition communication.
- ▶ Monitors the Service Processor and will perform a reset 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 HMC. It requires memory to support the resource assignment to the logical partitions on the server. The amount of memory required by the POWER Hypervisor firmware varies according to several factors. Factors influencing the POWER Hypervisor memory requirements include:

- ▶ 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 to create a partition is the size of the system's Logical Memory Block (LMB). The default LMB size varies according to the amount of memory configured in the CEC as shown in Table 3-1.

Table 3-1 Configured CEC memory-to-default Logical Memory Block size

Configurable CEC memory	Default Logical Memory Block
Greater than 8 GB up to 16 GB	64 MB
Greater than 16 GB up to 32 GB	128 MB
Greater than 32 GB	256 MB

In most cases, 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. The storage virtualization is accomplished using two, paired, adapters: a virtual SCSI server adapter and 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” on page 111. The Virtual I/O Server software is available with the optional PowerVM Edition features.

Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions on the same server to use a fast and secure communication without any need for physical interconnection. The virtual Ethernet allows a transmission speed in the range of 1 - 3 Gbps, depending on the maximum transmission unit (MTU) size and CPU entitlement. Virtual Ethernet support starts with IBM AIX Version 5.3, or an appropriate level of Linux supporting virtual Ethernet devices (see 3.4.9, “Operating System support for PowerVM” on page 119). The virtual Ethernet is part of the base system configuration.

Virtual Ethernet has the following major features:

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

Any virtual Ethernet can also have 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” on page 111 for more details about shared Ethernet), also known as Shared Ethernet Adapter.

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

Virtual 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 on page 96 depicts the connections between the client partition virtual Fibre Channel adapters and the external storage. For additional information, see 3.4.8, “NPIV” on page 118.

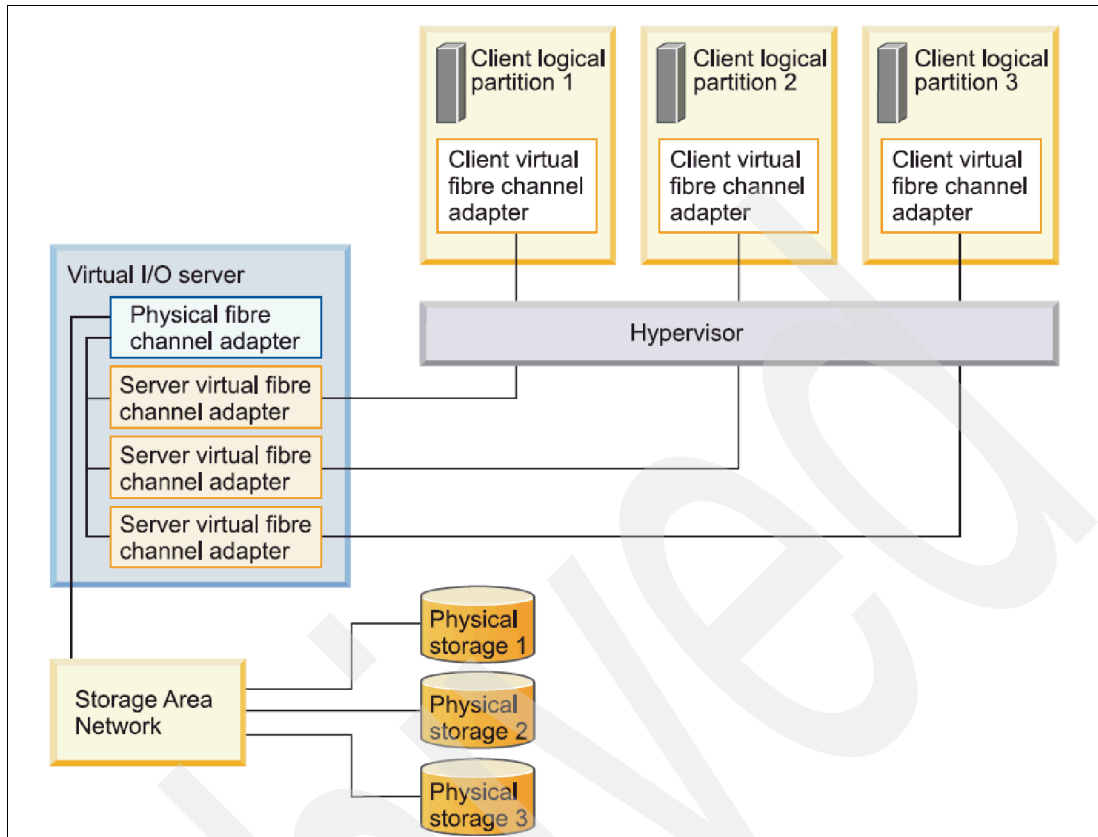


Figure 3-1 Connectivity between virtual Fibre Channels adapters and external SAN devices

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, strictly speaking, not a virtualization feature, the POWER modes are described here because they affect various virtualization features.

On a Power System servers, partitions can be configured to run in several modes, including:

- ▶ 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://www.power.org/resources/reading/PowerISA_V2.05.pdf

► POWER6+ compatibility mode

This mode is similar to POWER6, with 8 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://www.power.org/resources/downloads/PowerISA_V2.06_PUBLIC.pdf

The selection of the mode is made on a per partition basis, from the HMC, by editing the partition profile as presented in Figure 3-2.

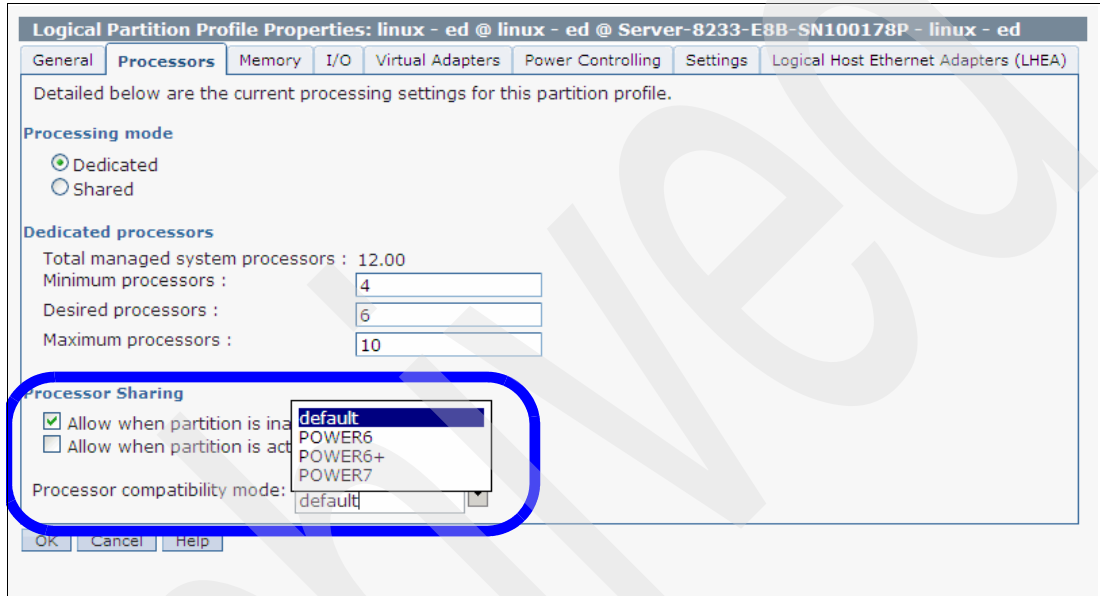


Figure 3-2 Configuring partition profile compatibility mode from the HMC

Table 3-2 lists the differences between these modes.

Table 3-2 Differences between POWER6 and POWER7 mode

POWER6 mode (and POWER6+)	POWER7 mode	Customer value
2-thread SMT	4-thread SMT	Throughput performance, processor core utilization
VMX (Vector Multimedia Extension / AltiVec)	VSX (Vector Scalar Extension)	High performance computing
Affinity OFF by default	3-tier memory, Micropartition Affinity	Improved system performance for system images spanning sockets and nodes
<ul style="list-style-type: none"> ▶ Barrier Synchronization ▶ Fixed 128-byte Array; Kernel Extension Access 	<ul style="list-style-type: none"> ▶ Enhanced Barrier Synchronization ▶ Variable Sized Array; User Shared Memory Access 	High performance computing parallel programming synchronization facility
<ul style="list-style-type: none"> ▶ 64-core and 128-thread scaling 	<ul style="list-style-type: none"> ▶ 32-core and 128-thread scaling ▶ 64-core and 256-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, 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 is an optional feature of POWER7 processor-based servers that must be specified when creating the configuration in the e-Config tool, as follows:

- IBM Power 750** #4792
- IBM Power 770** #4791
- IBM Power 780** #4791

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 an innovative POWER7 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.

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 per additional CPU utilized. Other test workloads had more modest results.

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 corresponds to a partition that has spare processing power capacity; the blue curve corresponds to a partition constrained in processing power.

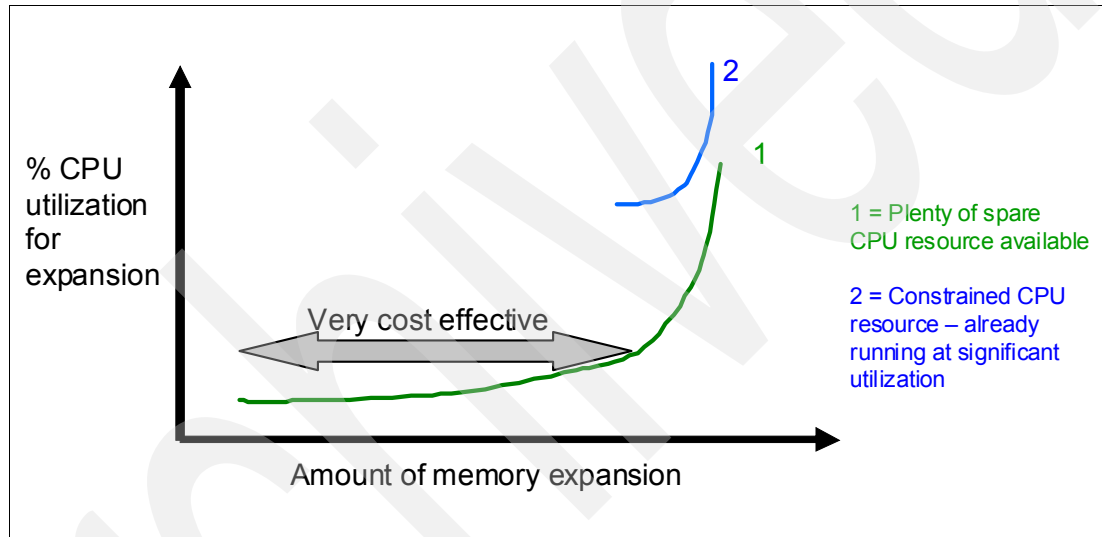


Figure 3-3 CPU usage versus memory expansion effectiveness

Both cases show that there is a knee-of-curve relationship for CPU resource required for memory expansion:

- ▶ Busy processor cores do not have resources to spare for expansion.
- ▶ The more memory expansion done, the more CPU resource is required.

The knee varies depending on how compressible that 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 perform 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 model Power System can run the planning tool. Figure 3-4 on page 100 shows an example of the output returned by this planning tool. The tool outputs various real memory and CPU resource combinations to achieve the desired effective memory. It also recommends one particular combination. In this example, the tool recommends that you allocate 58% of a processor, to benefit from 45% extra memory capacity.

