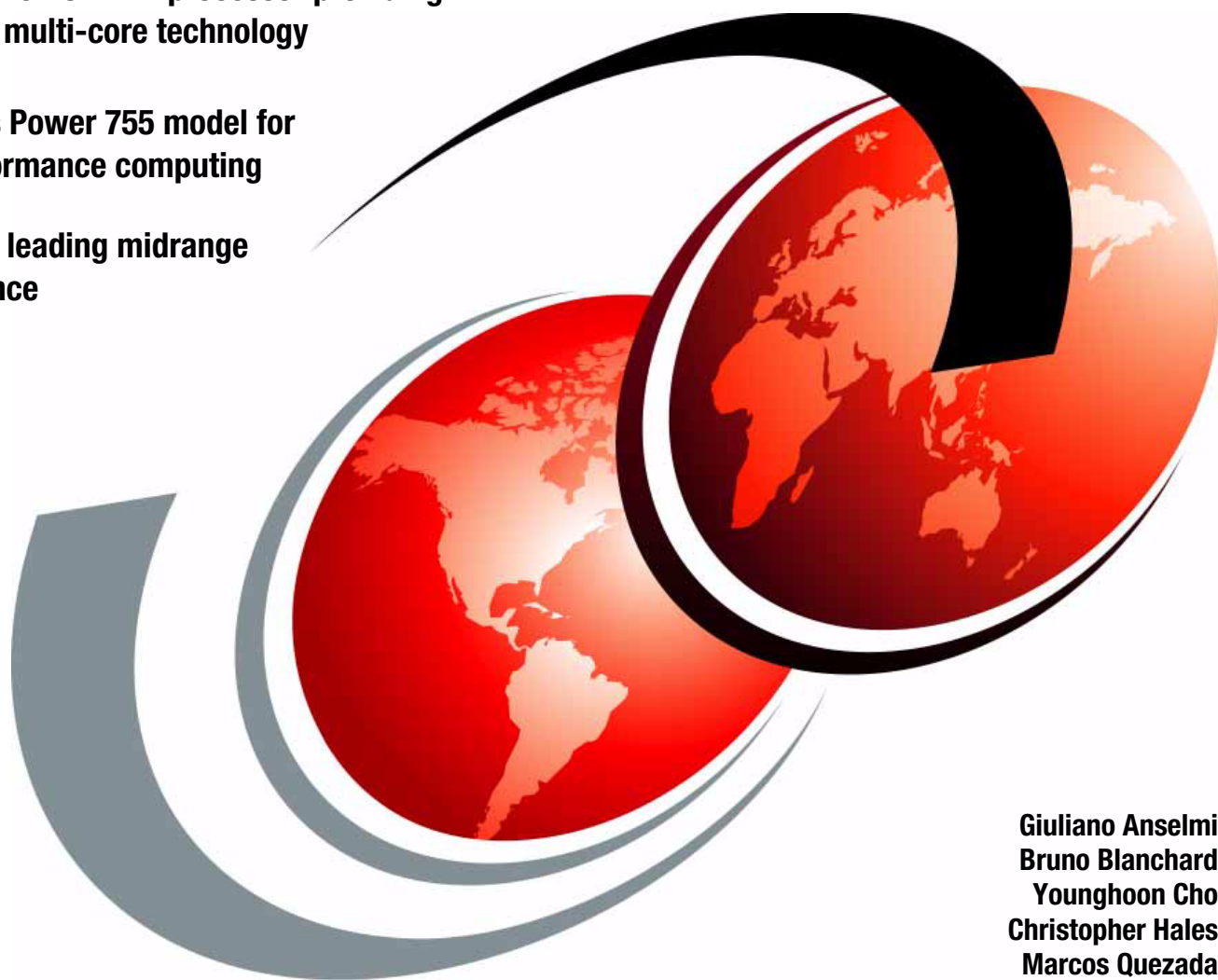


IBM Power 750 and 755 Technical Overview and Introduction

Features the POWER7 processor providing advanced multi-core technology

Discusses Power 755 model for high performance computing

Describes leading midrange performance



Giuliano Anselmi
Bruno Blanchard
Younghoon Cho
Christopher Hales
Marcos Quezada



International Technical Support Organization

IBM Power 750 and 755 Technical Overview and Introduction

March 2010

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

First Edition (March 2010)

This edition applies to the IBM Power 750 (8233-E8B) and IBM Power 755 (8236-E8C) 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 | 4 |
| 1.3 Physical package | 4 |
| 1.4 System features | 6 |
| 1.4.1 Power 750 Express system features | 6 |
| 1.4.2 Power 755 system features | 7 |
| 1.4.3 Minimum features | 8 |
| 1.4.4 Power supply features | 10 |
| 1.4.5 Processor card features | 11 |
| 1.4.6 Memory features | 13 |
| 1.5 Disk and media features | 14 |
| 1.6 I/O drawers for Power 750 | 15 |
| 1.6.1 PCI-DDR 12X Expansion Drawers (#5796) | 15 |
| 1.6.2 12X I/O Drawer PCIe (#5802 and #5877) | 15 |
| 1.6.3 I/O drawers and usable PCI slot | 16 |
| 1.7 Comparison between models | 16 |
| 1.8 Build to Order | 17 |
| 1.9 IBM Editions | 17 |
| 1.10 Model upgrades | 18 |
| 1.11 Hardware Management Console models | 18 |
| 1.12 System racks | 19 |
| 1.12.1 IBM 7014 Model T00 rack | 19 |
| 1.12.2 IBM 7014 Model T42 rack | 20 |
| 1.12.3 IBM 7014 Model S25 rack | 20 |
| 1.12.4 Feature number 0555 rack | 20 |
| 1.12.5 Feature number 0551 rack | 20 |
| 1.12.6 Feature number 0553 rack | 21 |
| 1.12.7 The AC power distribution unit and rack content | 21 |
| 1.12.8 Rack-mounting rules | 22 |
| 1.12.9 Useful rack additions | 22 |
| 1.12.10 OEM rack | 23 |
| Chapter 2. Architecture and technical overview | 25 |
| 2.1 The IBM POWER7 processor | 27 |
| 2.1.1 POWER7 processor overview | 28 |
| 2.1.2 POWER7 processor core | 29 |
| 2.1.3 Simultaneous multithreading | 30 |
| 2.1.4 Memory access | 31 |
| 2.1.5 Flexible POWER7 processor packaging and offerings | 31 |

| | | |
|-------------------|---|-----------|
| 2.1.6 | On-chip L3 cache innovation and Intelligent Cache | 32 |
| 2.1.7 | POWER7 processor and Intelligent Energy | 33 |
| 2.1.8 | Comparison of the POWER7 and POWER6 processors | 34 |
| 2.2 | POWER7 processor cards | 34 |
| 2.3 | Memory subsystem | 36 |
| 2.3.1 | Registered DIMM | 36 |
| 2.3.2 | Memory placement rules. | 36 |
| 2.3.3 | Memory throughput. | 38 |
| 2.4 | Capacity on Demand. | 38 |
| 2.5 | Technical comparison of Power 750 and Power 755 | 38 |
| 2.6 | I/O buses and GX card | 39 |
| 2.7 | Internal I/O subsystem | 40 |
| 2.7.1 | Slot configuration | 40 |
| 2.7.2 | System ports | 41 |
| 2.8 | Integrated Virtual Ethernet adapter | 41 |
| 2.8.1 | IVE features | 41 |
| 2.8.2 | IVE subsystem | 43 |
| 2.9 | PCI adapters | 43 |
| 2.9.1 | LAN adapters | 44 |
| 2.9.2 | Graphics accelerators | 45 |
| 2.9.3 | SCSI and SAS adapters | 46 |
| 2.9.4 | iSCSI. | 47 |
| 2.9.5 | Fibre Channel adapter | 48 |
| 2.9.6 | Fibre Channel over Ethernet (FCoE) | 49 |
| 2.9.7 | InfiniBand Host Channel adapter | 50 |
| 2.9.8 | Asynchronous adapter | 51 |
| 2.10 | Internal storage | 51 |
| 2.10.1 | Dual-write cache RAID feature | 52 |
| 2.10.2 | External SAS port | 53 |
| 2.10.3 | Split DASD backplane feature | 53 |
| 2.10.4 | Media bays | 53 |
| 2.11 | External I/O subsystems | 54 |
| 2.11.1 | PCI-DDR 12X Expansion Drawer (#5796) | 54 |
| 2.11.2 | 12X I/O Drawer PCIe (#5802 and #5877). | 55 |
| 2.11.3 | Dividing SFF drive bays in 12X I/O drawer PCIe | 57 |
| 2.11.4 | 12X I/O Drawer PCIe and PCI-DDR 12X Expansion Drawer 12X cabling | 59 |
| 2.11.5 | 12X I/O Drawer PCIe and PCI-DDR 12X Expansion Drawer SPCN cabling | 61 |
| 2.12 | External disk subsystems | 62 |
| 2.12.1 | EXP 12S Expansion Drawer | 62 |
| 2.12.2 | IBM System Storage | 63 |
| 2.13 | Hardware Management Console (HMC) | 65 |
| 2.13.1 | HMC functional overview | 65 |
| 2.13.2 | HMC connectivity to the POWER7 processor based systems | 67 |
| 2.13.3 | High availability using the HMC | 69 |
| 2.13.4 | HMC code level. | 70 |
| 2.14 | IVM | 70 |
| 2.15 | Operating system support. | 72 |
| 2.16 | Compiler technology | 75 |
| 2.17 | Energy management. | 76 |
| 2.17.1 | IBM EnergyScale technology | 76 |
| 2.17.2 | Thermal power management device card (TPMD) | 78 |
| Chapter 3. | Virtualization | 81 |

| | | |
|---|--|------------|
| 3.1 | POWER Hypervisor | 82 |
| 3.2 | POWER processor modes | 85 |
| 3.3 | Active Memory Expansion | 86 |
| 3.4 | PowerVM | 90 |
| 3.4.1 | PowerVM editions | 90 |
| 3.4.2 | Logical partitions | 91 |
| 3.4.3 | Multiple Shared-Processor Pools | 94 |
| 3.4.4 | Virtual I/O Server | 98 |
| 3.4.5 | PowerVM Lx86 | 102 |
| 3.4.6 | PowerVM Live Partition Mobility | 102 |
| 3.4.7 | Active Memory Sharing | 104 |
| 3.4.8 | NPIV | 105 |
| 3.4.9 | Operating system support for PowerVM | 105 |
| 3.4.10 | POWER7 and Linux programming support | 106 |
| 3.5 | System Planning Tool | 107 |
| Chapter 4. Continuous availability and manageability | | 109 |
| 4.1 | Reliability | 111 |
| 4.1.1 | Designed for reliability | 111 |
| 4.1.2 | Placement of components | 112 |
| 4.1.3 | Redundant components and concurrent repair | 112 |
| 4.2 | Availability | 112 |
| 4.2.1 | Partition availability priority | 113 |
| 4.2.2 | General detection and deallocation of failing components | 113 |
| 4.2.3 | Memory protection | 115 |
| 4.2.4 | Cache protection | 118 |
| 4.2.5 | Special uncorrectable error handling | 119 |
| 4.2.6 | PCI enhanced error handling | 120 |
| 4.3 | Serviceability | 121 |
| 4.3.1 | Detecting | 122 |
| 4.3.2 | Diagnosing | 125 |
| 4.3.3 | Reporting | 126 |
| 4.3.4 | Notifying | 128 |
| 4.3.5 | Locating and servicing | 129 |
| 4.4 | Manageability | 133 |
| 4.4.1 | Service user interfaces | 133 |
| 4.4.2 | IBM Power Systems firmware maintenance | 139 |
| 4.4.3 | Electronic Services and Electronic Service Agent | 142 |
| 4.5 | Operating system support for RAS features | 142 |
| Related publications | | 145 |
| | IBM Redbooks | 145 |
| | Online resources | 145 |
| | How to get Redbooks | 147 |
| | Help from IBM | 147 |

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

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

| | | |
|--|-------------------------|---|
| Active Memory™ | Micro-Partitioning™ | pSeries® |
| AIX 5L™ | POWER Hypervisor™ | Rational® |
| AIX® | Power Systems™ | Redbooks® |
| DB2® | Power Systems Software™ | Redpaper™ |
| DS8000® | POWER4™ | Redbooks (logo)  ® |
| Electronic Service Agent™ | POWER4+™ | RS/6000® |
| EnergyScale™ | POWER5™ | Solid® |
| FlashCopy® | POWER5+™ | System p5® |
| Focal Point™ | POWER6+™ | System p® |
| HACMP™ | POWER6® | System Storage™ |
| i5/OS® | POWER7™ | System z® |
| IBM Systems Director Active Energy Manager™ | PowerPC® | Tivoli® |
| IBM® | PowerVM™ | TotalStorage® |
| | POWER® | XIV® |

The following terms are trademarks of other companies:

SnapManager, and the NetApp logo are trademarks or registered trademarks of NetApp, Inc. 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.