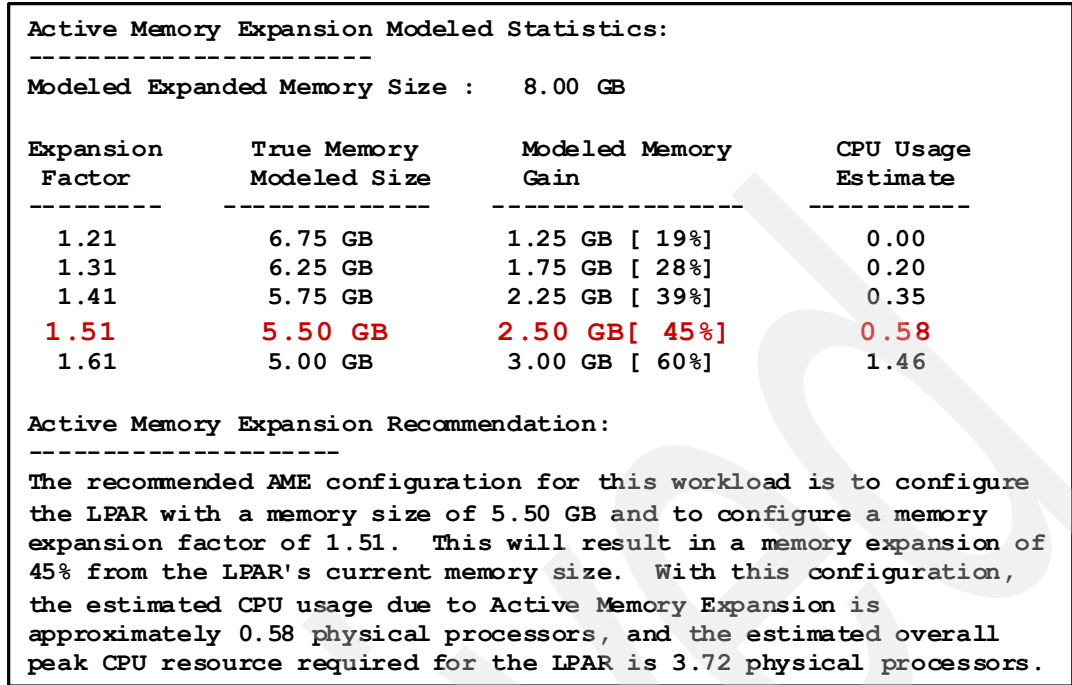


Figure 3-4 Output from Active Memory Expansion planning tool

After you select the value of the memory expansion factor you want to achieve, you can use this value to configure the partition from the HMC, as shown in Figure 3-5.

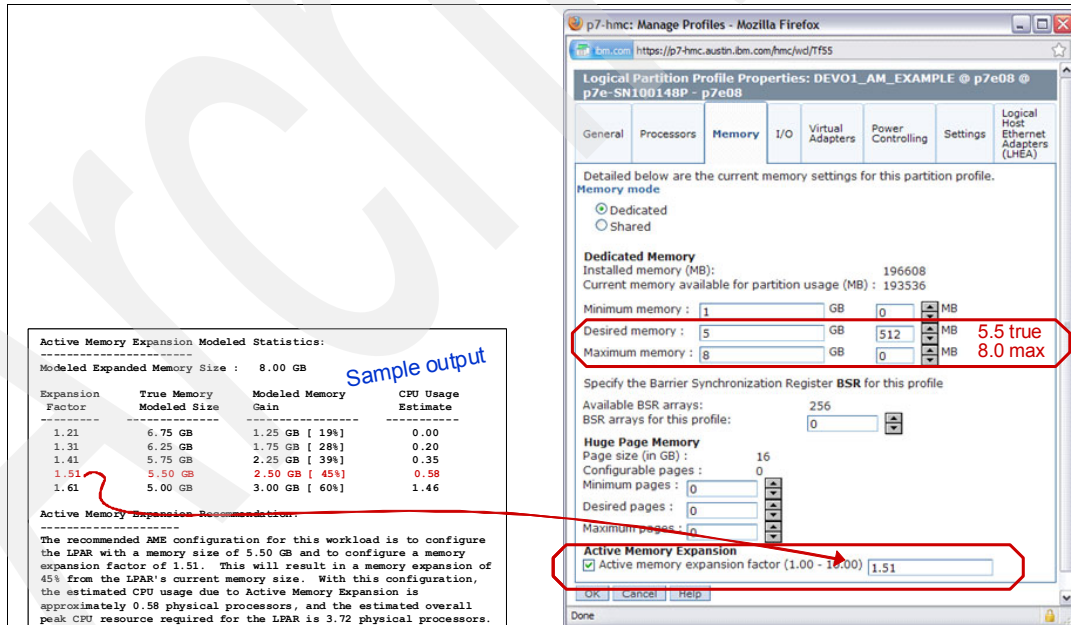


Figure 3-5 Using the planning tool result to configure the partition

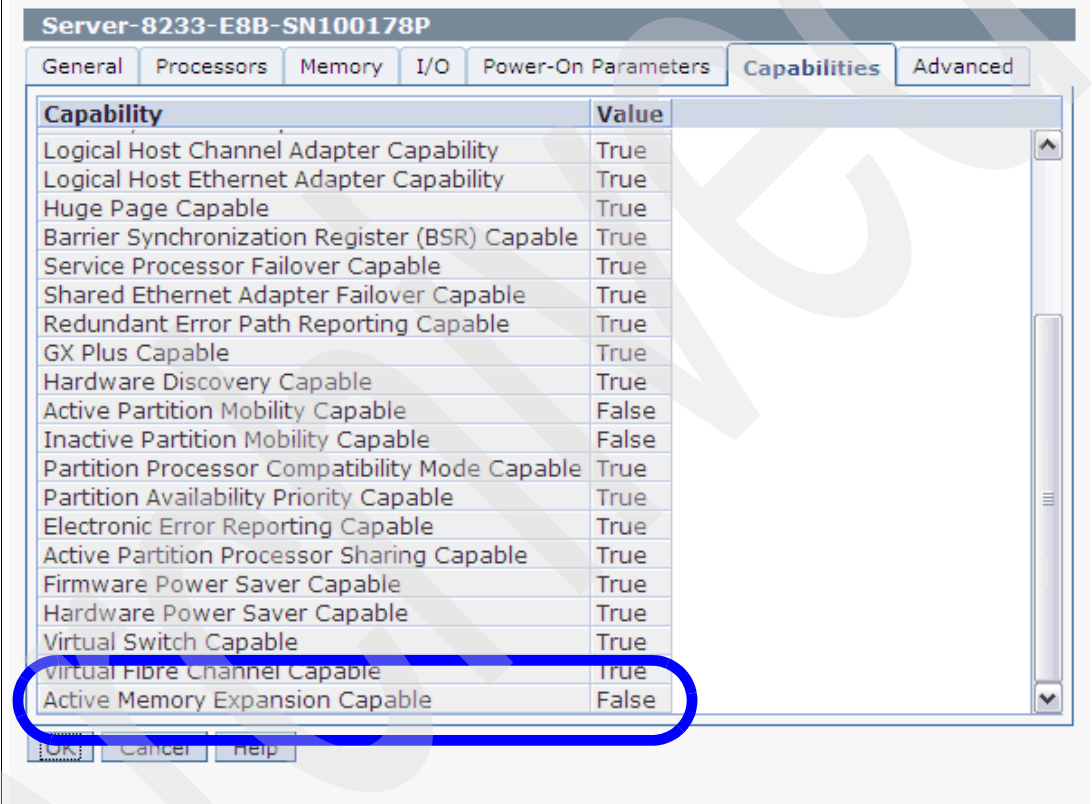
On the HMC menu describing the partition, check the **Active Memory Expansion** box and enter true and maximum memory, and the memory expansion factor. To turn off expansion, clear the check box. In both case, a reboot of the partition is needed to activate the change.

In addition, a 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 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. Rebooting is not required to enable the physical server. 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 partition using memory expansion.

From the HMC, you may view whether the Active Memory Expansion feature has been activated, as shown in Figure 3-6.



Capability	Value
Logical Host Channel Adapter Capability	True
Logical Host Ethernet Adapter Capability	True
Huge Page Capable	True
Barrier Synchronization Register (BSR) Capable	True
Service Processor Failover Capable	True
Shared Ethernet Adapter Failover Capable	True
Redundant Error Path Reporting Capable	True
GX Plus Capable	True
Hardware Discovery Capable	True
Active Partition Mobility Capable	False
Inactive Partition Mobility Capable	False
Partition Processor Compatibility Mode Capable	True
Partition Availability Priority Capable	True
Electronic Error Reporting Capable	True
Active Partition Processor Sharing Capable	True
Firmware Power Saver Capable	True
Hardware Power Saver Capable	True
Virtual Switch Capable	True
Virtual Fibre Channel Capable	True
Active Memory Expansion Capable	False

Figure 3-6 Server capabilities listed from the HMC

Note: To move by using Live Partition Mobility to an LPAR using Active Memory Expansion to a different system, the target system must support AME (the target system must have AME activated with the software key). If the target system does not have AME activated, the mobility operation will fail during the pre-mobility check phase, and an appropriate error message will be displayed to the user.

For detailed information regarding Active Memory Expansion, you can download the document *Active Memory Expansion: Overview and Usage Guide* from this location:

http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&appname=S TGE_PO_PO_USEN&htmlfid=POW03037USEN

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 Power Systems Virtualization (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. Section 3.4.1, “PowerVM editions” on page 102 discusses the licensed features of each of the three separate editions of PowerVM.

3.4.1 PowerVM editions

This section provides information about the virtualization capabilities of the PowerVM. The three editions of PowerVM are suited for various purposes, as follows:

- ▶ **PowerVM Express Edition**
This edition is intended for evaluations, pilots, proof of concepts, generally in single-server projects.
- ▶ **PowerVM Standard Edition**
This edition 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

Table 3-3 lists the version of PowerVM that are available on each model of POWER7 processor-based servers.

Table 3-3 Availability of PowerVM per POWER7 processor technology based server model

PowerVM editions	Express	Standard	Enterprise
IBM Power 750	#7793	#7794	#7795
IBM Power 755	No	No	No
IBM Power 770	No	#7942	#7995
IBM Power 780	No	#7942	#7995

Upgrading from the Express Edition to the Standard or Enterprise Edition, and from Standard to Enterprise Editions, are possible. Table 3-4 on page 103, outlines the functional elements of the three PowerVM editions.

Table 3-4 PowerVM capabilities

PowerVM editions	Express	Standard	Enterprise
Micro-partitions	Yes	Yes	Yes
Maximum LPARs	1+2 per server	10/core	10/core
Management	VMcontrol IVM	VMcontrol IVM, HMC	VMcontrol IVM, HMC
Virtual IO Server	Yes	Yes	Yes
NPIV	Yes	Yes	Yes
Multiple Shared Processor Pool	No	Yes	Yes
Live Partition Mobility	No	No	Yes
Active Memory Sharing	No	No	Yes

Note The IBM Power 770 and 780 have to be managed with 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.

Dynamic logical partitioning

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

Later, dynamic 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 are executing. AIX Version 5.2, with all the necessary enhancements to enable dynamic LPAR, was introduced in 2002. The ability to reconfigure dynamic LPARs encourages system administrators to dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

Micro-Partitioning

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 using fractions of processors is also known as a Shared Processor Partition or micro-partition. Micro-partitions run over a set of processors called Shared Processor Pool. And virtual processors are used to let the operating system manage the fractions of processing power 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 have to be defined:

- ▶ The minimum, desired, and maximum processing units
Processing units are defined as processing power, or the fraction of time 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 using the Shared Processor Pool Management window. Shared processor partitions use the default shared processor pool called DefaultPool by default. See 3.4.3, “Multiple Shared-Processor Pools” on page 106 for details about Multiple Shared Processor Pools.
- ▶ Whether the partition will or will not be able to access extra processing power to “fill up” its virtual processors above its capacity entitlement (selecting either to cap or uncapped your partition)
If there is spare processing power 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’s 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 units desired value but can exceed that value in the case of an uncapped partition and up to the number of virtual processors allocated.

A partition can be defined with a processor capacity as small as 0.10 processing units. This represents 0.10 of a physical processor. Each physical processor can be shared by up to 10 shared processor partitions and the partition’s entitlement can be incremented fractionally by as little as 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 or Integrated Virtualization Management.

This IBM Power 750 system can be configured with up to 32 cores, and the IBM Power 770 and 780 servers up to 64 cores. At the time of writing, these systems can support up to one of the following maximums:

- ▶ 32 (Power 770) and 64 (Power 780) dedicated partitions
- ▶ Up to 160 micro-partitions

An important point is that the maximums stated are supported by the hardware, but the practical limits depend on the application workload demands

Note: IBM plans for PowerVM to support up to 320 logical partitions on the Power 750 server and up to 640 logical partitions on the Power 770 and 780 servers. For future POWER7 systems, IBM plans for PowerVM to support up to 1,000 logical partitions per server.