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

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

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

Preface

This IBM® Redpaper™ publication is a comprehensive guide covering the IBM Power 750 and Power 755 servers supporting AIX®, IBM i, and Linux® operating systems. The goal of this paper is to introduce the major innovative Power 750 and 755 offerings and their prominent functions, including:

- ▶ The POWER7™ processor available at frequencies of 3.0 GHz, 3.3 GHz, and 3.55 GHz
- ▶ The specialized POWER7 Level 3 cache that provides greater bandwidth, capacity, and reliability
- ▶ The 1 Gb or 10 Gb Integrated Virtual Ethernet adapter, included with each server configuration, and providing native hardware virtualization
- ▶ 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
- ▶ EnergyScale™ technology that provides features such as power trending, power-saving, capping of power, and thermal measurement.

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

- ▶ 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 550 system.

This paper does not replace the latest marketing materials and configuration tools. It is intended as an additional source of information that, together with existing sources, 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, Craig G. Johnson, Deanna M. Johnson, Roxette Johnson, Ronald Kalla, Bob Kovacs, Phil N. Lewis, Casey McCreary, Jeff Meute, Michael Middleton, Bill Moran, Michael J Mueller, Steve Munroe, Duc Nguyen, Thoi Nguyen, Mark Olson, Patrick O'Rourke, Jan Palmer, Amartey Pearson, Audrey Romonosky, Jeffrey Scheel, Kimberly Schmid, Helena Sunny, Joel Tendler, Jeff Van Heuklon, Jonathan Van Niewaal, 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>



General description

The Power 750 Express and Power 755 systems utilize the innovative IBM POWER7 processor technology that is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding workloads.

The Power 750 Express is designed to address challenging commercial workloads, whereas the Power 755 is a compute node that is particularly suited to high performance computing (HPC) workloads.

1.1 Overview of systems

Figure 1-1 shows the Power 750 Express and Power 755 systems.

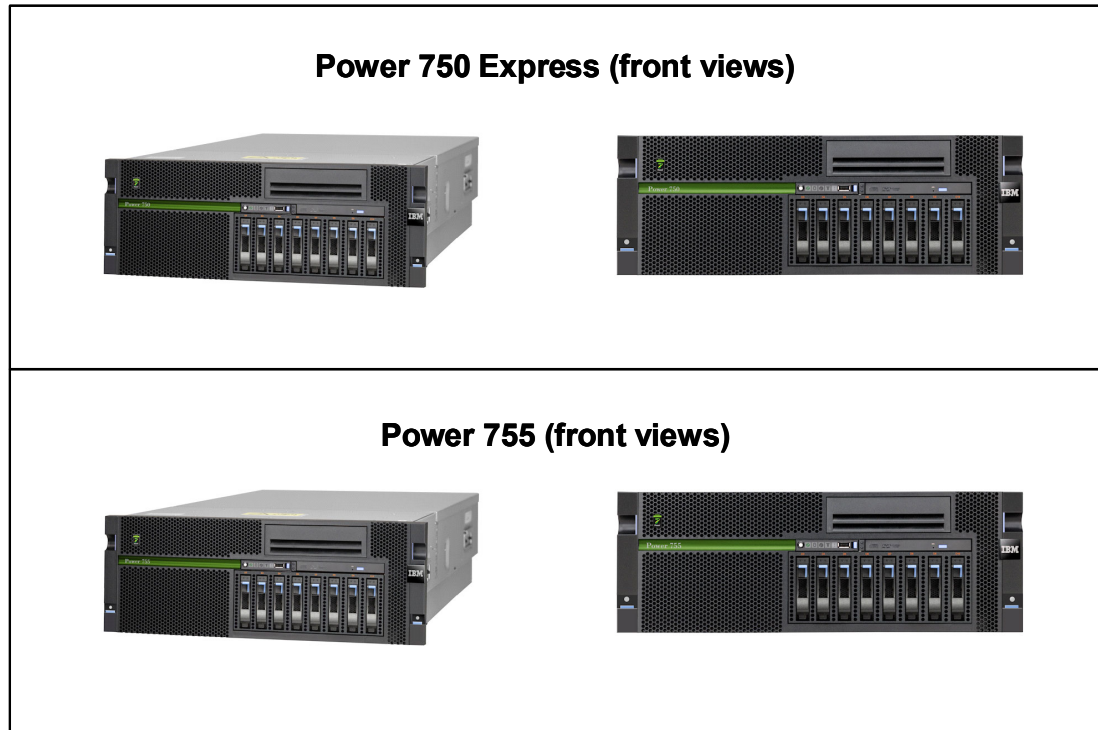


Figure 1-1 The Power 750 Express system and the Power 755 compute node

The Power 750 Express server

The Power 750 Express server (8233-E8B) supports up to four 6-core 3.3 GHz or four 8-core 3.0 GHz, 3.3 GHz, and 3.55 GHz POWER7 processor cards in a rack-mount drawer configuration. The POWER7 processors in this server are 64-bit, 6-core and 8-core modules that are packaged on dedicated processor cards with 4 MB of L3 cache per core and 256 KB of L2 cache per core.

The Power 750 Express server supports a maximum of 32 DDR3 DIMM slots, eight DIMM slots per processor card. Memory features (two memory DIMMs per feature) supported are 8 GB, 16 GB, and 32 GB, and run at speeds of 1066 MHz. A system with four processor cards installed has a maximum memory capacity of 512 GB.

The Power 750 Express server provides great I/O expandability. For example, with 12X-attached I/O drawers, the system can have up to 50 PCI-X slots or up to 41 PCIe slots. This combination can provide over 100 LAN ports or up to 576 disk drives (over 240 TB of disk storage). Extensive quantities of externally attached storage and tape drives and libraries can also be attached.

The Power 750 Express system unit without I/O drawers can contain a maximum of either eight small form factor (SFF) SAS disks or eight SFF SAS solid state drives (SSDs), providing up to 2.4 TB of disk storage.

All disks and SSDs are direct dock and hot pluggable. The eight SAS bays can be split into two sets of four bays for additional AIX or Linux configuration flexibility. The system unit also

contains a slimline DVD-RAM, plus a half-high media bay for an optional tape drive or removable disk drive.

Also available in the Power 750 Express system unit is a choice of quad-gigabit or dual-10 Gb integrated host Ethernet adapters. These native ports can be selected at the time of initial order. Virtualization of these integrated Ethernet adapters is supported.

IBM Power 755 compute node

The IBM Power 755 (8236-E8C) compute node is designed for organizations that require a scalable system with extreme parallel processing performance and dense packaging. Ideal workloads for the Power 755 include high performance computing (HPC) applications such as weather and climate modeling, computational chemistry, physics, and petroleum reservoir modeling that require highly intense computations where the workload is aligned with parallel processing methodologies.

The Power 755 server is a 3.3 GHz 32-core system based on the IBM POWER7 processor, and addresses workloads in HPC environments. A single Power 755 system provides four 64-bit, eight-core POWER7 processor modules with 4 MB of L3 cache per core and 256 KB of L2 cache per core. Each POWER7 processor module is packaged on its own processor card. The processor card has eight DDR3 DIMM slots, the Power 755 system has a total of 32 DIMM slots offering a maximum of 256 GB memory when all 32 DIMM slots are filled with 8 GB DIMMs.

Up to 64 Power 755 nodes (each with 32 POWER7 processor cores) can be clustered together using 12X InfiniBand adapters. This provides an HPC compute resource of up to 2,048 POWER7 processor cores. The IBM HPC software stack provides the necessary development tools, libraries, and system management software to manage a Power 755 server cluster.

The Power 755 system unit provides up to five PCI slots, one GX++ slot for a 12X adapter, eight SFF (small form factor) SAS bays, and a DVD-RAM. Three of the five Peripheral Component Interconnect (PCI) slots are PCI Express (PCIe) 8x and two are PCI-extended (PCI-X) DDR. The GX++ slot can hold a 12X InfiniBand adapter supporting 4x connection to other Power 755 systems.

The eight SAS bays contain a minimum of two disks and a maximum of eight disks or SSDs, providing up to 2.4 TB of storage capacity. Up to an additional 156 SAS bays are available using the EXP12S SAS disk/SSD drawer (#5886), providing up to 70 TB of additional capacity. All drives are direct dock and hot pluggable.

The Power 755 system unit also provides a choice of quad-gigabit or dual-10 Gb integrated host Ethernet adapters, which can be extensively virtualized. These ports are selected at the time of initial order and do not use a PCI slot.

1.2 Operating environment

The operating environment specifications for the servers can be seen in Table 1-1.

Table 1-1 Operating environment for Power 750 Express and Power 755

| Power 750 Express and Power 755 operating environment | | |
|---|--|--|
| Description | Operating | Non-operating |
| Temperature | 5 to 35 degrees C (41 to 95 degrees F) Recommended: 18 - 27 degrees C (64 - 80 degrees F) | 5 - 45 degrees C (41 - 113 degrees F) |
| Relative humidity | 20 - 60% | 8 - 80% |
| Maximum dew point | 29 degrees C (84 degrees F) | 28 degrees C (82 degrees F) |
| Operating voltage | 200 - 240 V ac | N/A |
| Operating frequency | 50 - 60 +/- 3 Hz | N/A |
| Power consumption | 1950 watts maximum | N/A |
| Power source loading | 2.0 kVA maximum | N/A |
| Thermal output | 6655 Btu/hr maximum | N/A |
| Maximum altitude | 3048 m (10,000 ft) | N/A |
| Noise-level reference point ^a : (12 cores at 3.3 GHz, 16x 8 GB DIMMs, 2x power supplies, 8x SFF disks, 1x DVD-RAM, 3x PCI adapters) | 6.2/6.4 bels (operating/idle) 5.8/5.6 bels (operating/idle) with acoustic rack doors | |
| Noise-level reference point ^a : (24 cores at 3.3 GHz, 16x 8 GB DIMMs, 2x power supplies, 8x SFF disks, 1x DVD-RAM, 3x PCI adapters) | 7.1/7.1 bels (operating/idle) 6.5/6.5bels (operating/idle) with acoustic rack doors | |

a. For noise-level reference points for specific server configurations, see the service manual for the relevant model.

1.3 Physical package

Table 1-2 on page 5 shows the physical dimensions of the Power 750 Express and Power 755 chassis. Both servers are available only in a rack-mounted form factor and each can take four EIA units (4U) of rack space.

Table 1-2 Physical dimensions of a Power 750 Express and Power 755 chassis

| Dimension | Power 750 Express (Model 8233-E8B) | Power 755 (Model 8236-E8C) |
|-----------|------------------------------------|----------------------------------|
| Width | 443 mm (17.3 in) | 443 mm (17.3 in) |
| Depth | 730 mm (28.7 in) | 730 mm (28.7 in) |
| Height | 173 mm (6.81 in), 4U (EIA units) | 173 mm (6.81 in), 4U (EIA units) |
| Weight | 48.7 kg (107.4 lbs) | 48.7 kg (107.4 lbs) |

The front and rear views of the Power 750 Express can be seen in Figure 1-2.

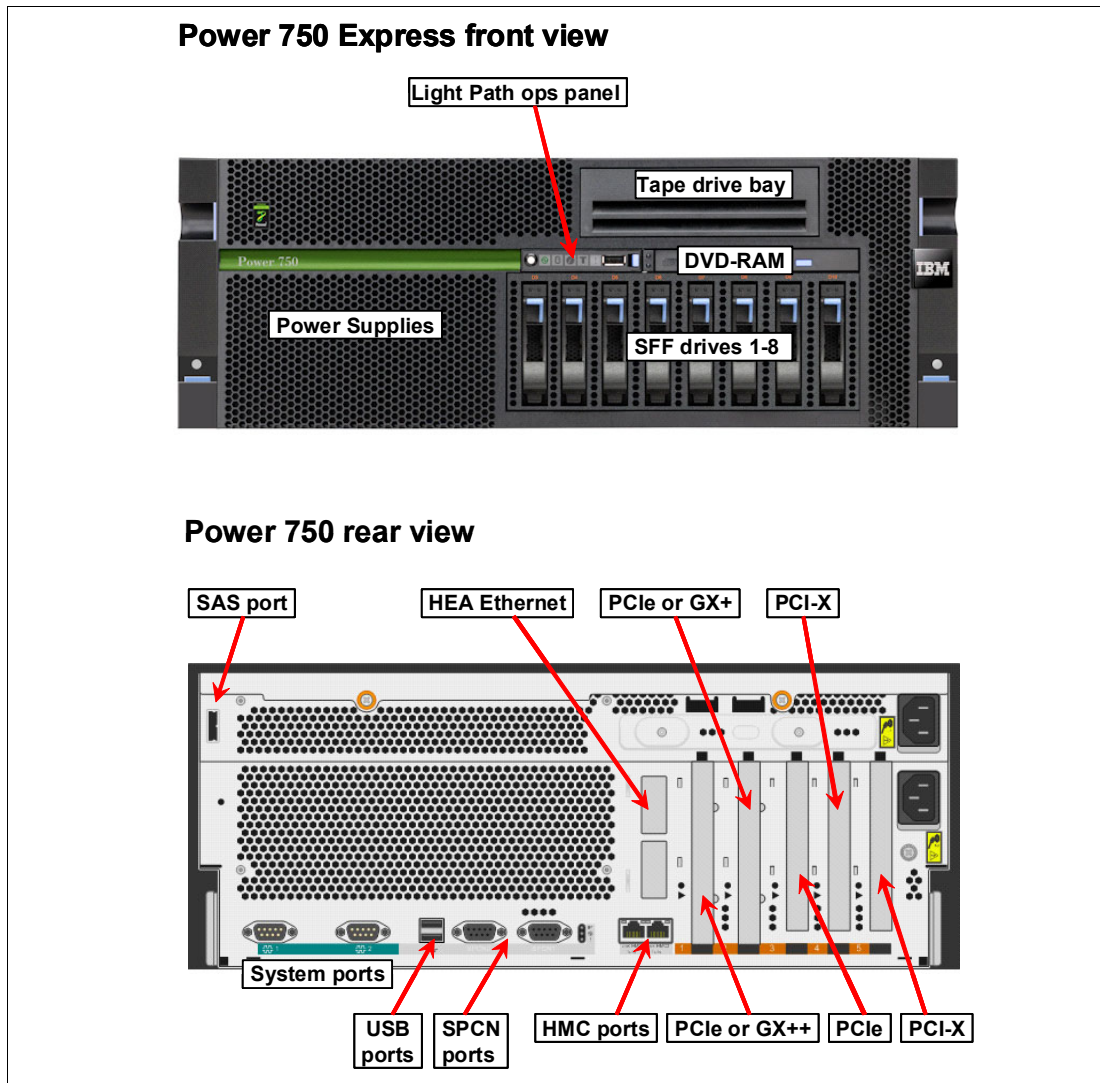


Figure 1-2 Front and rear views of the Power 750 Express system

Table 1-3 on page 8 shows the top view of the Power 750 Express/Power 755 systems. Clearly shown are the processor card locations, fans, and optional (only on Power 750 Express) RAID features.

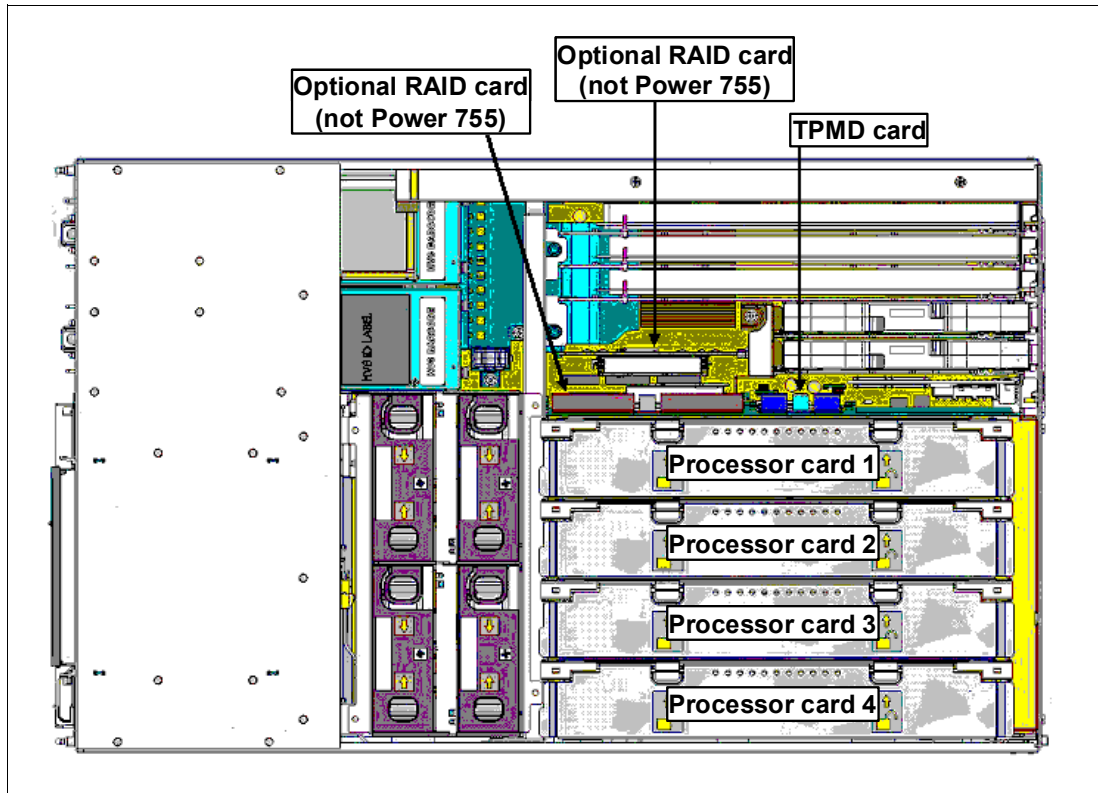


Figure 1-3 Top view of the Power 750 Express/Power 755 systems

1.4 System features

The system chassis contains one to four processor cards. Each card houses a POWER7 processor with either 6-cores or 8-cores. Each of the POWER7 processors in the server has a 64-bit architecture, 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 750 Express system features

The following is a summary of standard features:

- ▶ Rack-mount (4U) configuration
- ▶ Processors:
 - 6-, 12-, 18-, and 24-core design with one, two, three, or four 3.3 GHz processor cards
 - 8-, 16-, 24-, and 32-core design with one, two, three, or four 3.0 GHz or 3.3 GHz processor cards
 - 8-, 16-, 24-, and 32-core design with one, two, three, or four 3.55 GHz processor cards
- ▶ Up to 512 GB of 1066 MHz ECC (error checking and correcting) memory, expandable to 128 GB per processor card
- ▶ 8 x 2.5-inch DASD/SSD/Media backplane with an external SAS port
 - 1 to 8 SFF (Small Form Factor) DASD or solid state drives (mixing allowed)

- ▶ Choice of two Integrated Virtual Ethernet daughter cards:
 - Quad-port 1 Gb IVE
 - Dual-port 10 Gb IVE
- ▶ Two media bays:
 - One slim bay for a DVD-RAM (required)
 - One half-height bay for an optional tape drive or removable disk
- ▶ A maximum of five hot-swap slots:
 - Two PCIe x8 slots, short card length (slots 1 and 2)
 - One PCIe x8 slot, full card length (slot 3)
 - Two PCI-X DDR slots, full card length (slots 4 and 5)
 - One GX+ slot (shares same space as PCIe x8 slot 2)
 - One GX++ slot (shares same space as PCIe x8 slot 1)
- ▶ Integrated:
 - Service Processor
 - Quad-port 10/100/1000 Mb Ethernet
 - EnergyScale technology
 - Hot-swap and redundant cooling
 - Three USB ports; two system ports
 - Two HMC ports; two SPCN ports
- ▶ Two Power Supplies, 1725 Watt AC, Hot-swap

1.4.2 Power 755 system features

The following is a summary of standard features:

- ▶ Rack-mount (4U) configuration
- ▶ 32-core design with four 3.3 GHz processor cards
- ▶ Up to 256 GB of 1066 MHz ECC (error checking and correcting) memory, maximum of 64 GB per processor card
- ▶ 8 x 2.5-inch DASD/SSD/Media backplane with an external SAS port
 - 2 to 8 SFF (Small Form Factor) DASD or Solid® State drives (mixing allowed)
- ▶ Choice of two Integrated Virtual Ethernet daughter cards:
 - Quad-port 1 Gb IVE
 - Dual-port 10 Gb IVE
- ▶ One media bay:
 - Slim bay for a DVD-RAM (required)
- ▶ A maximum of five hot-swap slots:
 - Two PCIe x8 slots, short card length (slots 1 and 2)
 - One PCIe x8 slot, full card length (slot 3)
 - Two PCI-X DDR slots, full card length (Slots 4 and 5)
 - One GX++ slot (shares same space as PCIe x8 slot 1)

- ▶ Integrated:
 - Service Processor
 - Quad-port 10/100/1000 Mb Ethernet
 - EnergyScale technology
 - Hot-swap and redundant cooling
 - Three USB ports; two system ports
 - Two HMC ports; two SPCN ports
- ▶ Two Power Supplies, 1725 Watt AC, Hot-swap

Note: In the Power 755, the GX++ slot and corresponding adapter cannot be used for I/O expansion, it is for node clustering only.

1.4.3 Minimum features

The minimum Power 750 Express configuration must include a processor, processor activations, memory, two power supplies and power cords, one or two DASD, a DASD/SSD/Media backplanes, an operator panel cable, an Ethernet daughter card, a DVD-RAM, an operating system indicator, and a Language Group Specify.

Each system has a minimum feature-set in order to be valid. The minimum system configuration for a Power 750 Express is shown in Table 1-3.

Table 1-3 Minimum features for Power 750 Express system

| Power 750 Express minimum features | Additional notes |
|--|--|
| 1x CEC chassis (4U) | <ul style="list-style-type: none"> ▶ System chassis ▶ DASD/Media backplane with external SAS port, 8 x 2.5 inch-DASD (#8340) ▶ Cable for rack-mount drawer with 2.5 inch-DASD backplane (#1878) ▶ 2x Power Cords (selected by customer) ▶ 2x 1725 watt A/C Power Supply (#7740) ▶ 1x HEA Adapter (one of these): <ul style="list-style-type: none"> – 4-port 1 Gb daughter card (#5624) – Dual-port 10 Gb IVE daughter card (#5613) |
| 1x primary operating system (one of these) | <ul style="list-style-type: none"> ▶ AIX (#2146) ▶ Linux (#2147) ▶ IBM i (#2145) plus IBM i 6.1.1 (#0566 and #0040) |
| 1x Processor Card | <ul style="list-style-type: none"> ▶ 6-core 3.3 GHz POWER7 processor card (#8335) ▶ 8-core 3.0 GHz POWER7 processor card (#8334) ▶ 8-core 3.3 GHz POWER7 processor card (#8332) ▶ 8-core 3.55 GHz POWER7 processor card (#8336) |

| Power 750 Express minimum features | Additional notes |
|---|---|
| 6 or 8 Processor Activations: (all processor cores must be active) | <ul style="list-style-type: none"> ▶ For processor card #8335, one of the following items: <ul style="list-style-type: none"> – 6 x #7717 – 3 x #7717 and 3 x #2327 ▶ For processor card #8334 one of the following items: <ul style="list-style-type: none"> – 8 x #7714 – 4 x #7714 and 4 x #2324 ▶ For use with processor card #8336, one of the following items: <ul style="list-style-type: none"> – 8 x #7716 – 4 x #7716 and 4 x #2326 ▶ For use with 4x processor card #8332, one of the following items: <ul style="list-style-type: none"> – 8 x #7715 – 4 x #7715 and 4 x #2325 |
| 8 GB DDR3 Memory: | ▶ 8 GB (2 x 4 GB) Memory DIMMs, 1066 MHz (#4526) |
| For AIX and Linux: 1x disk drive For IBM i, 2x disk drives | <p>AIX/Linux/Virtual I/O Server:</p> <ul style="list-style-type: none"> ▶ 73.4 GB SAS 2.5-inch 15,000 RPM (#1883) ▶ 146.8 GB SAS 2.5-inch 10,000 RPM (#1882) ▶ 300 GB SAS 2.5-inch 15,000 RPM (#1885) ▶ 69 GB SAS 2.5-inch Solid State Drive (#1890) <p>IBM i</p> <ul style="list-style-type: none"> ▶ 69.7 GB SAS 2.5-inch 15,000 RPM (#1884) ▶ 139.5 GB SAS 2.5-inch 15,000 RPM (#1888) ▶ 69 GB SAS 2.5-inch Solid State Drive (#1909) <p>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.</p> <p>If #0837 (Boot from SAN) is selected a Fibre Channel or Fibre Channel over Ethernet adapter must also be ordered.</p> |
| 1X Language Group (selected by the customer) | No additional notes |
| 1X Removable Media Device | DVD-RAM (#5762) |

The minimum Power 755 configuration must include four processor cards, 32 processor activations, memory, two power supplies and power cords, two DASD, a DASD/SSD/Media backplane, an operator panel cable, an Ethernet daughter card, a DVD-RAM, an operating system indicator, and a Language Group Specify.

Each system has a minimum feature-set in order to be valid. The minimum system configuration for a Power 755 is shown in Table 1-3 on page 8.