Additional information about virtual processors includes:

- ▶ A virtual processor can be running (dispatched) either on a physical processor or as standby waiting for a physical processor to become available.
- ▶ Virtual processors do not introduce any additional abstraction level; they really are only a dispatch entity. When running on a physical processor, virtual processors run at the same speed as the physical processor.
- ▶ Each partition's profile defines CPU entitlement that determines how much processing power any given partition should receive. The total sum of CPU entitlement of all partitions cannot exceed the number of available physical processors in a shared processor pool.
- ▶ 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. Figure 3-7 shows a diagram of the concepts discussed in this section.

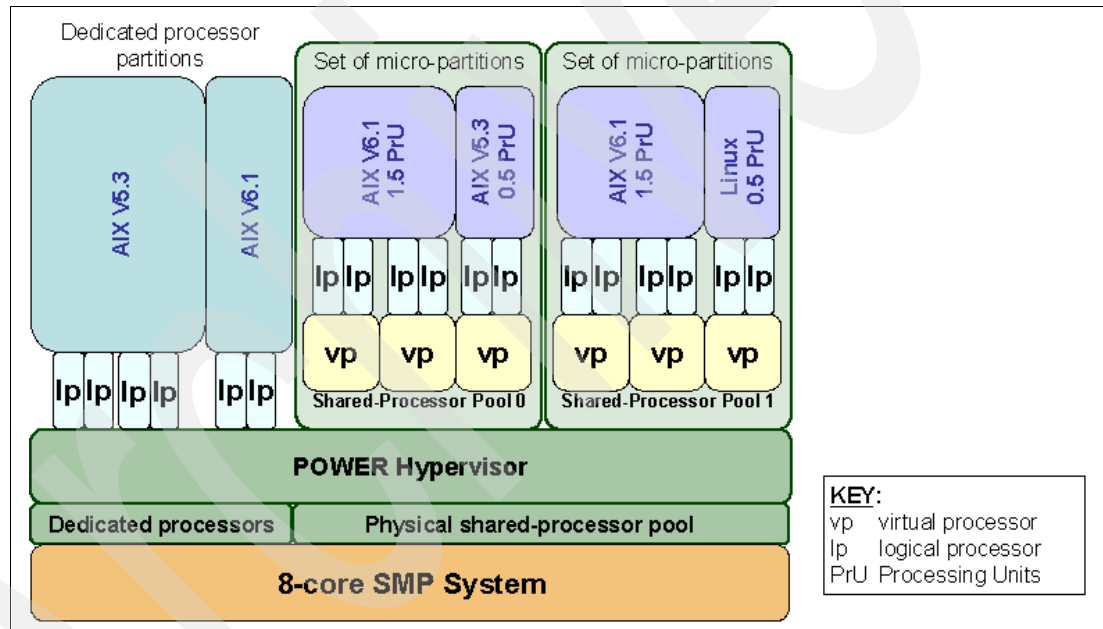


Figure 3-7 Logical partitioning concepts

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). 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 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 may 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 (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 partition's processing units entitlement. One processing unit represents one physical processor's processing capacity. At the end of the POWER Hypervisor's dispatch cycle (10 ms), all partitions should receive total CPU time equal to their processing units entitlement. The logical processors are defined on top of virtual processors. So, even with a virtual processor, the concept of logical processor exists and the number of logical processor depends whether the simultaneous multithreading is turned on or off.

3.4.3 Multiple Shared-Processor Pools

Multiple Shared-Processor Pools (MSPPs) is a capability supported on POWER7 processor and POWER6 processor based servers. This capability allows a system administrator to create a set of micro-partitions with the purpose of controlling the processor capacity that can be consumed from the physical shared-processor pool.

To implement MSPPs, there is a set of underlying techniques and technologies. An overview of the architecture of Multiple Shared-Processor Pools can be seen in Figure 3-8 on page 106.

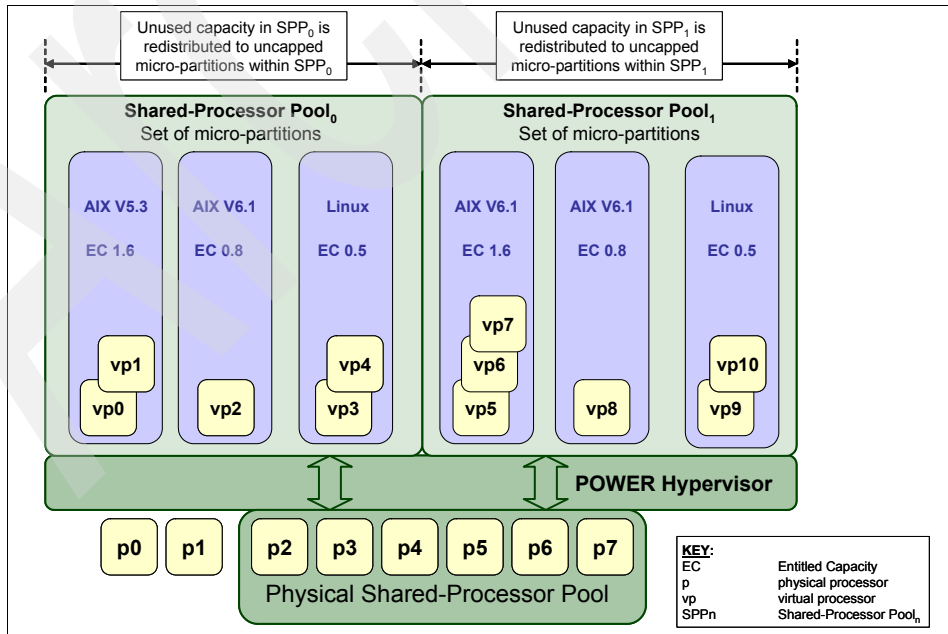


Figure 3-8 Overview of the architecture of Multiple Shared-Processor Pools

Micro-partitions are created and then identified as members of either the default Shared-Processor Pool₀ or a user-defined Shared-Processor Pool_n. The virtual processors that exist within the set of micro-partitions 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 micro-partition 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 micro-partition is uncapped.

If some micro-partitions in a Shared-Processor Pool do not use their capacity entitlement, the unused capacity is ceded and other uncapped micro-partitions 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 micro-partitions 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 (SPP₀)

On any Power Systems server supporting 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 SPP₀. 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. See Table 3-5 on page 108.

Table 3-5 Attribute values for the default Shared-Processor Pool (SPP₀)

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 micro-partitions in the default Shared-Processor Pool.

Creating Multiple Shared-Processor Pools

The default Shared-Processor Pool (SPP₀) 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 micro-partitions (either transferred from SPP₀ or newly created).

Levels of processor capacity resolution

The two levels of processor capacity resolution implemented by the POWER Hypervisor and Multiple Shared-Processor Pools are:

► Level₀

The first level, Level₀, 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 micro-partition within the same Shared-Processor Pool.

► Level₁

This is the second level of processor capacity resolution. When all Level₀ capacity has been resolved within the Multiple Shared-Processor Pools, the POWER Hypervisor harvests unused processor cycles and redistributes them to eligible micro-partitions regardless of the Multiple Shared-Processor Pools structure.

You can see the two levels of unused capacity redistribution implemented by the POWER Hypervisor in Figure 3-9 on page 109.

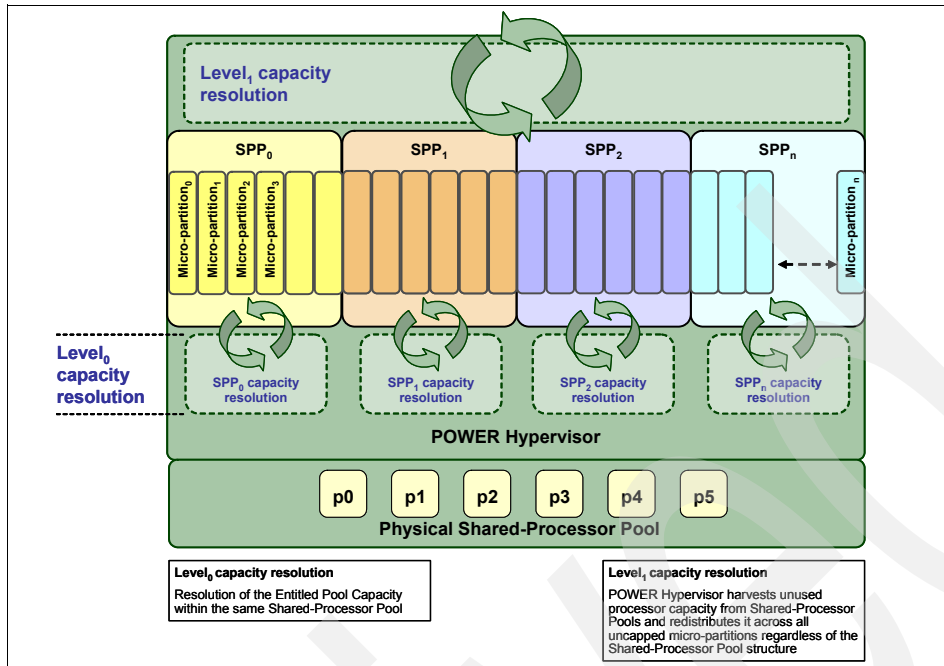


Figure 3-9 The two levels of unused capacity redistribution

Capacity allocation above the Entitled Pool Capacity (Level₁)

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 micro-partitions within the same Shared-Processor Pool. This level of processor capacity management is sometimes referred to as Level₀ 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 micro-partitions within it require additional processor capacity (above the Entitled Pool Capacity) then the POWER Hypervisor redistributes some of the extra capacity to the uncapped micro-partitions. This level of processor capacity management is sometimes referred to as Level₁ capacity resolution.

To redistribute unused processor capacity to uncapped micro-partitions in Multiple Shared-Processor Pools above the Entitled Pool Capacity, the POWER Hypervisor uses a higher level of redistribution, Level₁.

Important: 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 micro-partitions 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 micro-partitions within the Shared-Processor Pools cede the capacity to the POWER Hypervisor.

In busy Shared-Processor Pools, where the micro-partitions have used all of the Entitled Pool Capacity, the POWER Hypervisor allocates additional cycles to micro-partitions, in which *all* of the following statements are true:

- ▶ The Maximum Pool Capacity of the Shared-Processor Pool hosting the micro-partition has not been met.
- ▶ The micro-partition is uncapped.
- ▶ The micro-partition has enough virtual-processors to take advantage of the additional capacity.

Under these circumstances, the POWER Hypervisor allocates additional processor capacity to micro-partitions on the basis of their uncapped weights independent of the Shared-Processor Pool hosting the micro-partitions. This can be referred to as Level₁ 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 micro-partitions in the system into account, regardless of the Multiple Shared-Processor Pools structure.

Dynamic adjustment of Maximum Pool Capacity

The Maximum Pool Capacity of a Shared-Processor Pool, other than the default Shared-Processor Pool₀, can be adjusted dynamically from the HMC, using either the graphical or command-line interface (CLI).

Dynamic adjustment of Reserve Pool Capacity

The Reserved Pool Capacity of a Shared-Processor Pool, other than the default Shared-Processor Pool₀, can be adjusted dynamically from the HMC, using either the graphical or CLI interface.

Dynamic movement between Shared-Processor Pools

A micro-partition can be moved dynamically from one Shared-Processor Pool to another using the HMC using either the graphical or CLI interface. Because the Entitled Pool Capacity is partly made up of the sum of the entitled capacities of the micro-partitions, removing a micro-partition 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 micro-partition 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 will still exist but will not be active. Use the HMC interface to deactivate a Shared-Processor Pool. A Shared-Processor Pool cannot be deactivated unless all micro-partitions hosted by the Shared-Processor Pool have been removed.

Live Partition Mobility and Multiple Shared-Processor Pools

A micro-partition may leave a Shared-Processor Pool because of PowerVM Live Partition Mobility. Similarly, a micro-partition may 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 micro-partition.

Because several simultaneous micro-partition migrations are supported by PowerVM Live Partition Mobility, it is conceivable to migrate the entire Shared-Processor Pool from one server to another.

3.4.4 Virtual I/O Server

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.