Table 1-4 Minimum features for Power 755 system

| Power 755 minimum features | Additional notes |
|---|---|
| 1x CEC chassis (4U) | <ul style="list-style-type: none"> ▶ System chassis ▶ DASD/Media backplane with external SAS port, 8 x 2.5 inch DASD (#8340) ▶ Cable for rack-mount drawer with 2.5 inch DASD backplane (#1878) ▶ 2x Power Cords (selected by customer) ▶ 2x 1725 watt A/C Power Supply (#7740) ▶ 1x HEA Adapter, one of the following items: <ul style="list-style-type: none"> – 4-port 1 Gb IVE daughter card (#5624) – Dual-port 10 Gb IVE daughter card (#5613) |
| 1x primary operating system (one of these) | <ul style="list-style-type: none"> ▶ AIX (#2146) ▶ Linux (#2147) |
| 4x processor cards | ▶ Four processor cards of: 8-core 3.3 GHz POWER7 processor card (#8332) |
| 32 x processor activations: (all processor cores must be active) | 32 zero-priced processor activations (#2325) |
| 128 GB DDR3 Memory | 128 GB minimum memory from one type: <ul style="list-style-type: none"> ▶ 8 GB (2 x 4 GB) Memory DIMMs, 1066 MHz (#4526) ▶ 16 GB (2 x 8 GB) Memory DIMMs, 1066 MHz (#4527) |
| For AIX and Linux: 2 x disk drive | <ul style="list-style-type: none"> ▶ 73.4 GB SAS 2.5-inch 15,000 RPM (#1883) ▶ 146.8 GB SAS 2.5-inch 10,000 RPM (#1882) ▶ 146.8 GB SAS 2.5-inch 15,000 RPM (#1886) ▶ 300 GB SAS 2.5-inch 15,000 RPM (#1885) ▶ 69 GB SAS 2.5-inch Solid State Drive (#1890) <p>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.</p> <p>If #0837 (Boot from SAN) is selected a Fibre Channel or Fibre Channel over Ethernet adapter must also be ordered.</p> |
| 1X Language Group (selected by the customer) | No additional notes |
| 1x Removable Media Device | DVD-RAM(#5762) |

1.4.4 Power supply features

Two system 1725 watt AC power supplies (#7740) are required for the Power 750 Express and Power 755; the second power supply provides redundant power for enhanced system availability. To provide full redundancy, the two power supplies must be connected to separate PDUs.

The server 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.

1.4.5 Processor card features

Each of the possible four processor cards within the system house a single POWER7 processor. The processor has either 6-cores or 8-cores and eight DDR3 DIMM slots. Figure 1-4 shows the processor card.

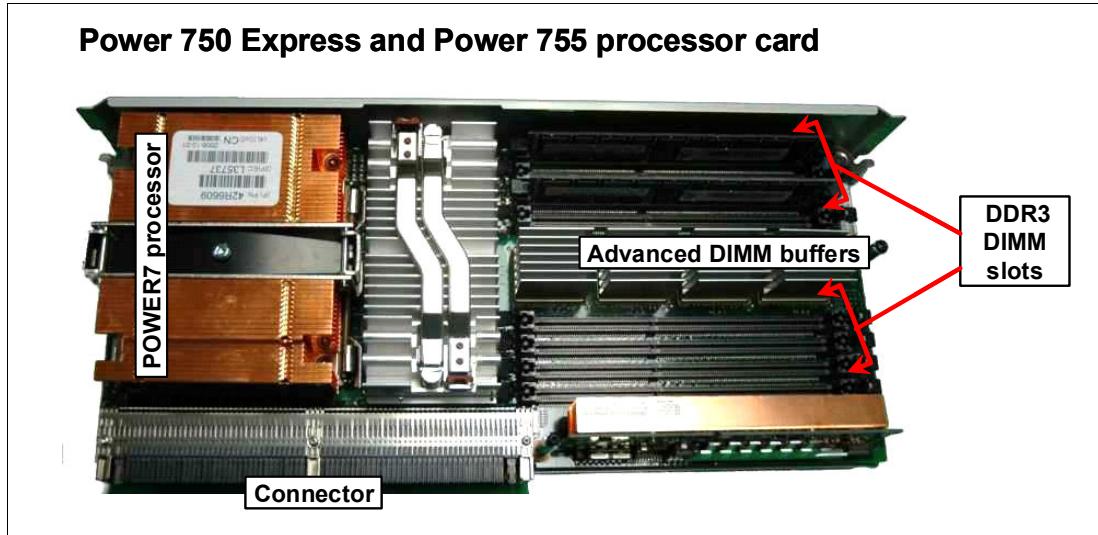


Figure 1-4 Processor card for Power 750 Express and Power 755 systems

Processor features: Power 750 Express

A minimum of one processor card is required on an order, with a maximum of 32 processor cores on four processor cards. One, two, three, or four 6-core 3.3 GHz (#8335), or 8-core 3.0 GHz (#8334)/3.3 GHz (#8332) processor cards may be installed in a system. Four 8-core 3.55 GHz (#8336) processor cards may be installed in a system.

Note: Processor cards (#8332, #8334, #8335, and #8336) may not be mixed in the system and all processor cores in the systems must be activated.

Table 1-5 summarizes the processor features for the power 750 Express.

Table 1-5 Summary of processor features for the Power 750 Express

| Processor card feature | Processor card description | Processor activation | Min./Max. cards |
|--|--------------------------------|---|-----------------|
| 8335 | 6-core 3.3 GHz processor card | The 6-core 3.3 GHz processor card (#8335) requires that six processor activation codes be ordered. Six processor activation code features (6 x #7717, or 3 x #7717 and x #2327) are required per processor card. | 1/4 |
| 8334 | 8-core 3.0 GHz processor card | The 8-core 3.0 GHz processor card (#8334) requires that eight processor activation codes be ordered. Eight processor activation code features (8 x #7714, or 4 x #7714 and 4 x #2324) are required per processor card. | 1/4 |
| 8332 | 8-core 3.3 GHz processor card | The 8-core 3.3 GHz processor card (#8332) requires that eight processor activation codes be ordered. Eight processor activation code features (8 x #7715, or 4 x #7715 and 4 x #2325) are required per processor card. | 1/4 |
| 8336 | 8-core 3.55 GHz processor card | The 8-core 3.55 GHz processor card (#8336) requires that eight processor activation codes be ordered. Eight processor activation code features (8 x #7716, or 4 x #7716 and 4 x #2326) are required per processor card. | 1/4 |
| Note: POWER7 processor features in the system cannot be mixed; all must have the same number of cores and run at the same frequency. | | | |

Processor features: Power 755

A minimum of four processor cards are required on an order. Four 8-core 3.3 GHz (#8332) processor cards are installed in a system.

Note: Only one processor feature type is allowed in a Power 755: #8332. All four processor cards must be ordered so the system is fully populated with processor cards.

Table 1-6 summarizes the processor features for the Power 755.

Table 1-6 Summary of processor features for the Power 755

| Processor card feature | Processor card description | Processor activation | Min./Max. cards |
|------------------------|-------------------------------|--|-----------------|
| 8332 | 8-core 3.3 GHz processor card | The 8-core 3.3 GHz processor card (#8332) requires that eight processor activation codes be ordered. Eight processor activation code features (8 x #7715, or 4 x #7715 and 4 x #2325) are required per processor card. | 4/4 |

1.4.6 Memory features

In POWER7 processor based 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 larger memory configurations.

Figure 1-5 outlines the memory connectivity specific to the Power 750 Express and the Power 755. The four memory channels can be clearly seen.

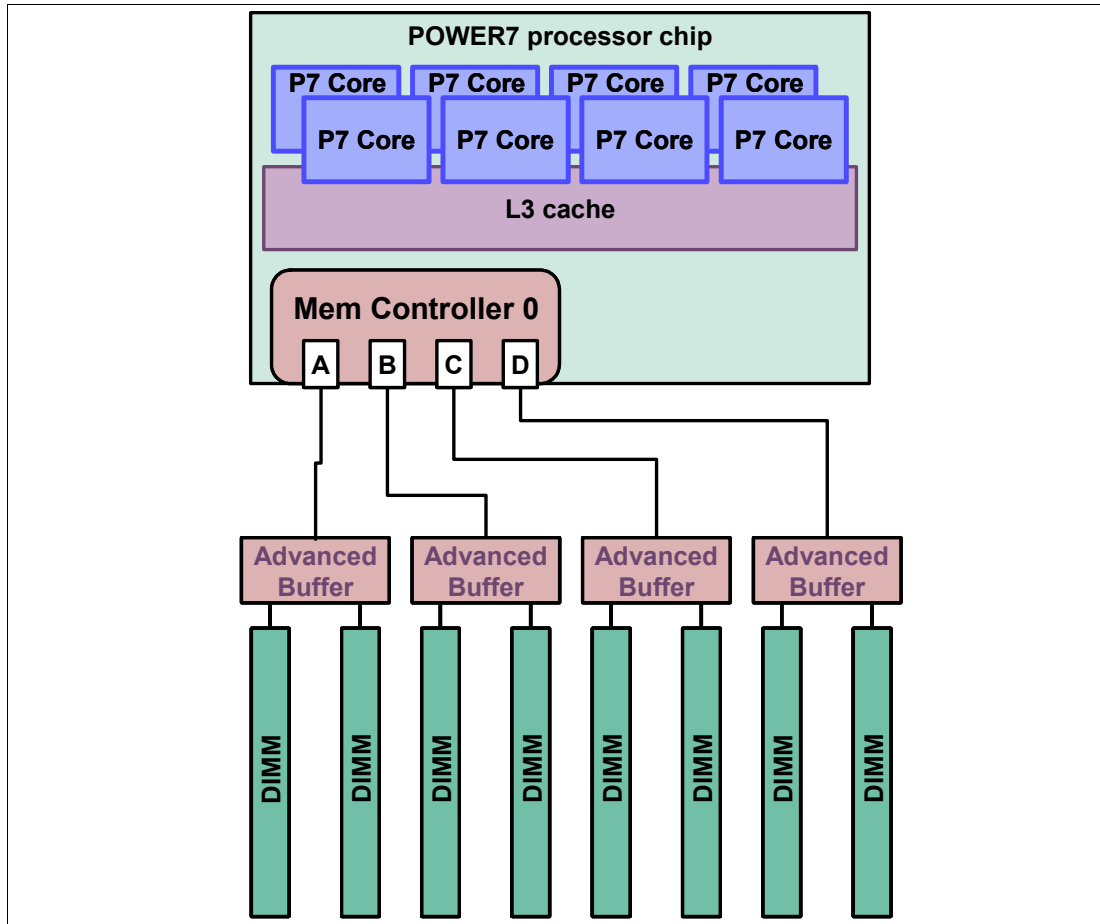


Figure 1-5 Outline of memory connectivity to Power 750 Express and Power 755 DDR3 DIMMs

On each processor card for the Power 750 Express and Power 755 there is a single POWER7 SCM which allows a total of eight DDR3 memory DIMM slots to be connected (eight DIMM slots per processor card).

For the Power 750 Express the DIMM cards can have an 4 GB, 8 GB, or 16 GB capacity. The Power 755 utilizes 4 GB and 8 GB DIMMs. They are connected to the POWER7 processor memory controller through an advanced memory buffer ASIC (application specific integrated circuit).

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

The Power 750 Express has memory features in 8 GB, 16 GB, and 32 GB capacities whereas the Power 755 has memory features in 8 GB and 16 GB capacities. Table 1-7 summarizes the capacities of the memory features and highlights other characteristics.

Mixing different-sized memory features on the same server is possible. For example, the 1 x 16 GB memory feature (#4527) can replace the 2 x 8 GB feature (#4526). However, all memory features on an individual processor card must be identical.

Table 1-7 Summary of memory features

| Feature Code | Feature capacity | Access rate | DIMMs | DIMM slots | Support |
|--------------|------------------|-------------|-----------------|------------|--------------------------------|
| 4526 | 8 GB | 1066 MHz | 2 x 4 GB DIMMs | 2 | Power 750 Express Power 755 |
| 4527 | 16 GB | 1066 MHz | 2 x 8 GB DIMMs | 2 | Power 750 Express Power 755 |
| 4528 | 32 GB | 1066 MHz | 2 x 16 GB DIMMs | 2 | Power 750 Express |

1.5 Disk and media features

Each system features one SAS DASD controller with a maximum of either eight SFF SAS disks or eight SFF SAS SSDs. All disks and SSDs are direct dock and hot pluggable.

Table 1-8 shows the available disk drive feature codes.

Table 1-8 Disk drive feature code description

| Feature code | Description | OS support |
|-------------------|-------------------------------------|------------|
| 1882 | 146.8 GB 10K RPM SAS SFF Disk Drive | AIX, Linux |
| 1883 | 73.4 GB 15K RPM SAS SFF Disk Drive | AIX, Linux |
| 1884 ^a | 69.7 GB 15K RPM SAS SFF Disk Drive | IBM i |
| 1885 | 300 GB 10K RPM SFF SAS Disk Drive | AIX, Linux |
| 1886 | 146 GB 15K RPM SFF SAS Disk Drive | AIX, Linux |
| 1888 ^a | 139 GB 15K RPM SFF SAS Disk Drive | IBM i |
| 1890 | 69 GB SFF SAS Solid State Drive | AIX, Linux |
| 1909 ^a | 69 GB SFF SAS Solid State Drive | IBM i |

a. Supported on Power 750 only.