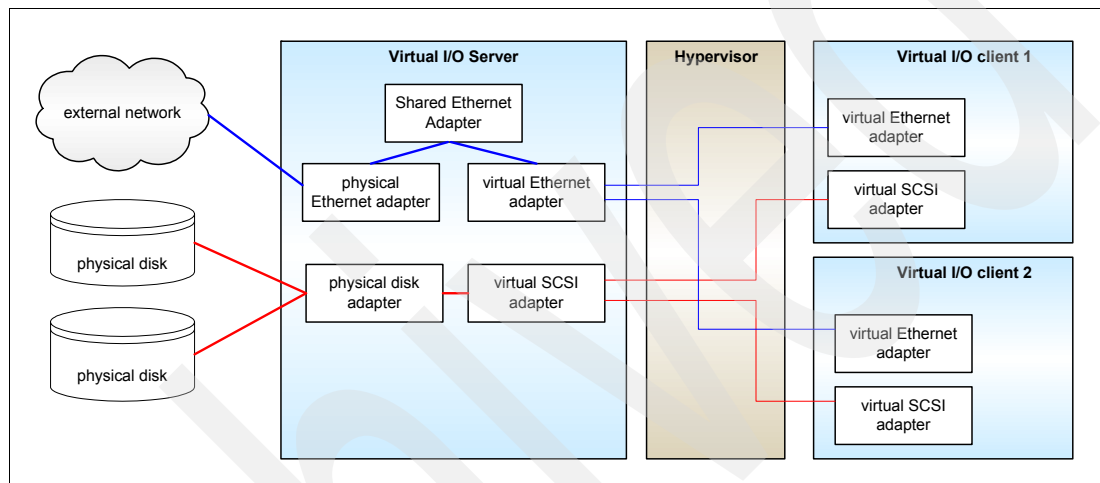


Figure 3-10 Architectural view of the Virtual I/O Server

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 only supported in special Virtual I/O Server partitions. Three major virtual devices are supported by the Virtual I/O Server: a Shared Ethernet Adapter, Virtual SCSI, and Virtual Fibre Channel adapter. The Virtual Fibre Channel adapter is used with the NPIV feature, described in 3.4.8, “NPIV” on page 118.

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 an SEA, you can connect internal and external VLANs using a physical adapter. The Shared Ethernet Adapter service can only be hosted 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 an SEA with one physical and two virtual Ethernet adapters. An SEA can include up to 16 virtual Ethernet adapters on the Virtual I/O Server that share the same physical access.

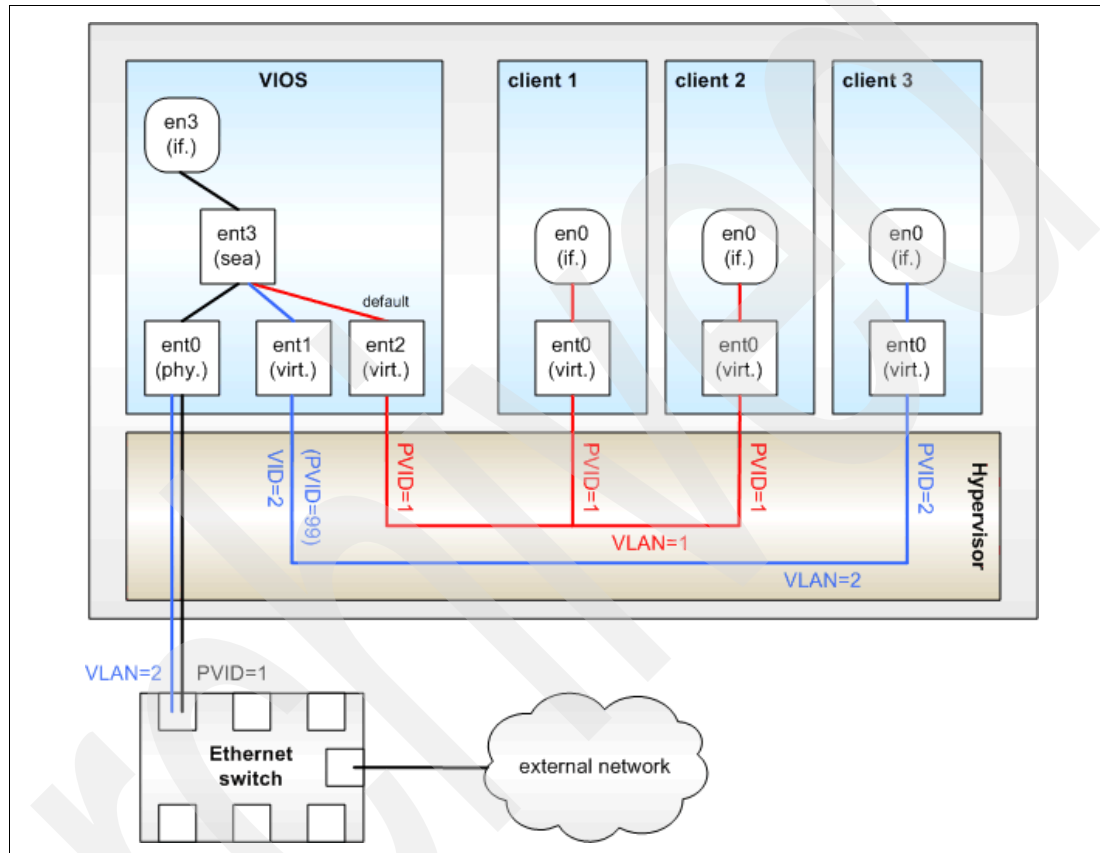


Figure 3-11 Architectural view of a Shared Ethernet Adapter

A single SEA setup can have up to 16 Virtual Ethernet trunk adapters and each virtual Ethernet trunk adapter can support up to 20 VLAN networks. Therefore, 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 an SEA.

Note: A Shared Ethernet Adapter does not need to have an IP address configured 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 SEA device, or on an additional virtual Ethernet adapter in the Virtual I/O Server. This leaves the SEA without the IP address, allowing for maintenance on the SEA without losing IP connectivity in case SEA failover is configured.

For a more detailed discussion about virtual networking, see:

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

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 server or, in SCSI terms, 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 using an HMC or through the Integrated Virtualization Manager on smaller systems. The virtual SCSI server (target) adapter is responsible for executing any SCSI commands it receives. It is owned by the Virtual I/O Server partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the Virtual I/O Server.

Physical disks presented to the Virtual I/O Server can be exported and assigned to a client partition in a number of 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 on page 114 shows an example where one physical disk is divided into two logical volumes by the Virtual I/O Server. Each of the two client partitions is assigned one logical volume, which is then accessed through a virtual I/O adapter (VSCSI Client Adapter). Inside the partition, the disk is seen as a normal hdisk.

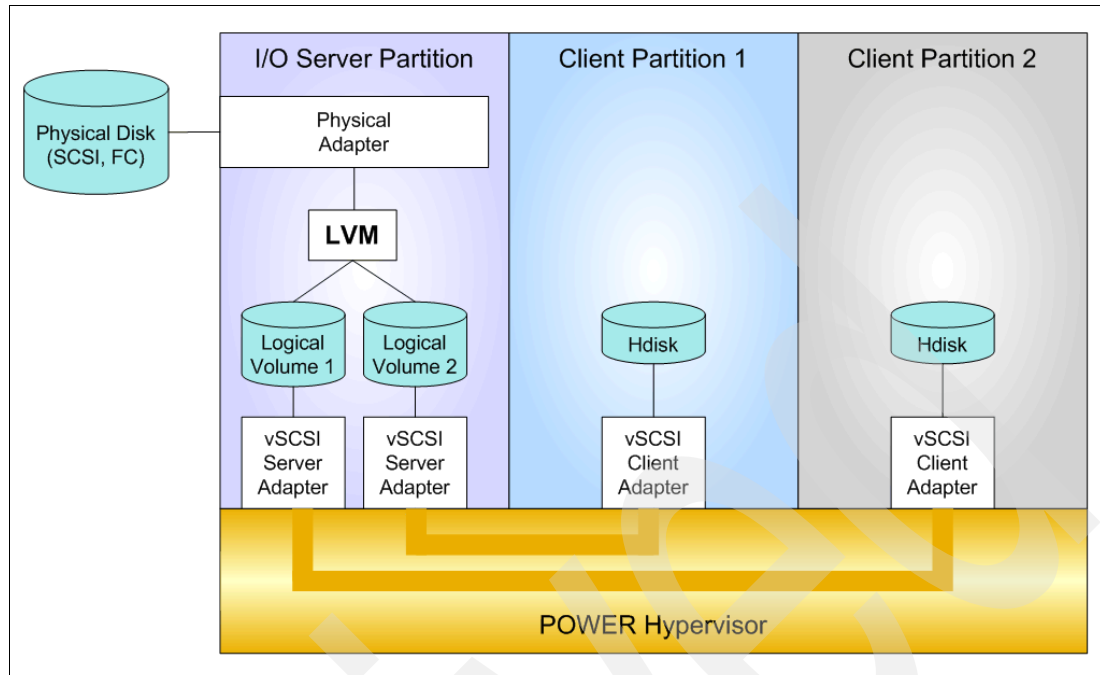


Figure 3-12 Architectural view of virtual SCSI

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:

<http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html>

Virtual I/O Server functions

Virtual I/O Server has a number of features, including monitoring solutions:

- ▶ Support for Live Partition Mobility on POWER6 processor-based systems with the PowerVM Enterprise Edition. For more information about Live Partition Mobility, see 3.4.6, “PowerVM Live Partition Mobility” on page 115.
- ▶ Support for virtual SCSI devices backed by a file, which 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), SNMP v3 (Simple Network Management Protocol) and LDAP (Lightweight Directory Access Protocol client functionality)
- ▶ System Planning Tool (SPT) and Workload Estimator, which are designed to ease the deployment of a virtualized infrastructure. For more information about System Planning Tool, see 3.5, “System Planning Tool” on page 121.
- ▶ Includes IBM Systems Director and a number of pre-installed Tivoli® agents, such as: Tivoli Identity Manager (TIM), to allow easy integration into an existing Tivoli Systems Management infrastructure; Tivoli Application Dependency Discovery Manager (ADDM), which creates and maintains automatically 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 Lx86

Note: IBM plans for PowerVM Lx86 to support POWER7 systems in second quarter 2010.

3.4.6 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility allows you to 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. This approach allows your users to 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's 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 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 there are any physical devices 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. POWER 6 processor-based hypervisors have this capability; firmware must be at firmware level eFW3.2 or later. All POWER7 processor-based hypervisors support 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 leverages 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 or 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, you might perform these 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://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7hc3/iphc3pcmco mbosact.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 HMC 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 should not be running on battery power, and shared storage (external hdisk with `reserve_policy=no_reserve`). In addition, all logical partitions should be on the same open network with RMC established to the HMC.

The HMC is used to configure, validate, and orchestrate. You use the HMC to configure the Virtual I/O Server as an MSP and to configure the VASI device. An HMC wizard validates your configuration and identifies things that can cause the migration to fail. During the migration, the HMC 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 POWER6 server. 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 whole 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.

3.4.7 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 only available with the Enterprise version of PowerVM.

The physical memory of an IBM Power System can be assigned to multiple partitions either in a dedicated or in a 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.

For additional information regarding Active Memory Sharing, see *PowerVM Virtualization Active Memory Sharing*, REDP-4470.

3.4.8 NPIV

N_Port ID virtualization (NPIV) is a technology that 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 about NPIV, see:

- ▶ *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

NPIV is supported in PowerVM Express, Standard, and Enterprise Editions, on the IBM Power System 750, 755, 770, and 780 servers, and for partitions using AIX 5.3, AIX 6.1, IBM i 6.1, SLES 11, and RHAT 5.4.

3.4.9 Operating System support for PowerVM

Table 3-6 summarizes the PowerVM features supported by the operating systems compatible with the POWER7 processor-based servers.

Table 3-6 PowerVM features supported by AIX, IBM i and Linux

Feature	AIX V5.3	AIX V6.1	IBM i 6.1.1	RHEL V5.4	SLES V10SP3	SLES V11
Simultaneous multithreading (SMT)	Yes ^a	Yes ^b	Yes ^c	Yes ^a	Yes ^a	Yes
DLPAR I/O adapter add/remove	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR processor add/remove	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR memory add	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR memory remove	Yes	Yes	Yes	No	No	Yes
Capacity Upgrade on Demand ^d	Yes	Yes	Yes	Yes	Yes	Yes
Micro-Partitioning	Yes	Yes	Yes	Yes	Yes	Yes
Shared Dedicated Capacity	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Shared Processor Pools	Yes	Yes	Yes	Yes	Yes	Yes
Virtual I/O Server	Yes	Yes	Yes	Yes	Yes	Yes
IVM	Yes	Yes	Yes	Yes	Yes	Yes
Virtual SCSI	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Ethernet	Yes	Yes	Yes	Yes	Yes	Yes
NPIV	Yes	Yes	Yes	Yes	No	Yes
Live Partition Mobility	Yes	Yes	No	Yes	Yes	Yes
Workload Partitions	No	Yes	No	No	No	No
Active Memory Sharing	No	Yes	Yes	No	No	Yes
Active Memory Expansion	No	Yes ^e	No	No	No	No

a. Support for only two threads

b. AIX 6.1 up to TL4 SP2 only supports two threads; supports four threads as of TL4 SP3

c. IBM i 6.1.1 and up support SMT4

d. Available on selected models

e. On AIX 6.1 with TL4 SP2 and later

3.4.10 POWER7 Linux programming 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 3-7 lists the support of specific programming features for various versions of Linux.

Table 3-7 Linux support for POWER7 features

Features	Linux Releases		Comments
	SLES 10 SP	SLES 11	
POWER6 compatibility mode	Yes	Yes	-
POWER7 mode	No	Yes	-
Strong Access Ordering	No	Yes	Can Improve Lx86 performance
Scale to 256 cores / 1024 threads	No	Yes	Base OS support available
4-way SMT	No	Yes	-
VSX Support	No	Partial	Full exploitation requires Advance Toolchain
Distro toolchain mcpu/mtune=p7	No	No	SLES11/GA toolchain has minimal P7 enablement necessary to support kernel build
Advance Toolchain Support	Yes; execution restricted to Power6 instructions	Yes	Alternative IBM GNU Toolchain
64k base page size	No	Yes	-
Tickless idle	No	Yes	Improved energy utilization and virtualization of partially to fully idle partitions.

Note: IBM is working with Red Hat on POWER7 support. Red Hat plans to support the Power 750, 755, 770, and 780 models in an upcoming release that is targeted for availability during the first half of 2010. For additional questions about the availability of this release, contact Red Hat

For information regarding Advance Toolchain, go to the following address:

<http://www.ibm.com/developerworks/wikis/display/hpccentral/How+to+use+Advance+Toolchain+for+Linux+on+POWER>

You may also visit the University of Illinois Linux on Power Open Source Repository:

- ▶ <http://ppclinux.ncsa.illinois.edu>
- ▶ 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 a system or systems to be partitioned with logical partitions. You may 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 the Integrated Virtualization Manager (IVM).

Note: Ask your IBM Representative or Business Partner to use the Customer Specified Placement manufacturing option if you want to automatically deploy your partitioning environment on a new machine. SPT looks for the resource's allocation to be 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 the following 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 may use the SPT before you order a system to determine what you must order to support your workload. You may also use the SPT to determine how you can partition a system that you already have.

Be sure to use the IBM System Planning Tool to estimate POWER Hypervisor requirements and determine the memory resources that are required for all partitioned and non-partitioned servers.

Figure 3-13 shows the estimated Hypervisor memory requirements, based on sample partition requirements.

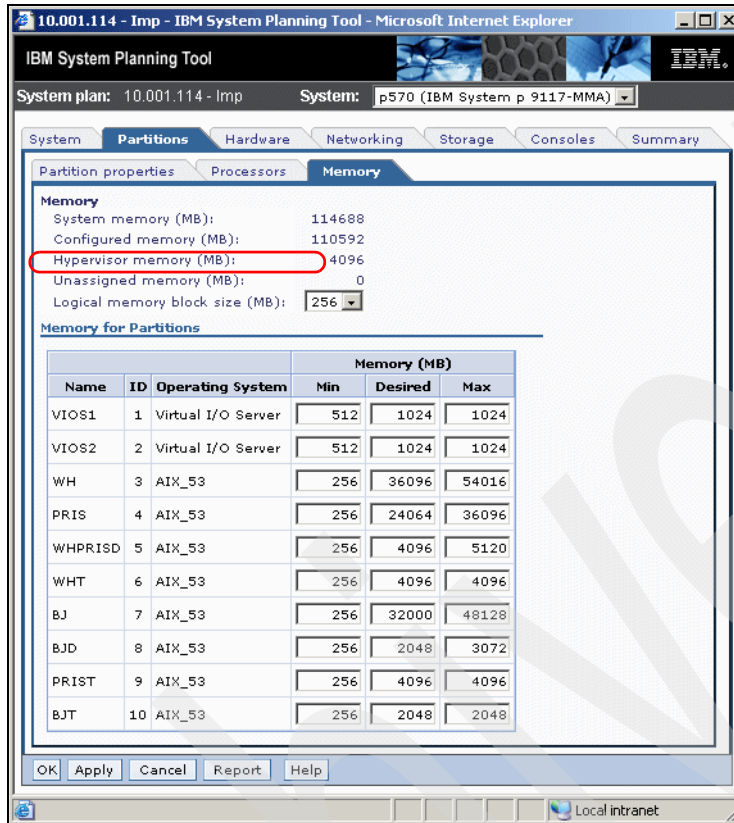


Figure 3-13 IBM System Planning Tool window showing Hypervisor memory requirements

The SPT and its supporting documentation are on the IBM System Planning Tool site:

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

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.

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 impacted by a fault or defect.
- ▶ **Serviceability:** Indicates how well faults and their impacts 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 (SER) 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 a number of features designed to avoid or correct environmentally induced, single-bit, intermittent failures as well as handle solid faults in components, including selective redundancy to tolerate certain faults without requiring an outage or parts replacement.

IBM is the only vendor that designs, manufactures, and integrates its most critical server components, including:

- ▶ POWER processors
- ▶ Caches
- ▶ Memory buffers
- ▶ Hub-controllers
- ▶ Clock cards
- ▶ Service processors

Design and manufacturing verification and integration, as well as field support information is used as feedback for continued improvement on the final products.

This chapter also includes a manageability section describing the means to successfully manage your systems.

Several software-based availability features exist that are based on the benefits available when using AIX and IBM i as the operating system. Support of these features when using Linux can vary.

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 heat generated 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; grade 3 defined as industry standard (off-the-shelf). As shown in Figure 4-1, using stringent design criteria and an extensive testing program, the IBM manufacturing team can produce grade 1 components that are expected to be 10 times more reliable than industry standard. Engineers select grade 1 parts for the most critical system components. Newly introduced organic packaging technologies, rated grade 5, achieve the same reliability as grade 1 parts.

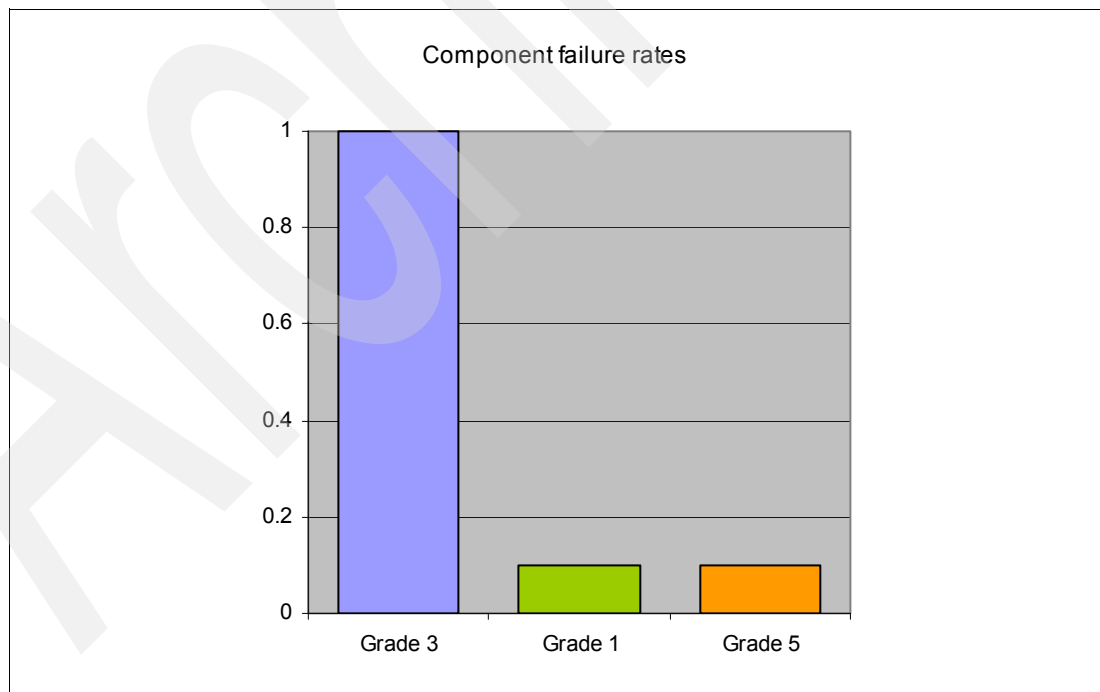


Figure 4-1 Component Failure rates

4.1.2 Placement of components

Packaging is designed to deliver both high performance and high reliability. For example, the reliability of electronic components is directly related to their thermal environment, that is, large decreases in component reliability 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 use of redundant parts allows the system to remain operational. Among the parts are:

- ▶ POWER7 cores, which include redundant bits in L1-I, L1-D, and L2 caches, and in L2 and L3 directories
- ▶ Power 770 and 780 main memory DIMMs, which contain an extra DRAM chip for improved redundancy
- ▶ Power 770 and 780 redundant system clock and service processor for configurations with two or more central electronics complex drawers (CEC)
- ▶ Redundant and hot-swap cooling
- ▶ Redundant and hot-swap power supplies
- ▶ Redundant 12X loops to I/O subsystem

For maximum availability, be sure to connect power cords from the same system to two separate Power Distribution Units (PDUs) in the rack, and to connect each PDU to independent power sources. Deskside form factor power cords must be plugged to two independent power sources in order to achieve maximum availability.

Note: 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 off-line before they reach the point of hard failure, causing a system outage, and without the need to re-create 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 are able to perform the following functions:

- ▶ Self-diagnose and self-correct during run time
- ▶ Automatically reconfigure to mitigate potential problems from suspect hardware
- ▶ Can self-heal or can 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 central electronics complex.

Throughout this chapter, we describe IBM POWER technology's capabilities that are focused on keeping a system environment up and 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 137.

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 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 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.

Persistent deallocation functions include:

- ▶ Processor
- ▶ L2/L3 cache lines (cache lines are dynamically deleted)
- ▶ Memory
- ▶ Deconfigure or bypass failing I/O adapters

Note: The auto-restart (reboot) option has to be enabled from the Advanced System Manager Interface or from the Control (Operator) Panel. Figure 4-2 shows this option using the ASMI.

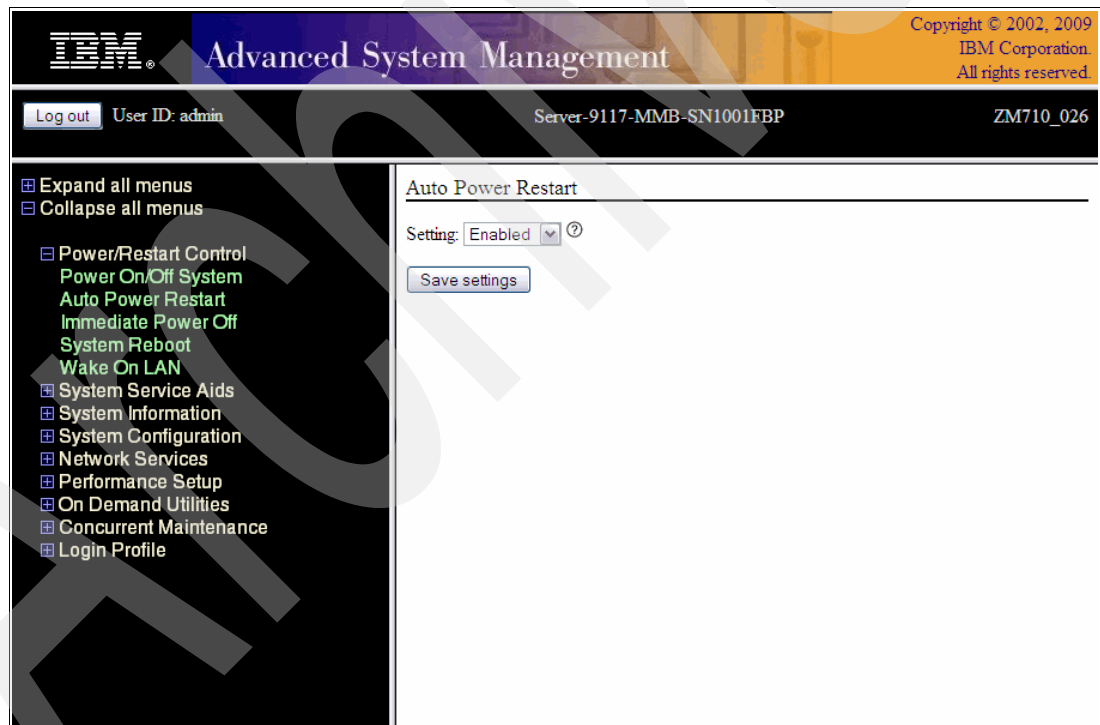


Figure 4-2 ASMI Auto Power Restart setting panel

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. This ability 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 this 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 continues as before.