The Power 750 has a slimline media bay, and a half-high bay that can contain an optional tape drive or removable disk drive equipped with USB Internal Docking Station for Removable Disk Drive (#1103). However, a slimline media bay is available only for a SATA Slimline DVD-RAM Drive (#5762) on the Power 755.

Table 1-9 shows the available media device feature codes for Power 750.

Table 1-9 Media device feature code description for Power 750

| Feature code | Description |
|-------------------|---|
| 5743 ^a | SATA Slimline DVD-ROM Drive |
| 5762 ^b | SATA Slimline DVD-RAM Drive |
| 5746 | Half High 800 GB/1.6 TB LTO4 SAS Tape Drive |
| 5619 | 80/160 GB DAT160 SAS Tape Drive |
| 5661 ^c | DAT320 160/320 GB Tape Drive |
| 1106 | USB 160 GB Removable Disk Drive |
| 1107 | USB 500 GB Removable Disk Drive |

a. Supported, but no longer orderable.

b. Supported on Power 755 also.

c. Requires the internal SAS cable (#3657) with right angle SAS connector.

1.6 I/O drawers for Power 750

The Power 750 has a GX+ and a GX++ slots that are shared with the first two PCIe slots. Optional GX Dual-port 12X Channel Attach (#5609) that plugs into only GX++ slot or The GX Dual-port 12x Channel Attach (#5616) that plug into only GX+ slot are used for I/O Drawer expansion. If more PCI slots are needed, such as to extend the number of LPARs, up to eight PCI-DDR 12X Expansion Drawers (#5796), and up to four 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 drawer (height) and mounts in a 19-inch rack. Feature 5796 takes up half the width of the 4U rack space. Feature 5796 requires the use of a #7314 drawer mounting enclosure. The 4U enclosure (height) 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:

- ▶ 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. Unit #5877 is the same as #5802 except it does not support any disk bays. Unit #5877 can be on the same loop as #5802. Unit #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 various 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 supported is different.

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

| Processor cards | Maximum #5796 drawers | Maximum #5802 and #5877 drawers ^a | Total number of slots | | | |
|-----------------|-----------------------|--|-----------------------|----------------|-----------------|-----------------|
| | | | #5796 | | #5802 and #5877 | |
| | | | PCI-X | PCIe | PCI-X | PCIe |
| One | 4 | 2 | 26 | 2 ^a | 2 | 22 ^a |
| Two | 8 | 4 | 50 | 1 ^b | 2 | 41 ^b |
| Three | 8 | 4 | 50 | 1 ^b | 2 | 41 ^b |
| Four | 8 | 4 | 50 | 1 ^b | 2 | 41 ^b |

a. One PCIe slot is reserved for the GX expansion card.

b. Two PCIe slots are reserved for the GX expansion cards.

1.7 Comparison between models

The Power 750 offers a variety of configuration options where the POWER7 processor has 6-cores at 3.3 GHz, or 8-cores at 3.0 GHz, 3.3 GHz, or 3.55 GHz.

The Power 755 is a specialized systems focusing on high performance computing and offers only one configuration based on an 8-core POWER7 processor running at 3.3 GHz.

The POWER7 processor has 4 MB of on-chip L3 cache per core. For the 6-core version there is 24 MB of L3 cache available, whereas for the 8-core version there is 32 MB of L3 cache available.

Table 1-11 summarizes the processor core options and frequencies, and matches them to the L3 cache sizes.

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

| System | Cores per POWER7 SCM | Frequency (GHz) | L3 cache per POWER7 SCM | Min./Max. cores per system |
|-----------|----------------------|--------------------|-------------------------|----------------------------|
| Power 750 | 6 | 3.3 | 24 MB | 6/24 |
| Power 750 | 8 | 3.0 3.3 3.55 | 32 MB | 8/32 |
| Power 755 | 8 | 3.3 | 32 MB | 32/32 |

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.

Be sure to start with one of several available starting configurations, such as the IBM Editions. These solutions are available at initial system-order time with a starting configuration that is ready to run as is.

1.9 IBM Editions

IBM Editions are available only as initial order for the IBM Power 750.

If you order a Power 750 Express server IBM Edition as defined in this section, you can qualify for half the initial configuration's processor core activations at no addition charge.

The total memory (based on the number of cores) and the quantity or size of disk, SSD, Fibre Channel adapters, or Fibre Channel over Ethernet (FCoE) adapters included with the server are the only features that determine whether a customer is entitled to a processor activation at no additional charge.

When you purchase an IBM Edition, you may purchase an AIX, IBM i, or Linux operating system license, or you may choose to purchase the system with no operating system. The AIX, IBM i, or Linux operating system is processed using a feature number on AIX 5.3 or 6.1, IBM i 6.1.1, and SUSE Linux Enterprise Server. If you choose AIX 5.3 or 6.1 for your primary operating system, you may also order IBM i 6.1.1 and SUSE Linux Enterprise Server. The converse is true if you choose an IBM i or Linux subscription as your primary operating system.

These sample configurations can be changed as needed and still qualify for processor entitlements at no additional charge. However, selection of total memory or DASD, SSD, Fibre Channel, and FCoE adapter quantities, which are three smaller than the totals that are defined as the minimums, disqualifies the order as an IBM Edition and the no-charge processor activations are then removed.

The Edition minimum definition details are as follows:

- ▶ A minimum of 4 GB memory per core is needed to qualify for the IBM Edition.
- ▶ You must meet *one* of the following disk, SSD, FC, FCoE criteria minimums; partial criteria cannot be combined:
 - Two DASD
 - Two SSD
 - Two Fibre Channel adapters
 - Two FCoE adapters

1.10 Model upgrades

The Power 750 is a new serial-number server. There are no upgrades from POWER5™ or POWER6 servers into the Power 750 and 755, which retain the same serial number.

However, excluding RIO and HSL I/O drawers, much of the I/O from the POWER5 or POWER6 server can be reused on the Power 750. Note, however, that Power 750 servers that have only one processor card have a maximum of one I/O loop (one available GX slot). The capability for a second loop or a second GX slot requires two or more processor cards. Feature #5796 and 7314-G30 12X drawers that have PCI-X slots cannot be on the same 12X loop as the newer feature #5802 or #5877 12X I/O drawers which have PCIe slots. Thus, a 12-core, 16-core, or higher configuration with two loops provides more I/O migration flexibility.

In addition to I/O drawers, note these other key points:

- ▶ 15,000 rpm SCSI drives that are 69 GB, or higher, are supported
- ▶ 10,000 rpm SCSI or 15,000 rpm SCSI drives of less than 69 GB are not supported

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 the following functions:

- ▶ 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

Several HMC models are supported to manage systems based on POWER7. Licensed Machine Code Version 7 Revision 710 (#0962) is required to support POWER7 processor-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, but you can also use one of the withdrawn models listed in Table 1-12 on page 19.

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.

Be sure to upgrade the HMC memory to 4 GB if it will manage more than 254 partitions.

1.12 System racks

The Power 750 and 755 and its I/O drawers will mount in the 25U 7014-S25 (#0555), 36U 7014-T00 (#0551), or the 42U 7014-T42 (#0553) rack. These racks are built to the 19-inch EIA standard.

If a system is to be installed in a rack or cabinet other than from IBM, you must ensure that the rack meets requirements that are described in 1.12.10, “OEM rack” on page 23.

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

1.12.1 IBM 7014 Model T00 rack

The 1.8 meter (71-inch) Model T00 is compatible with past and present IBM Power Systems. The T00 rack has the following features:

- ▶ The usable space is 36 EIA units (36U).
- ▶ Optional side panels are removable.
- ▶ A highly perforated front door is an option.
- ▶ Side-to-side mounting hardware for joining multiple racks is an option.
- ▶ Standard business black or optional white color in OEM format is available.
- ▶ Power distribution and weight capacity have been increased.
- ▶ Both AC and DC configurations are supported.

- ▶ 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 1-6 on page 21), but others can fit inside the rack. See 1.12.7, “The AC power distribution unit and rack content” on page 21.
- ▶ Weights are as follows:
 - 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:

- ▶ It has 42 EIA units (42U) 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.)

1.12.3 IBM 7014 Model S25 rack

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

- ▶ It has 25U (EIA units).
- ▶ Weights are:
 - 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 will 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.

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 be different. Rack #0555 is supported, but is 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 be different. Certain 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 that is delivered when you order the 7014-T42 or B42 rack, although the included features might be different. Certain features that are included 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 1-6 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, a good practice is to use fillers in the EIA units occupied by these PDUs to facilitate proper air-flow and ventilation in the rack.

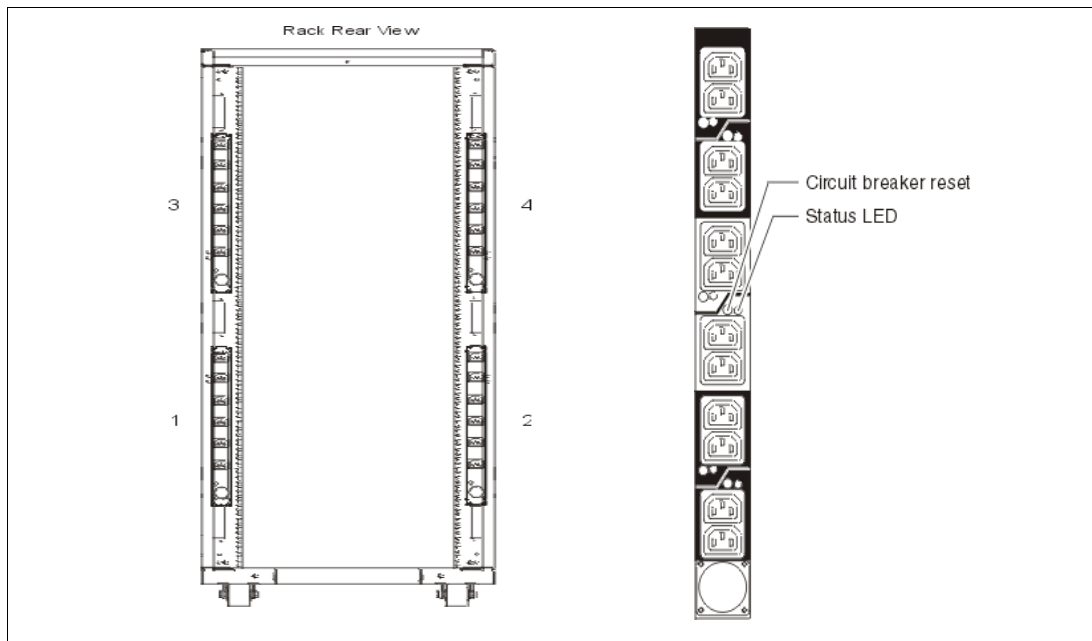


Figure 1-6 PDU placement and PDU view

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

<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 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 A, but each group of two outlets is fed from one 15 A circuit breaker.

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

The Universal PDUs are compatible with previous models.

Note: Each 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 connect each PDU to independent power sources.

1.12.8 Rack-mounting rules

The primary rules to follow when mounting the system into a rack are:

- ▶ The system is designed to be placed at any location in the rack. For rack stability, an advisable approach is to start filling a rack from the bottom.
- ▶ Any remaining space in the rack can be used to install other systems or peripherals, if the maximum permissible weight of the rack is not exceeded and the installation rules for these devices are followed.
- ▶ Before placing the system into the service position, you must follow the rack manufacturer's safety instructions regarding rack stability.

1.12.9 Useful rack additions

This section highlights several solutions that are available 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. It can be configured with one or two tape drives, or 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

1.12.10 OEM rack

The system can be installed in a suitable OEM rack, if the rack conforms to the EIA-310-D standard for 19-inch racks. This standard is published by the Electrical Industries Alliance. For detailed information, see the IBM Power Systems Hardware Information Center:

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

The Web site mentions the following key points:

- ▶ The front rack opening must be 451 mm wide + 0.75 mm (17.75 in. + 0.03 in.), and the rail-mounting holes must be 465 mm + 0.8 mm (18.3 in. + 0.03 in.) apart on center (horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges). Figure 1-7 is a top view showing the specification dimensions.