On IBM systems prior to POWER6, this error caused a checkstop.

Alternate processor retry

Hard failures are more difficult, being 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 there are available inactivated processor cores or CoD processor cores, 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 there are no CoD processor cores available system-wide, total processor capacity is lowered below the licensed number of cores.

Single processor checkstop

As in POWER6, POWER7 provides single-processor 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 are in play, 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 very 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:

- ▶ 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. These capabilities include:

- ▶ 64-byte ECC code

This innovative ECC algorithm from IBM research allows a full 8-bit device kill to be corrected dynamically. This ECC code mechanism works on DIMM pairs on a rank basis. (Depending on the size, a DIMM might have one, two, or four ranks). With this ECC code, an entirely bad DRAM chip can be marked as bad (chip mark). After marking the DRAM as bad, the code corrects all the errors in the bad DRAM; it can additionally mark a 2-bit symbol as bad and correct the 2-bit symbol. Providing a double-error detect or single-error correct ECC, or a better level of protection in addition to the detection or correction of a chipkill event.

This improvement in the ECC word algorithm replaces the redundant bit steering used on POWER6 systems.

The Power 770 and 780, and future POWER7 high end machines, have a spare DRAM chip per rank on each DIMM that can be spared out. Effectively this protection means that on a rank basis, a DIMM pair can detect and correct two and sometimes three chipkill events and still provide better protection than ECC, explained in the previous paragraph.

- ▶ Hardware scrubbing

Hardware scrubbing is a method 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 spare data bit-line, for which is determined to be faulty.

► Chipkill

Chipkill is an enhancement that enables a system to sustain the failure of an entire DRAM chip. Chipkill spreads the bit lines from a DRAM over multiple ECC words, so that a catastrophic DRAM failure does not affect more of what is protected by the ECC code implementation. The system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced. Figure 4-3 shows an example of how chipkill technology spreads bit lines across multiple ECC words.

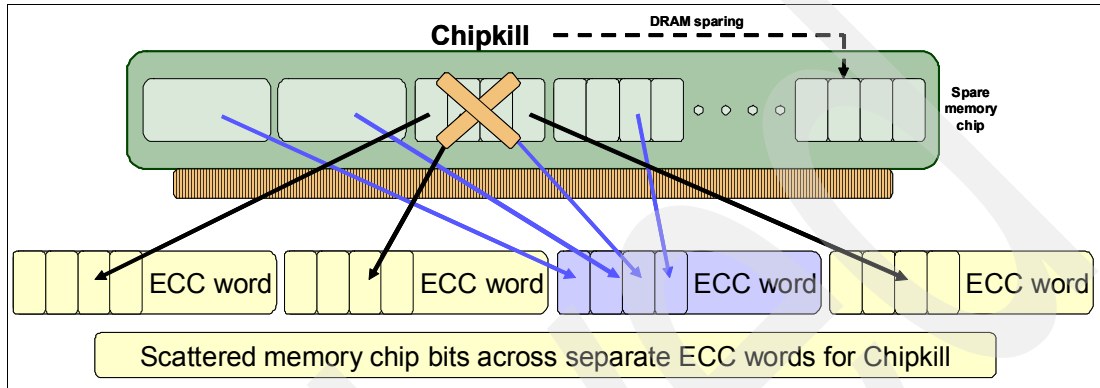


Figure 4-3 Chipkill in action with a spare memory DRAM chip on a Power 770 and 780

POWER7 memory subsystem

The POWER7 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, for which is determined to be faulty.

Figure 4-4 on page 132 shows a POWER7 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.

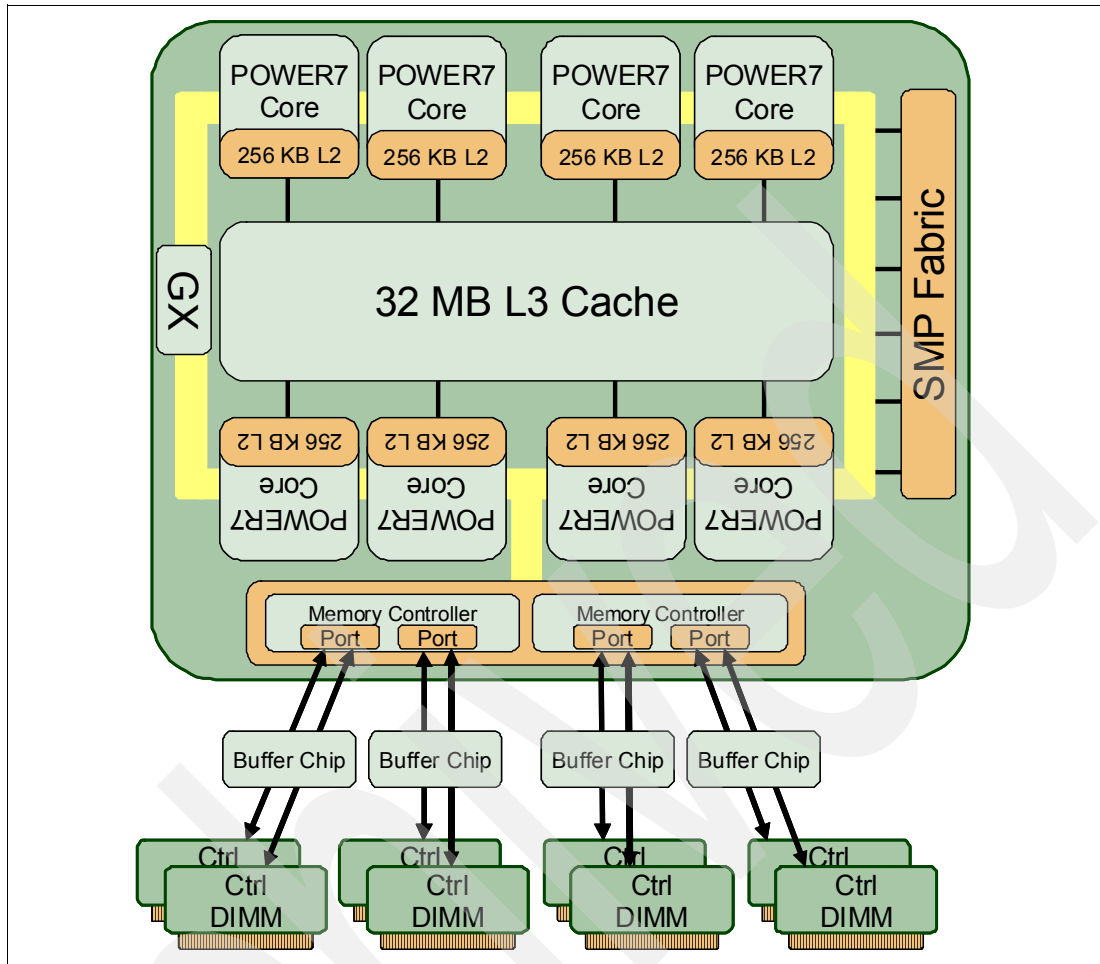


Figure 4-4 POWER7 memory subsystem

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, as well as for memory pages owned by the POWER Hypervisor. If a memory address experiences an uncorrectable or repeated correctable single cell error, the service processor sends the memory page address to the POWER Hypervisor to be marked for deallocation.

Pages 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 should be deallocated. Where possible, the operating system moves any data 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 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 deallocates the entire memory group that is associated with the error on all subsequent system reboots until the memory is repaired. This way is intended to guard against future uncorrectable errors while waiting for parts replacement.

Note: Memory page deallocation handles single cell failures, but, because of the sheer size of data in a data bit line, it may be inadequate for dealing with more catastrophic failures.

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 should 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™, the OS Service Focal Point, or the service processor receives a notification of the failed component, and triggers a service call.

4.2.4 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's instruction and data caches are protected against intermittent errors by using Processor Instruction Retry and against permanent errors by Alternate Processor Recovery, both mentioned earlier. 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 perform 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 detected in the cache can also trigger a purge and deleting of the cache line. This 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.

4.2.5 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 using a well-defined strategy that first considers the data source. Sometimes, an uncorrectable error is temporary in nature and occurs in data that can be recovered from another repository. For example:

- ▶ Data in the instruction L1 cache is never modified within the cache itself. Therefore, an uncorrectable error discovered in the cache is treated like an ordinary cache-miss, and correct data is loaded from the L2 cache.
- ▶ The 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. The system, instead, 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 owning 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 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 fail-over availability, redundant adapters spanning multiple PHB chips can enable the system to recover transparently, without partition loss.

4.2.6 PCI enhanced error handling

IBM estimates that PCI adapters can account for a significant portion of the hardware-based errors on a large server. Although servers that rely on boot-time diagnostics can identify failing components to be replaced by hot-swap and reconfiguration, runtime errors pose a more significant problem.

PCI adapters are generally complex designs involving 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 may 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, which 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.

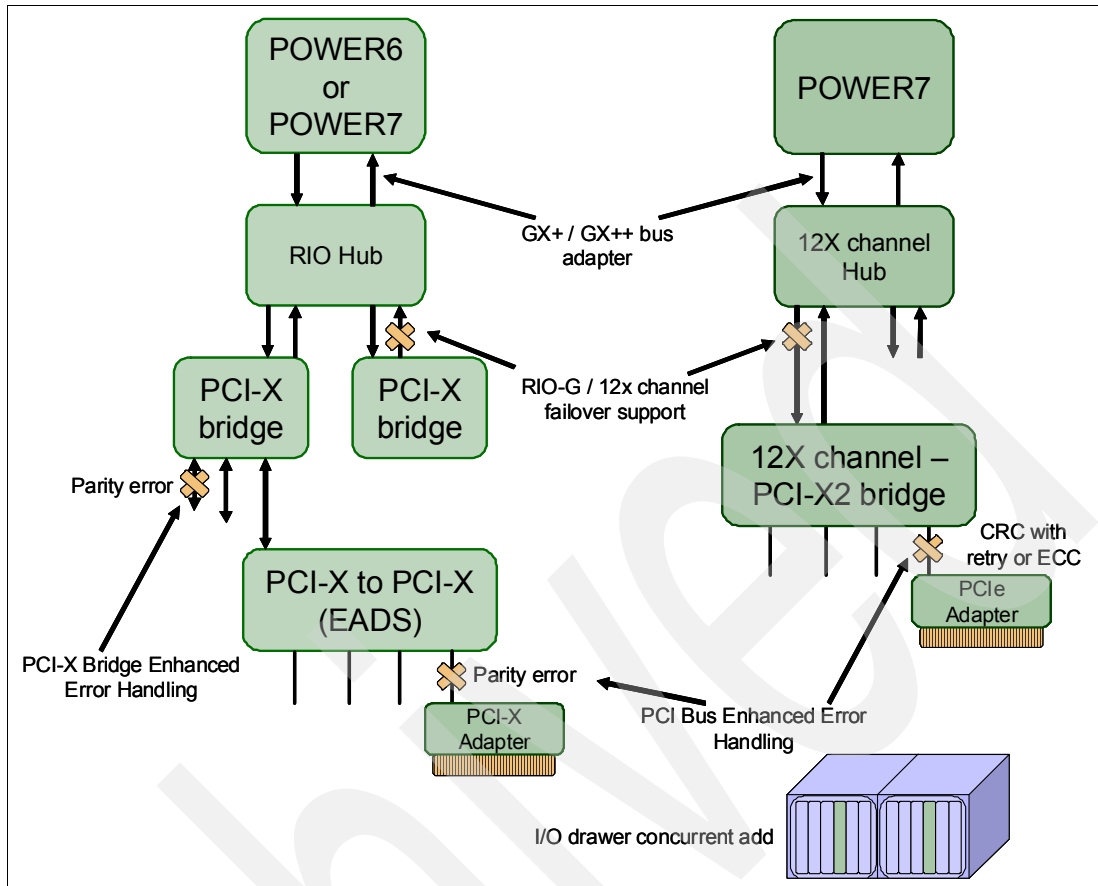


Figure 4-5 PCI enhanced error handling

4.3 Serviceability

IBM Power Systems design considers both IBM and client needs. The IBM Serviceability Team has enhanced the base service capabilities and continues to implement a strategy that incorporates best-of-breed 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 that includes:

- ▶ Easy access to service components; design for Customer Set Up (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.