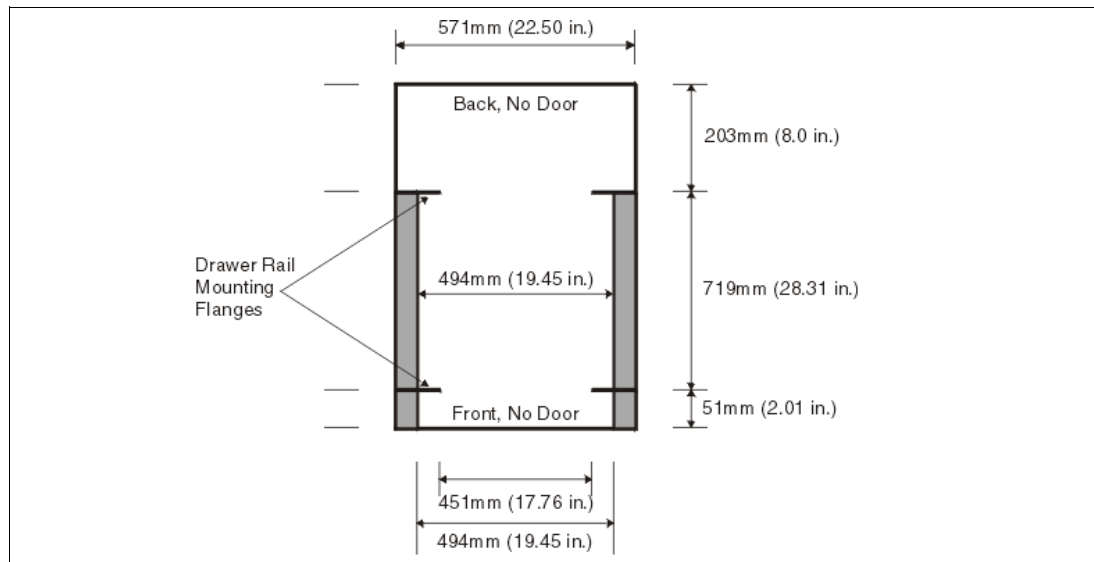


Figure 1-7 Top view of rack specification dimensions (not IBM)

- ▶ The vertical distance between the mounting holes must consist of sets of three holes spaced (from bottom to top) 15.9 mm (0.625 in.), 15.9 mm (0.625 in.), and 12.67 mm (0.5 in.) on center, making each three-hole set of vertical hole spacing 44.45 mm (1.75 in.) apart on center. Rail-mounting holes must be 7.1 mm + 0.1 mm (0.28 in. + 0.004 in.) in diameter. See Figure 1-8 on page 24 and Figure 1-7 for the top and bottom front specification dimensions.

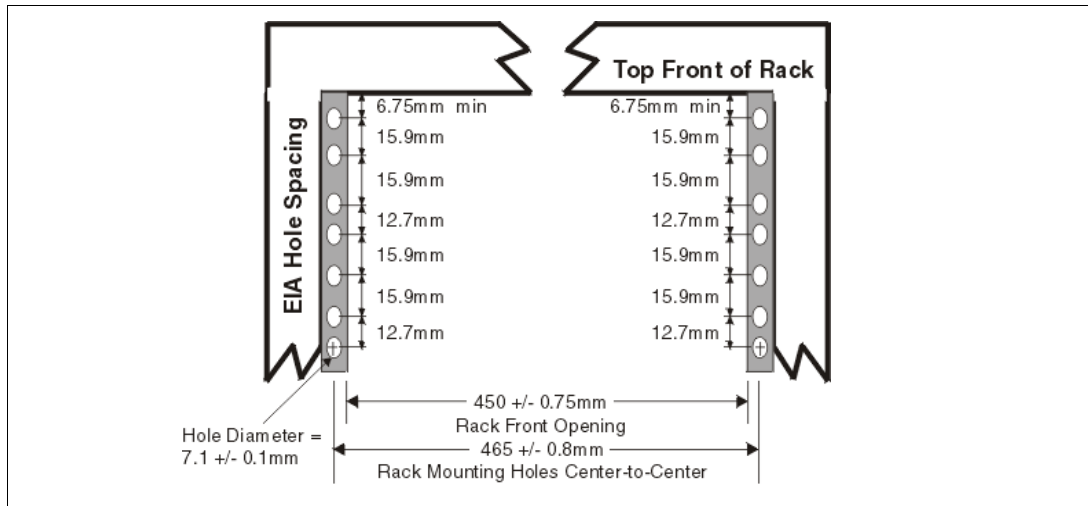


Figure 1-8 Rack specification dimensions, top front view

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.

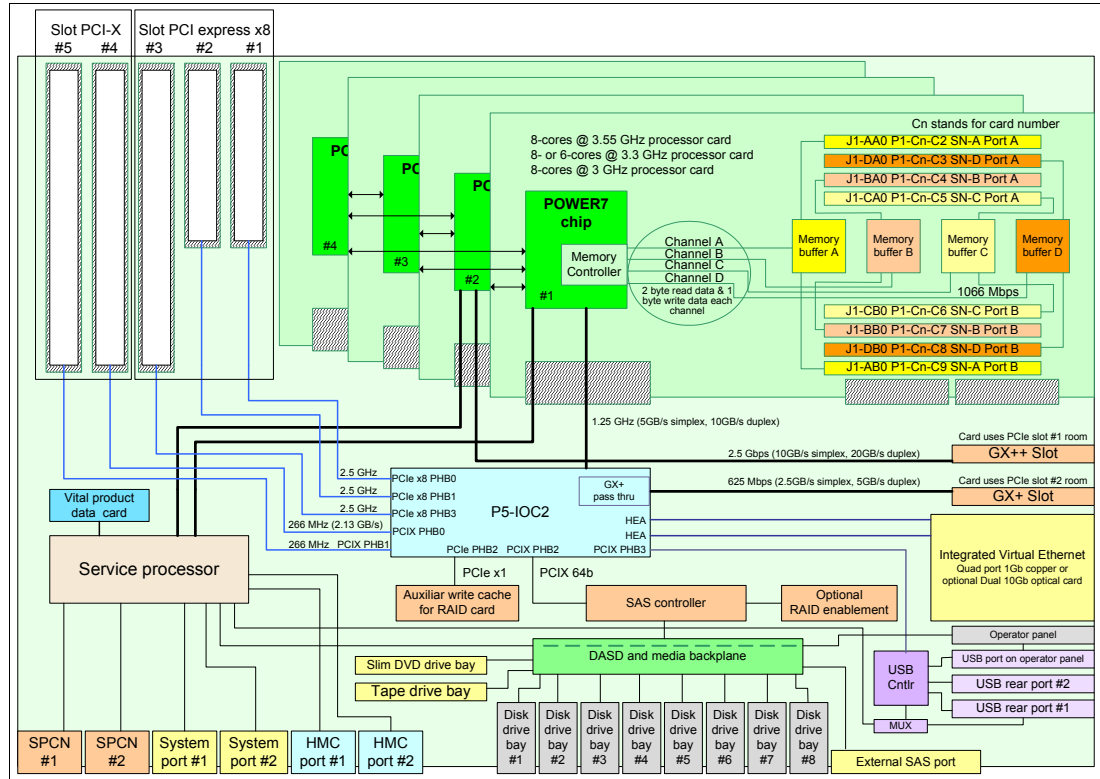


Figure 2-1 Power 750 and Power 755 logical data flow

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 and production workloads.

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 reliability, availability, and serviceability (RAS).

Although the processor is an important component in delivering outstanding servers, many elements and facilities have to be balanced across 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 signalling
- ▶ Exploitation of long-term investment in coherence innovation

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

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

Figure 2-2 on page 28 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.

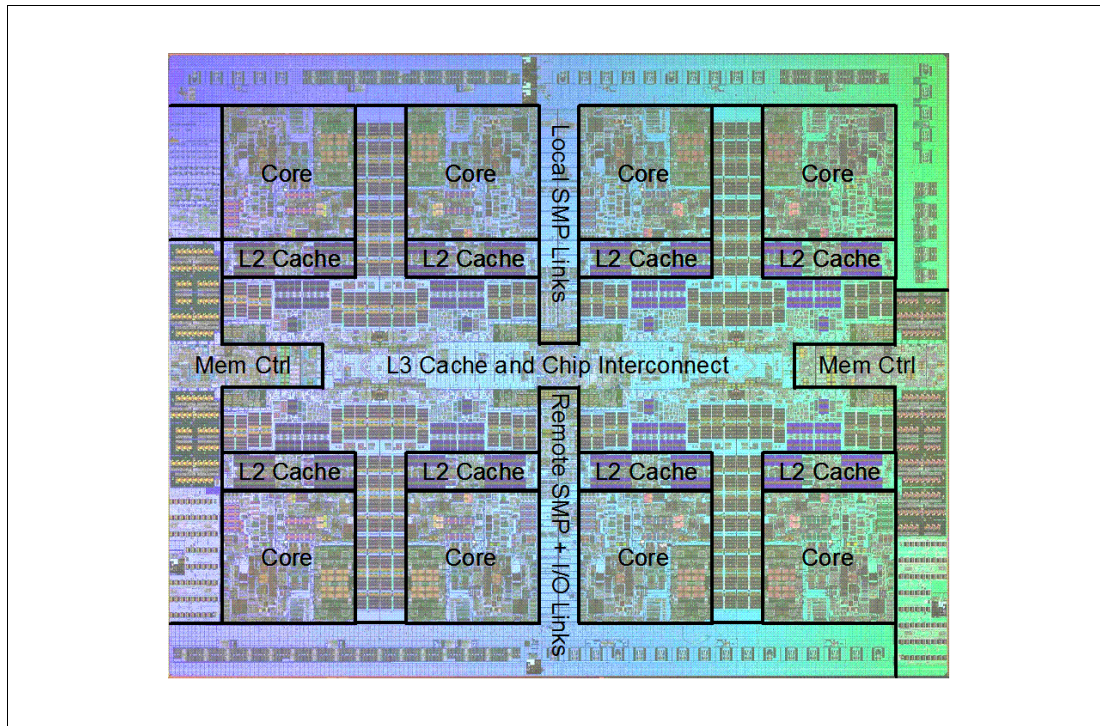


Figure 2-2 POWER7 processor die with key areas indicated

2.1.1 POWER7 processor overview

The POWER7 processor chip is fabricated with 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 leverages 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 32) |
| 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.

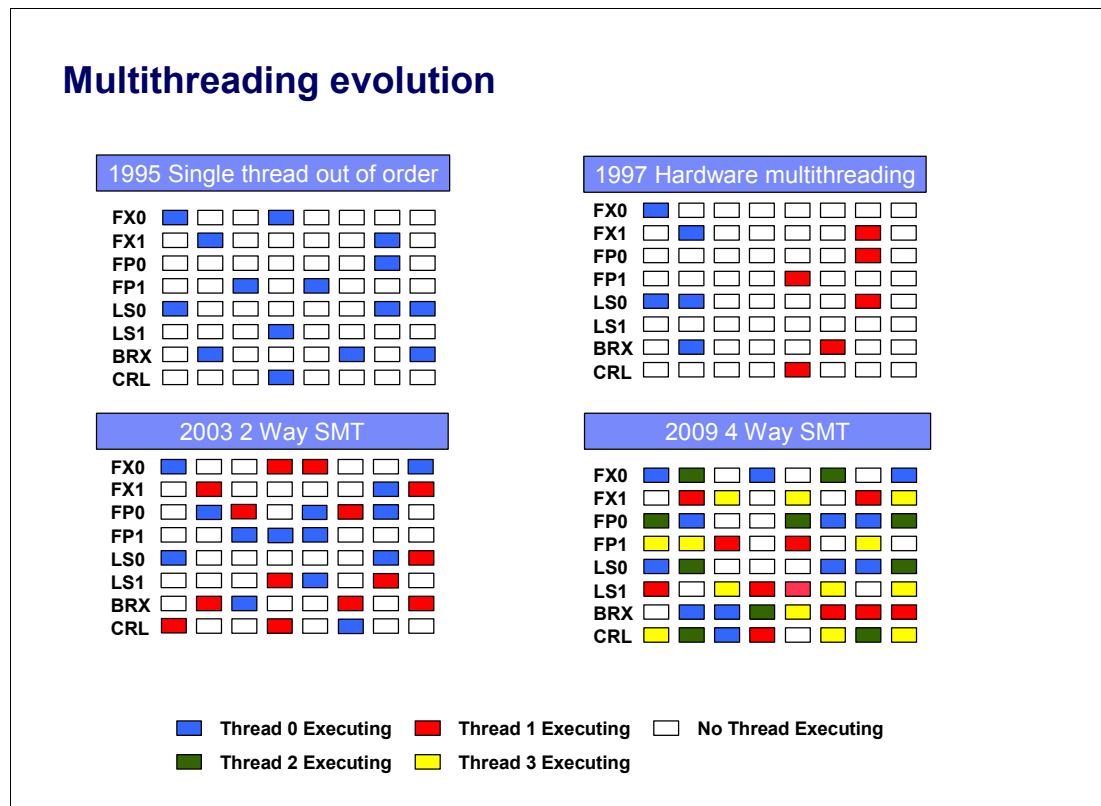


Figure 2-3 Evolution of simultaneous multithreading

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 Gbps 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 some 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.

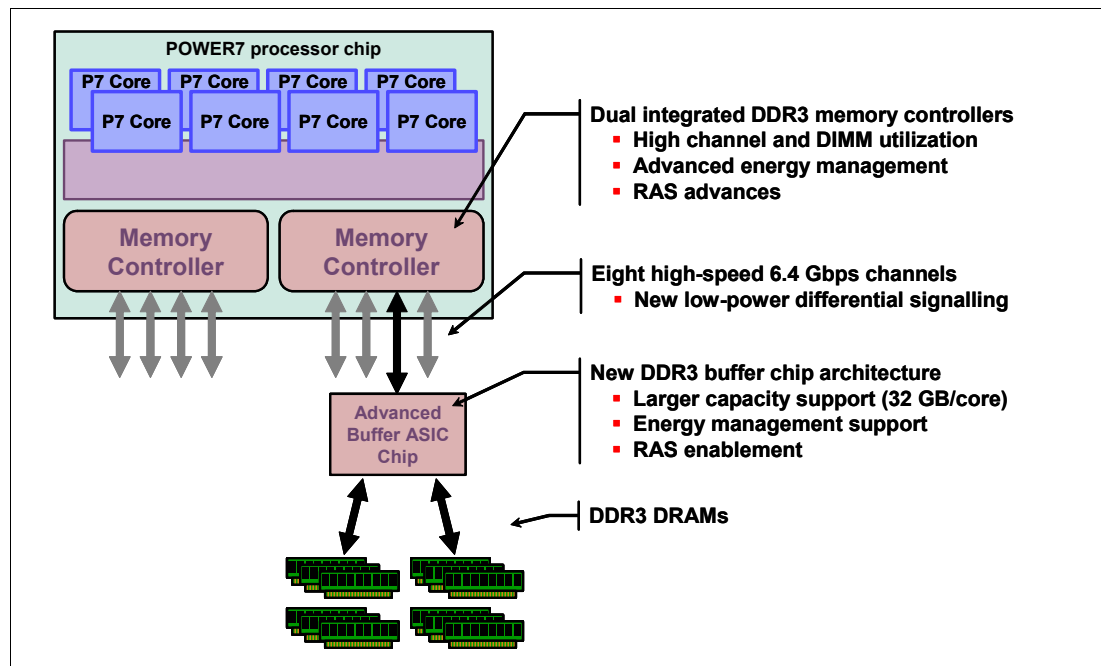


Figure 2-4 Overview of POWER7 memory access structure

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 break 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.

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.

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 appropriate to the scaling-point of particular server models.

Figure 2-5 shows the various physical packaging options that are supported with POWER7 processors.

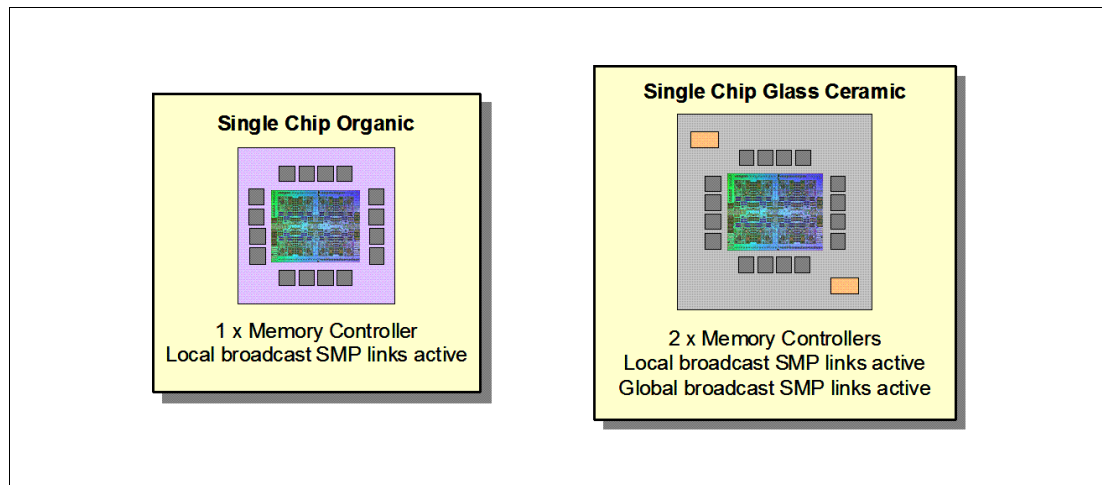


Figure 2-5 Outline of the POWER7 processor physical packaging

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 signalling between the L3 cache and other elements of the hierarchy such as the L2 cache or SMP interconnect.

The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core 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 on page 33 shows the FLR-L3 cache regions for each of the cores on the POWER7 processor die.

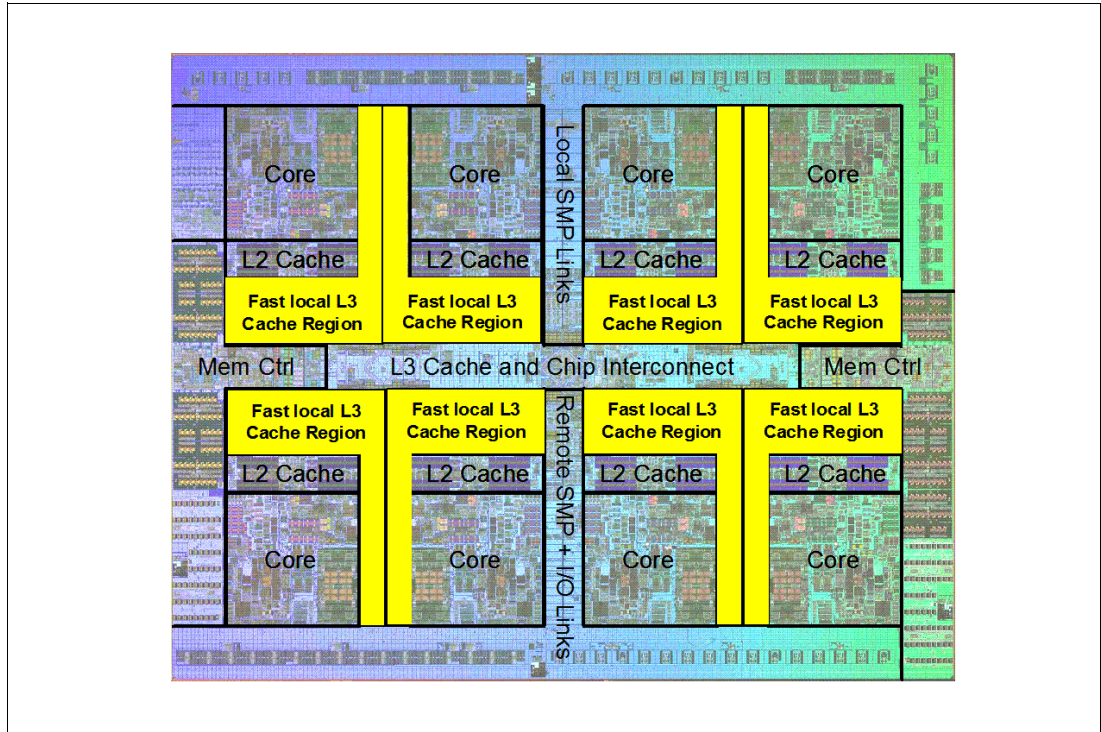


Figure 2-6 Fast Local Regions of L3 cache on the POWER7 processor

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 and 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 six 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 | 567mm ² | 341mm ² | 341mm 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

Each Power 750 Express chassis houses up to four POWER7 processor cards, which hosts one populated POWER7 processor socket (SCM) and eight DDR3 memory DIMM slots. The Power 755, being a high-performance compute node, always houses four POWER7 processor cards each with 8-cores at 3.3 GHz.

Note: All POWER7 processors in the system must be the same frequency and have the same number of processor cores. POWER7 processor types cannot be mixed within a system.

Figure 2-7 shows the processor card highlighting the POWER7 processor socket and the DDR3 DIMM slots.

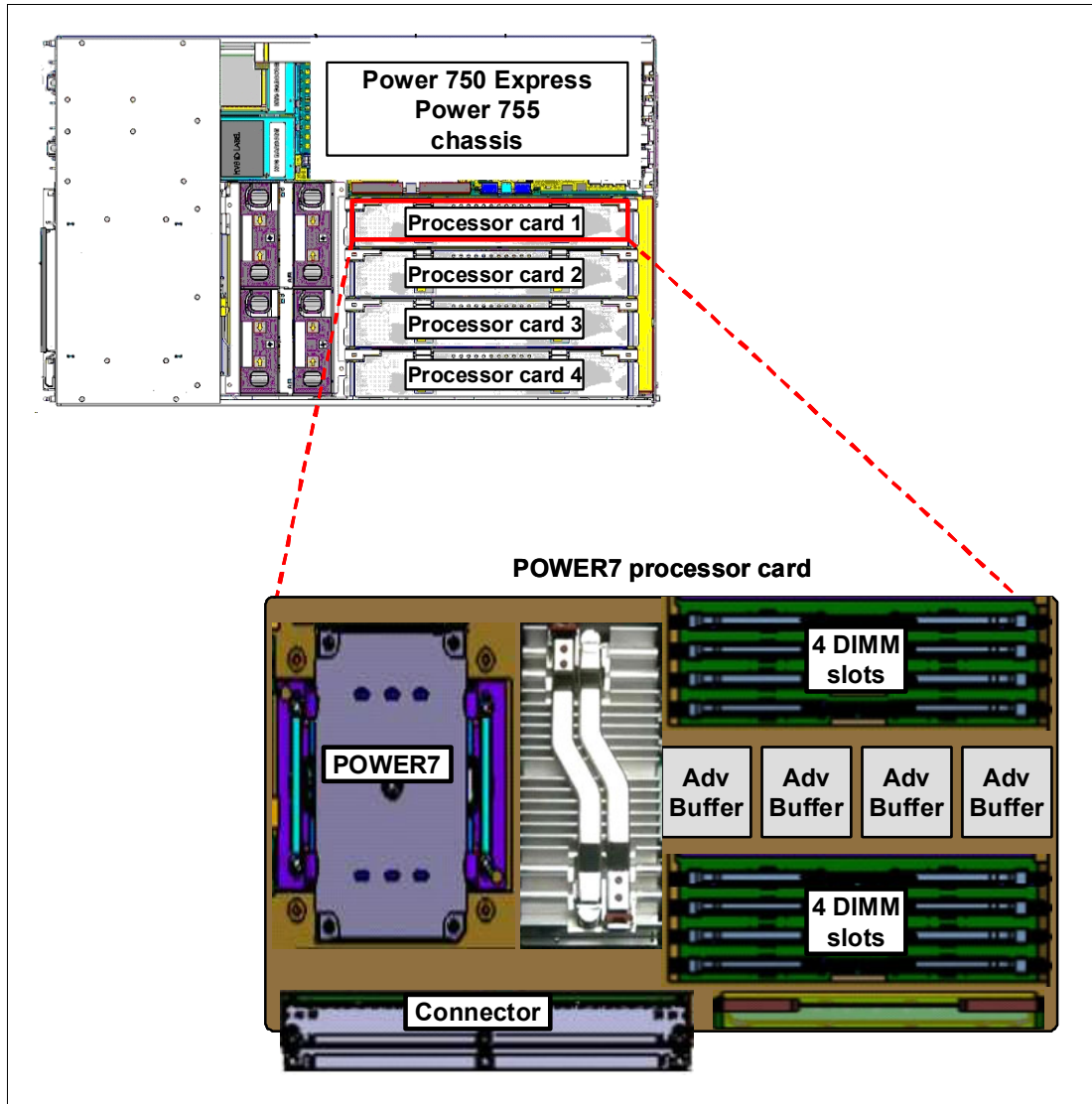


Figure 2-7 POWER7 processor card with processor socket, DIMM slots and advance buffer chips

Processor card slots

There are four processor card slots in the system chassis; slot 1 is on the left when looking at the system from the front view. Processor cards must be installed in sequential order, one to four.

Power 750 systems

Power 750 systems support POWER7 processors with various processor core-counts. Table 2-3 on page 36 summarizes the POWER7 processor options for the Power 750 system.

Table 2-3 Summary of POWER7 processor options for the Power 750 server

| Cores per POWER7 processor | Frequency (GHz) | L3 cache size per POWER7 processor (MB) | Minimum/maximum cores per system | Minimum/maximum processor cards |
|----------------------------|-----------------|---|----------------------------------|---------------------------------|
| 6 | 3.3 | 24 | 6/24 | 1/4 |
| 8 | 3.0 | 32 | 8/32 | 1/4 |
| 8 | 3.3 | 32 | 8/32 | 1/4 |
| 8 ^a | 3.55 | 32 | 32/32 | 4/4 |

a. Available only when all four processor cards installed at initial order (4 x 8-cores at 3.55 GHz).

Power 755 systems

Power 755 systems are for high performance computing environments and only support POWER7 processors with 8-cores at 3.3 GHz only and populated with all four processor cards.

Table 2-4 summarizes the POWER7 processor options for the Power 755 system.