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 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:

- ▶ POWER Hypervisor (system firmware) and Hardware Management Console connection surveillance
- ▶ Several 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 when:

- 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:

- 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 a Hardware Management Console, 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, 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 run time 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

Error checker signals are captured and stored in hardware fault isolation registers (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, run-time 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)

FFDC is an error isolation technique, which ensures that when a fault is detected in a system through error checkers or other types of detection methods, the root cause of the fault 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 related to the fault is captured and saved for analysis. In hardware, FFDC data is collected from the fault isolation registers 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 on page 139 shows a schematic of a fault isolation register implementation.

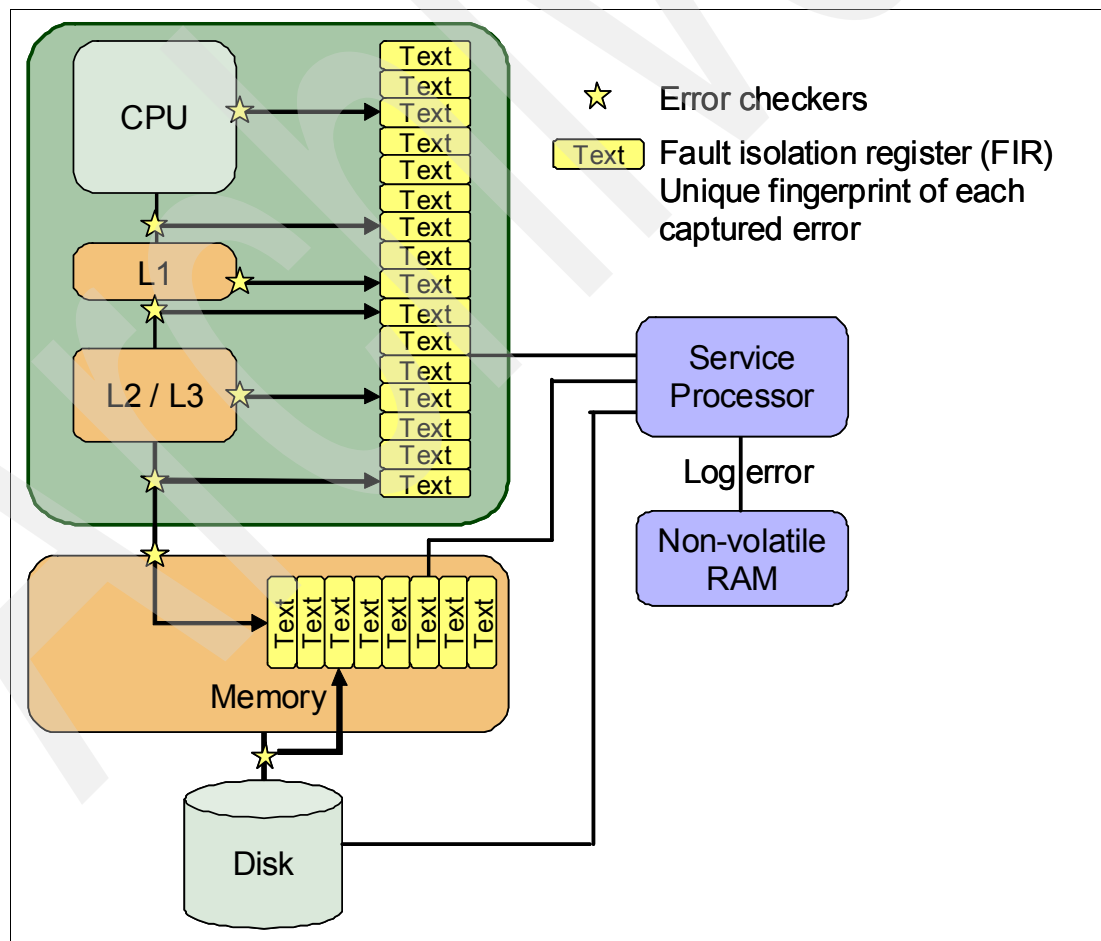


Figure 4-6 Schematic of FIR implementation

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) in order 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 has not been reached. Alternatively, it could indicate that a service action point has been reached, where the event exceeded a pre-determined threshold or was unrecoverable. Based on the isolation analysis, recoverable error threshold counts may 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. Boot-time diagnostic routines include:

- ▶ 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 may 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 in order 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 have been 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 thresholding) and predict when corrective actions must be automatically initiated by the system. These actions can include:

- ▶ Requests for a part to be replaced
- ▶ Dynamic 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 out an alert through 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 has 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:

- ▶ An error code uniquely describing the error event
- ▶ The location of the failing component
- ▶ The part number of the component to be replaced, including pertinent data 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 Remote Management and Control (RMC) application is delivered as part of the base operating system, including the operating system running on the Hardware Management Console. RMC provides a secure transport mechanism across the LAN interface between the operating system and the Hardware Management Console 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 should only be reported once, regardless of how many logical partitions experience the potential effect of the error. The Manage Serviceable Events task on the 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 Remote Management 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 like firmware levels, operating system levels, additional fault isolation register values, recoverable error threshold register values, system status, and any other pertinent data.

The data is formatted and prepared for transmission back to IBM to assist the service support organization with preparing a service action plan for the service representative or for additional analysis.

System dump handling

In 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 has detected, diagnosed, and reported 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 could 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 include:

- ▶ 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 in order to obtain the full value of IBM service enhancements.

Vital product data (VPD) and inventory management

Power Systems store vital product data (VPD) internally, which keeps 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, is 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):
 - Terracotta-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 may 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

Midrange and high-end systems, including models 770 and 780 and later, usually are repaired by IBM Support personnel.

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 some very complex high-end configurations.

In these situations, Guiding Light awaits 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.

Service labels

Service providers use these 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.

Several of these 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 are used to 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 several buttons, enabling a service support representative (SSR) or client to change various boot-time options and other limited service functions.

Concurrent maintenance

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 IBM POWER7 processor-based systems are designed for redundant or concurrently maintainable power, fans, physical storage, and I/O towers.

The most recent members of the IBM Power Systems family, based on the POWER7 processor, continue to support concurrent maintenance of power, cooling, PCI adapters, 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.

Hot-node add, hot-node repair, and memory upgrade

With the proper configuration and required protective measures, the Power 770 and 780 servers are designed for node add, node repair, or memory upgrade without powering down the system.

The Power 770 and 780 servers support the addition of another CEC enclosure (node) to a system (hot-node add) or adding more memory (memory upgrade) to an existing node. The additional Power 770 and 780 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 up the processors and memory in the node and its attached I/O resources (node evacuation).

To guard against any potential impact to system operation during hot-node add, memory upgrade, or node repair, clients must comply with the following protective measures:

- ▶ For memory upgrade and node repair, ensure that the system has sufficient inactive or spare processors and memory. Critical I/O resources must be configured with redundant paths.
- ▶ Schedule upgrades or repairs during non-peak operational hours.
- ▶ Move business applications to another server by using the PowerVM Live Partition Mobility feature or quiesce them. The user of LPM means that all critical applications must be halted or moved to another system before the operation begins. Non-critical applications can remain running. The partitions can be left running at the operating system command prompt.
- ▶ Back up critical application and system state information.
- ▶ Checkpoint the databases.

Blind-swap cassette

Blind-swap PCIe adapters represent significant service and ease-of-use enhancements in I/O subsystem design, while maintaining high PCIe 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, as on the Power 750 and 755.

Blind-swap allows PCIe adapters to be concurrently replaced without having to put the I/O drawer into a service position. Since first delivered, minor carrier design adjustments have improved an already well-thought-out service design.

For PCIe adapters on the POWER7 processor-based servers, blind swap cassettes include the PCIe slot, in order to avoid the top to bottom movement for inserting the card on the slot that was required on previous designs. The adapter is correctly connected by just sliding the cassette in.

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 using any of the following methods:

- ▶ IBM distributed media, such as a CD-ROM
- ▶ A Problem Fix distribution from the IBM Service and Support repository
- ▶ FTP from another server
- ▶ Download from the IBM Fix Central Web site:

<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 should plan to update in order 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 used to guide a service provider step-by-step through the process of repairing a system and verifying that the problem has been repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being repaired. Repair scenarios covered by repair and verify include:

- ▶ Replacing a defective 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

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 repair and verify procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event, to ensure that the guided maintenance procedures are operating correctly.

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 enable 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 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:

- ▶ Light Path and Guiding Light
For more information, see “Light Path” on page 144 and “Guiding Light” on page 144.
- ▶ 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 main system unit's state. The system unit can be in the following states:

- ▶ Standby (power off)
- ▶ Operating, ready to start partitions
- ▶ Operating with running logical partitions

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring the connection to the HMC for manageability purposes and accepting 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.

Note: The service processor enables a system that does not boot to be analyzed. The error log analysis can be performed from either the ASMI or the HMC.

The service processor uses two Ethernet 10/100Mbps ports. Note the following information:

- ▶ Both Ethernet ports are only visible 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:
 - 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 functions available through service processor include:

- ▶ 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 steps:

- ▶ 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, follow these 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 mobile computer 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 of the menu options in the Web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides a Secure Sockets Layer (SSL) Web connection to the service processor. To establish an SSL connection, open your browser using `https://`.

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 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.

The operator panel

The service processor provides an interface to the operator panel, which is used to display system status and diagnostic information.

The operator panel can be accessed in two ways:

- ▶ By using the normal operational front view
- ▶ By pulling it out to access the switches and view the LCD display. Figure 4-7 shows that the operator panel on a Power 770 and 780 is pulled out.

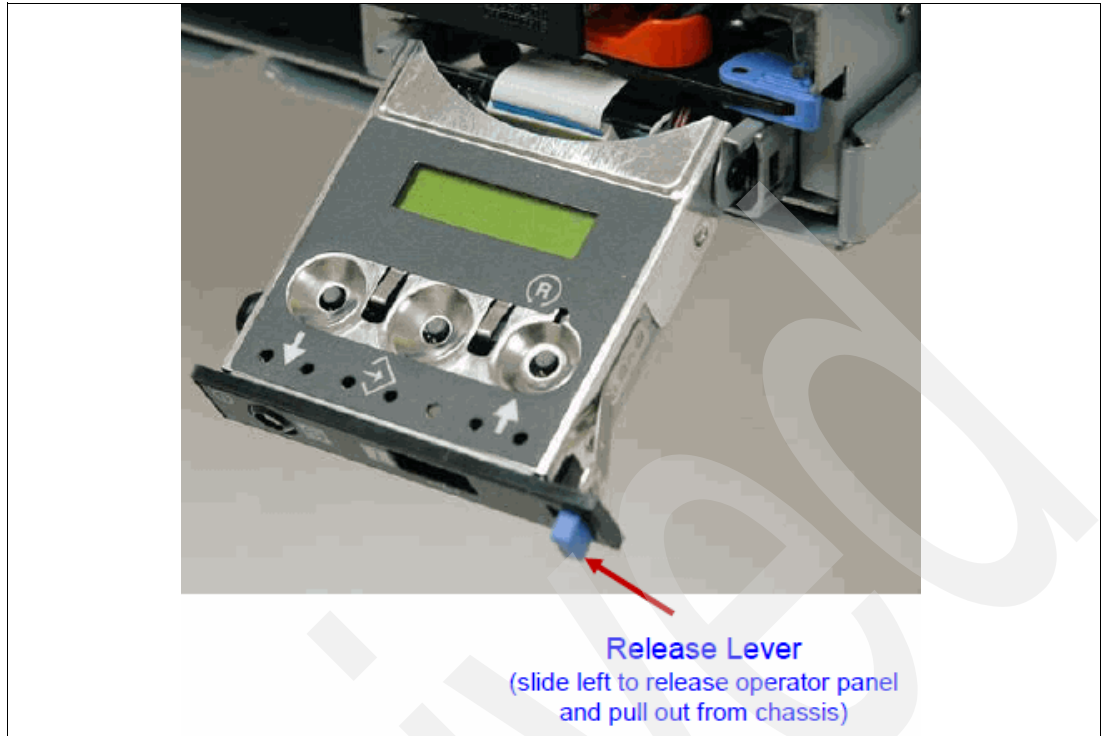


Figure 4-7 Operator panel is pulled out from the chassis

Several of the operator panel features include:

- ▶ A 2 x 16 character LCD display
- ▶ Reset, enter, power On/Off, increment and decrement buttons
- ▶ Amber System Information/Attention, green Power LED
- ▶ Blue Enclosure Identify LED on the Power 750 and 755
- ▶ Altitude sensor
- ▶ USB Port
- ▶ Speaker/Beeper

The functions available through the operator panel include:

- ▶ Error Information
- ▶ Generate dump
- ▶ View Machine Type, Model and Serial Number
- ▶ Limited set of repair functions

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

Requires a service mode boot of the system, enables the checking of system devices and features. Service mode provides the most complete checkout of the system resources. All system resources, except the SCSI adapter and the disk drives used for paging, can be tested.

► Concurrent mode

Enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, certain devices might require additional actions by the user or diagnostic application before testing can be done.

► Maintenance mode

Enables the checking of most system resources. Maintenance mode provides the same test coverage as service mode. The difference between the two modes is the way they are invoked. Maintenance mode requires that all activity on the operating system be stopped. The **shutdown -m** command is used to stop all activity on the operating system and put the operating system into maintenance mode.

The System Management Services (SMS) error log is accessible on the SMS menus. This error log contains errors that are found by partition firmware when the system or partition is booting.

The service processor's error log can be accessed on the ASMI menus.

You can also access the system diagnostics from a Network Installation Management (NIM) server.

Note: When you order a Power System, 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 up and running, 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, the service-level functions that you typically see when using the operating system service menus are as follows:

- ▶ 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
- ▶ Call Home/Customer Notification (both not available yet)
- ▶ 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 a Power 750 and 755, if that system is not managed by an HMC.

Power Systems has a permanent firmware boot side, or A side, and a temporary firmware boot side, or B side. New levels of firmware must be installed on the temporary side first in order to test the update's compatibility with existing applications. When the new level of firmware has been 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 on page 155 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 use 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 has been 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 nondisruptive 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. For deferred fixes within a service pack, only the fixes in the service pack, which cannot be concurrently activated, are deferred. Table 4-1 shows the file-naming convention for system firmware.

Table 4-1 Firmware naming convention

PPNNSSS_FFF_DDD			
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
SSS	Release indicator		
FFF	Current fix pack		
DDD	Last disruptive fix pack		

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 has 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.

- ▶ 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:

<https://www.ibm.com/support/electronic/portal>

4.5 Operating system support for RAS features

Table 4-2 gives an overview of a number of features for continuous availability that is supported by the various operating systems running on the POWER7 processor-based systems.

Table 4-2 Operating system support for RAS features

RAS feature	AIX 5.3	AIX 6.1	IBM i	RHEL 5	SLES 10	SLES 11
System deallocation of failing components						
Dynamic Processor Deallocation	X	X	X	X	X	X
Dynamic Processor Sparing	X	X	X	X	X	X
Processor Instruction Retry	X	X	X	X	X	X
Alternate Processor Recovery	X	X	X	X	X	X
Partition Contained Checkstop	X	X	X	X	X	X
Persistent processor deallocation	X	X	X	X	X	X
GX++ bus persistent deallocation	X	X	X	-	-	X
PCI bus extended error detection	X	X	X	X	X	X
PCI bus extended error recovery	X	X	X	Most	Most	Most
PCI-PCI bridge extended error handling	X	X	X	-	-	-
Redundant RIO or 12x Channel link	X	X	X	X	X	X
PCI card hot-swap	X	X	X	X	X	X
Dynamic SP failover at run-time	X	X	X	X	X	X
Memory sparing with CoD at IPL time	X	X	X	X	X	X
Clock failover runtime or IPL	X	X	X	X	X	X
Memory availability						
64-byte ECC code	X	X	X	X	X	X
Hardware scrubbing	X	X	X	X	X	X
CRC	X	X	X	X	X	X
Chipkill	X	X	X	X	X	X
L1 instruction and data array protection	X	X	X	X	X	X
L2/L3 ECC & cache line delete	X	X	X	X	X	X
Special uncorrectable error handling	X	X	X	X	X	X
Fault detection and isolation						
Platform FFDC diagnostics	X	X	X	X	X	X
Run-time diagnostics	X	X	X	Most	Most	Most
Storage Protection Keys	-	X	X	-	-	-

RAS feature	AIX 5.3	AIX 6.1	IBM i	RHEL 5	SLES 10	SLES 11
Dynamic Trace	X	X	X	-	-	X
Operating System FFDC	-	X	X	-	-	-
Error log analysis	X	X	X	X	X	X
Service Processor support for:						
Built-in-Self-Tests (BIST) for logic and arrays	X	X	X	X	X	X
Wire tests	X	X	X	X	X	X
Component initialization	X	X	X	X	X	X
Serviceability						
Boot-time progress indicators	X	X	X	Most	Most	Most
Firmware error codes	X	X	X	X	X	X
Operating system error codes	X	X	X	Most	Most	Most
Inventory collection	X	X	X	X	X	X
Environmental and power warnings	X	X	X	X	X	X
Hot-plug fans, power supplies	X	X	X	X	X	X
Extended error data collection	X	X	X	X	X	X
SP “call home” on non-HMC configurations	X	X	X	X	X	X
I/O drawer redundant connections	X	X	X	X	X	X
I/O drawer hot add and concurrent repair	X	X	X	X	X	X
Concurrent RIO/GX adapter add	X	X	X	X	X	X
Concurrent cold-repair of GX adapter	X	X	X	X	X	X
Concurrent add of powered I/O rack to Power 595	X	X	X	X	X	X
SP mutual surveillance w/ POWER Hypervisor	X	X	X	X	X	X
Dynamic firmware update with HMC	X	X	X	X	X	X
Service Agent Call Home Application	X	X	X	X	X	X
Guiding light LEDs	X	X	X	X	X	X
Lightpath LEDs	X	X	X	X	X	X
System dump for memory, POWER Hypervisor, SP	X	X	X	X	X	X
Infocenter / Systems Support Site service publications	X	X	X	X	X	X
System Support Site education	X	X	X	X	X	X
Operating system error reporting to HMC SFP	X	X	X	X	X	X
RMC secure error transmission subsystem	X	X	X	X	X	X
Health check scheduled operations with HMC	X	X	X	X	X	X
Operator panel (real or virtual)	X	X	X	X	X	X

RAS feature	AIX 5.3	AIX 6.1	IBM i	RHEL 5	SLES 10	SLES 11
Concurrent operator panel maintenance	X	X	X	X	X	X
Redundant HMCs	X	X	X	X	X	X
Automated server recovery/restart	X	X	X	X	X	X
High availability clustering support	X	X	X	X	X	X
Repair and Verify Guided Maintenance	X	X	X	Most	Most	Most
Concurrent kernel update	-	X	X	-	-	-
Hot-node Add ^a	-	-	-	-	-	-
Cold-node Repair ^a	-	-	-	-	-	-
Concurrent-node Repair ^a	-	-	-	-	-	-

a. eFM 3.2.2 and later.

Related publications

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

IBM Redbooks

For information about ordering these publications, see “How to get Redbooks” on page 163. Note that some of the documents referenced here may be available in softcopy only.

- ▶ *IBM Power 750 and 755 Technical Overview and Introduction*, REDP-4638
- ▶ *IBM PowerVM Live Partition Mobility*, SG24-7460
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- ▶ *IBM System p Advanced POWER Virtualization (PowerVM) Best Practices*, REDP-4194
- ▶ *IBM System Storage DS8000: Copy Services in Open Environments*, SG24-6788
- ▶ *PowerVM and SAN Copy Services*, REDP-4610
- ▶ *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- ▶ *PowerVM Virtualization on IBM System p: Introduction and Configuration Fourth Edition*, SG24-7940
- ▶ *SAN Volume Controller V4.3.0 Advanced Copy Services*, SG24-7574

Online resources

The POWER7 server data sheets and other resources can be found on the following Web pages:

- ▶ Active Memory Expansion: Overview and Usage
http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&apnname=STGE_PO_PO_USEN&htmlfid=POW03037USEN
- ▶ Advance Toolchain for Linux
<http://www.ibm.com/developerworks/wikis/display/hpccentral/How+to+use+Advance+Toolchain+for+Linux+on+POWER>
- ▶ Capacity on Demand
<http://www.ibm.com/systems/power/hardware/cod/>
- ▶ Download from the IBM Fix Central
<http://www.ibm.com/support/fixcentral/>
- ▶ Electronic Services information
<https://www.ibm.com/support/electronic/portal>
- ▶ IBM Power Systems Facts and Features: POWER7 Servers
http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=PM&subtype=BR&apnname=STGE_PO_PO_USEN&htmlfid=POB03022USEN&attachment=POB03022USEN.PDF

- ▶ IBM Power Systems Hardware Information Center
<http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp>
- ▶ IBM Storage U.S.A Web site
<http://www.ibm.com/systems/storage/>
- ▶ IBM System Planning Tool
<http://www.ibm.com/systems/support/tools/systemplanningtool/>
- ▶ Partition Mobility and migration compatibility modes
<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7hc3/iphc3pcmcombosact.htm>
- ▶ Power 750
http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=PM&subtype=SP&appname=STGE_PO_PO_USEN&htmlfid=POD03034USEN&attachment=POD03034USEN.PDF
- ▶ Power 755
http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=PM&subtype=SP&appname=STGE_PO_PO_USEN&htmlfid=POD03035USEN&attachment=POD03035USEN.PDF
- ▶ Power 770
http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=PM&subtype=SP&appname=STGE_PO_PO_USEN&htmlfid=POD03031USEN&attachment=POD03031USEN.PDF
- ▶ Power 780
http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=PM&subtype=SP&appname=STGE_PO_PO_USEN&htmlfid=POD03032USEN&attachment=POD03032USEN.PDF
- ▶ POWER7 2.06 of the Power Instruction Set Architecture
http://www.power.org/resources/downloads/PowerISA_V2.06_PUBLIC.pdf
- ▶ Specific storage devices supported for Virtual I/O Server
<http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html>
- ▶ Support for IBM Systems (access to the initial Web pages that address support)
<http://www.ibm.com/systems/support>
- ▶ Version 2.05 of the Power Instruction Set Architecture (ISA)
http://www.power.org/resources/reading/PowerISA_V2.05.pdf
- ▶ Virtual networking on AIX
http://www.ibm.com/servers/aix/whitepapers/aix_vn.pdf

How to get Redbooks

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

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Archived

Archived



IBM Power 770 and 780 Technical Overview and Introduction



Features the POWER7 processor for advanced multi-core technology

Describes MaxCore and TurboCore for redefining performance

Discusses enterprise-level RAS in a midrange footprint

This IBM Redpaper publication is a comprehensive guide covering the IBM Power 770 and Power 780 servers supporting IBM AIX, IBM i, and Linux operating systems. The goal of this paper is to introduce the major innovative Power 770 and 780 offerings and their prominent functions, including:

- ▶ Unique modular server packaging
- ▶ The specialized IBM POWER7 Level 3 cache that provides greater bandwidth, capacity, and reliability
- ▶ The 1 Gb or 10 Gb Integrated Virtual Ethernet adapter that brings native hardware virtualization up to 64 logical ports on this server
- ▶ IBM PowerVM virtualization including PowerVM Live Partition Mobility and PowerVM Active Memory Sharing
- ▶ 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
- ▶ Enterprise-ready reliability, serviceability, and availability

Professionals who want to acquire a better understanding of IBM Power Systems products should read this Redpaper.

This Redpaper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the 770 and 780 systems.

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, may be used to enhance your knowledge of IBM server solutions.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

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

For more information:
ibm.com/redbooks