Table 2-4 Summary of POWER7 processor options for the Power 755 server

| Cores per POWER7 processor | Frequency (GHz) | L3 cache size per POWER7 processor (MB) | Min./Max. cores per system | Min./Max. processor cards |
|----------------------------|-----------------|---|----------------------------|---------------------------|
| 8 ^a | 3.3 | 32 | 32/32 | 4/4 |

a. Only available when all four processor cards are installed at initial order (4 x 8-cores at 3.3 GHz)

2.3 Memory subsystem

For Power 750 and Power 755 systems, each of the four processor card houses one POWER7 single chip modules (SCMs). Each POWER7 processor has one on-chip DDR3 memory controller, which can interface with a total of eight DDR3 DIMMs.

2.3.1 Registered DIMM

Industry standard DDR3 Registered DIMM (RDIMM) technology is used to increase reliability, speed, and density of memory subsystems.

2.3.2 Memory placement rules

The minimum DDR3 memory capacity for the Power 750 and Power 755 systems is 8 GB (2 x 4 GB DIMMs). The maximum memory supported is as follows:

- ▶ Power 750: 512 GB (eight DIMMs per processor card 4 x 16 GB per DIMM)
- ▶ Power 755: 256 GB (eight DIMMs per processor card 4 x 8 GB per DIMM)

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

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

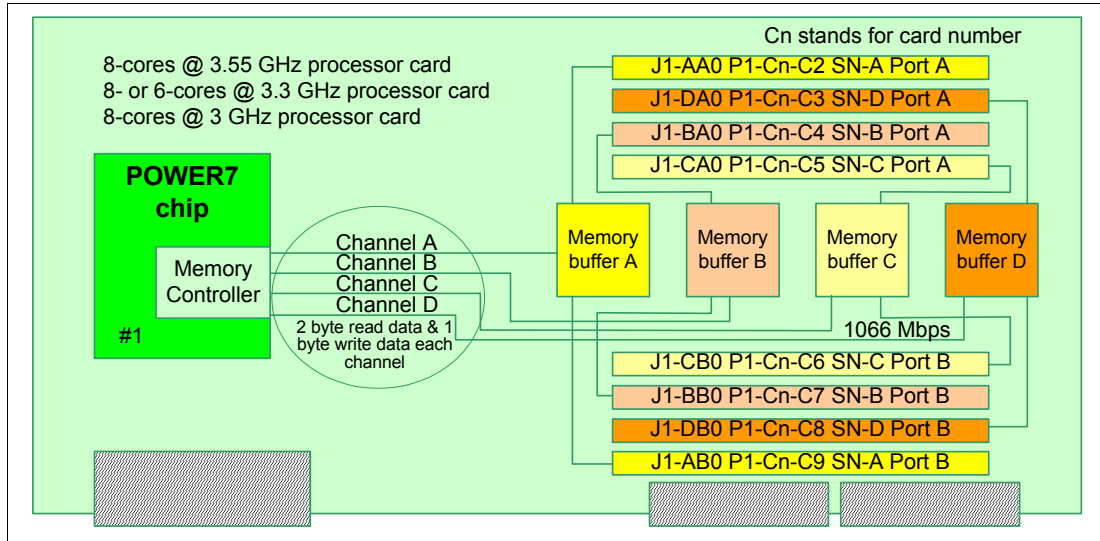


Figure 2-8 Memory DIMM topology for the Power 750 Express and Power 755 processor card

There are eight buffered DIMM slots per processor card; labelled one to eight in Figure 2-8. Note that the slots numbers in the diagram are not consecutive.

The memory-placement rules are as follows:

- ▶ Memory can be pluggable in *DIMM-pairs* (2 x DIMMs), or *DIMM-quads* (4 x DIMMs) for the first two DIMM-pairs or one DIMM-quad.
- ▶ After the second DIMM-pair is installed, all DIMMs must be installed in DIMM-quads (no third DIMM-pair is supported).
- ▶ Minimum memory requirement is for 8 GB (2 x 4 GB DIMMs).
- ▶ All DIMMs on a processor card must be identical (mixing of DIMM capacity and speed is not permitted).

Table 2-5 shows the installation slots for memory DIMMs (slot numbers are identified in Figure 2-8).

Table 2-5 Memory DIMM installation sequence and slots

| Installation slots for DIMM-pairs | | Installation slots for DIMM-quads | |
|-----------------------------------|--|-----------------------------------|--|
| DIMM-pair | Installation slot locations | DIMM-quad | Installation slot locations |
| First | <ul style="list-style-type: none"> ▶ P1-Cn-C2 ▶ P1-Cn-C4 | First | <ul style="list-style-type: none"> ▶ P1-Cn-C2 ▶ P1-Cn-C3 ▶ P1-Cn-C4 ▶ P1-Cn-C5 |
| Second | <ul style="list-style-type: none"> ▶ P1-Cn-C3 ▶ P1-Cn-C5 | Second | <ul style="list-style-type: none"> ▶ P1-Cn-C6 ▶ P1-Cn-C7 ▶ P1-Cn-C8 ▶ P1-Cn-C9 |

Note: No third DIMM-pair is supported. After the second DIMM-pair, memory can only be installed in DIMM-quads.

2.3.3 Memory throughput

POWER7 has exceptional cache, memory, and interconnect bandwidths. Table 2-6 shows the bandwidth estimate for the Power 750 Express and Power 755 systems.

Table 2-6 Power 750 and Power 755 memory and I/O bandwidth estimates

| Memory | | Bandwidth at processor core frequencies | | | |
|--|---------------------------------|---|---|---|---|
| | | Power 750 3.0 GHz | Power 750 3.3 GHz | Power 750 3.55 GHz | Power 755 3.3 GHz |
| L1 (data) cache per core | | 144 GBps | 158.4 GBps | 170.4 GBps | 158.4 GBps |
| L2 cache per core | | 144 GBps | 158.4 GBps | 170.4 GBps | 158.4 GBps |
| L3 cache per core | | 96 GBps | 105.6 GBps | 113.6 GBps | 105.6 GBps |
| System memory per POWER7 socket and system | | 68.22 GBps per socket 272.90 GBps per system | 68.22 GBps per socket 272.90 GBps per system | 68.22 GBps per socket 272.90 GBps per system | 68.22 GBps per socket 272.90 GBps per system |
| GX++ bus | | 20 GBps | 20 GBps | 20 GBps | 20 GBps |
| GX+ bus | | 10 GBps (shared) | 10 GBps (shared) | 10 GBps (shared) | 10 GBps (shared) |
| I/O Bandwidth | GX++ slot 1 | 20 GBps | 20 GBps | 20 GBps | 20 GBps |
| | GX+ bus slot 2 ^a | 5 GBps ^a | 5 GBps ^a | 5 GBps ^a | 5 GBps ^a |
| | Internal I/O slots ^a | 5 GBps ^a | 5 GBps ^a | 5 GBps ^a | 5 GBps ^a |
| | Total I/O bandwidth | 30 GBps | 30 GBps | 30 GBps | 30 GBps |

a. The internal GX+ bus is a pass-through bus and the bandwidth is shared between the internal I/O resources (PCIe, PCI-X, and so on) and the external PCIe/PCI-X drawers that attach to it. For the purposes of this table, 50% of the GX+ bandwidth is consumed by the internal I/O slots.

2.4 Capacity on Demand

Capacity on Demand is not supported on the Power 750 Express or Power 755 system.

2.5 Technical comparison of Power 750 and Power 755

Table 2-7 on page 39 shows a comparison of the technical aspects of the two systems, Power 750 Express and the Power 755 high-performance computing node.

Table 2-7 Comparison of technical characteristics between Power 750 Express and Power 755

| Systems characteristic | Power 750 Express | Power 755 |
|--|---|---|
| Processor | 6-cores at 3.3 GHz 8-cores at 3.0 GHz, 3.3 GHz, 3.55 GHz | 8-cores at 3.3 GHz |
| Pluggable processor cards | 1 - 4 | 4 only |
| Min./Max. processor cores | 6/24 (6-core) or 8/32 (8-core) | 32/32 |
| L3 cache | On-chip eDRAM | On-chip eDRAM |
| Max memory slots and type | 8 slots per processor card (32 slots max.), DDR3 at 1066 MHz | 8 slots per processor card (32 slots max.), DDR3 at 1066 MHz |
| Memory chipkill | Yes | Yes |
| Memory spare | Yes | Yes |
| Memory hotplug | No | No |
| TPMD card | Yes | Yes |
| PCIe x8 slots (in chassis) | 3 | 3 |
| PCI-X 2.0 slots in chassis | 2 | 2 |
| PCIe and PCI-X hot plug | Yes | Yes |
| PowerVM support | Yes | No |
| Capacity on Demand | No | No |
| Redundant hotplug power | Yes | Yes |
| DASD bays (hot-plug, front access, SFF) | 8 | 8 |
| GX slot (GX+ slot does not support RIO2) | 1 x GX+ slot and 1 x GX++ slot (not hot pluggable) | 1 x G++ slot for clustering support only |

2.6 I/O buses and GX card

Each POWER7 processor provides a GX+ bus which is used to connect to an I/O subsystem or Fabric Interface card. The processor card that populates the first processor slot is connected to the GX+ multifunctional host bridge chip, which provides the following major interfaces:

- ▶ One GX+ pass-through bus: This port is used in Power 750 only.
- ▶ 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 ports

The GX+ multifunctional host bridge provide a dedicated GX+ bus routed to the GX+ slot through GX+ pass-through bus. The other GX++ slot is not active unless the second processor card is installed.

Both GX+ and GX++ slots that do not support hot-plug are shared with the first two PCIe slots. Optional GX Dual-port 12X Channel Attach (#5609) that plugs into only GX++ slot or the GX Dual-port 12x Channel Attach (#5616) that plugs into only the GX+ slot is used for I/O Drawer expansion for Power 750 only. The GX Dual-port 12X Channel Attach (#5609) can be used in clustering only on Power 755.

Table 2-8 provides I/O bandwidth of Power 750 3.55 GHz processors configuration.

Table 2-8 I/O bandwidth

| I/O | Bandwidth |
|-----------|---|
| GX++ Bus | 20 GBps |
| GX+ Bus | 10 GBps (Shared with GX+ slot and Internal I/O) |
| Total I/O | 30 GBps |

2.7 Internal I/O subsystem

The internal I/O subsystem resides on the system planar, which supports a mixture of both PCIe and PCI-X slots. All PCIe or PCI-X slots are hot pluggable and Enhanced Error Handling (EEH). PCI EEH-enabled adapters respond to a special data packet 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. For Linux, EEH support extends to the majority of frequently used devices, although various third-party PCI devices might not provide native EEH support.

2.7.1 Slot configuration

Table 2-9 lists the slot configuration of a 750 and 755.

Table 2-9 Slot configuration of a 750 and 755

| Slot# | Description | Location code | PCI host bridge (PHB) | Maximum card size |
|--------|-------------------------------|----------------|-----------------------|-------------------|
| Slot 1 | PCIe x8 GX++ Slot | P1-C1 P1-C7 | PCIe PHB0 | Short |
| Slot 2 | PCIe x8 GX+ Slot | P1-C2 P1-C8 | PCIe PHB1 | Short |
| Slot 3 | PCIe x8 | P1-C3 | PCIe PHB3 | Long |
| Slot 4 | PCI-X DDR, 64-bit, 266 MHz | P1-C4 | PCI-X PHB0 | Long |
| Slot 5 | PCI-X DDR, 64-bit, 266 MHz | P1-C5 | PCI-X PHB1 | Long |

Note the following information:

- ▶ Slot 1 can be used for either a PCIe x8 adapter in connector P1-C1, or a GX++ adapter in connector P1-C7.
- ▶ Slot 2 can be used for either a PCIe x8 adapter in connector P1-C2, or a GX+ adapter in connector P1-C8.

2.7.2 System ports

The system planar has two serial ports that are called *system ports*. When an HMC is connected to the system, the integrated system ports are rendered nonfunctional. In this case, you must install an asynchronous adapter, which is described in Table 2-16 on page 51, for serial port usage. Note the following information:

- ▶ Integrated system ports are not supported under AIX or Linux when the HMC ports are connected to an HMC. Either the HMC ports or the integrated system ports can be used, but not both.
- ▶ The integrated system ports are supported for modem and asynch terminal connections. Any other application using serial ports requires a serial port adapter to be installed in a PCI slot. The integrated system ports do not support IBM HACMP™ configurations.

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. The Integrated Virtual Ethernet adapter card options are:

- ▶ #5613: Dual port (SR) Integrated Virtual Ethernet 10 Gb daughter card
- ▶ #5624: 4-port 1 Gb Integrated Virtual Ethernet daughter card (lowest cost and default card in base system configuration)

2.8.1 IVE features

IVE includes special hardware features to provide logical Ethernet ports that can communicate to logical partitions (LPAR) reducing the use of IBM 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 an physical Ethernet port. LPARs can share IVE ports with improved performance.

Figure 2-9 shows the difference between IVE and SEA implementations.

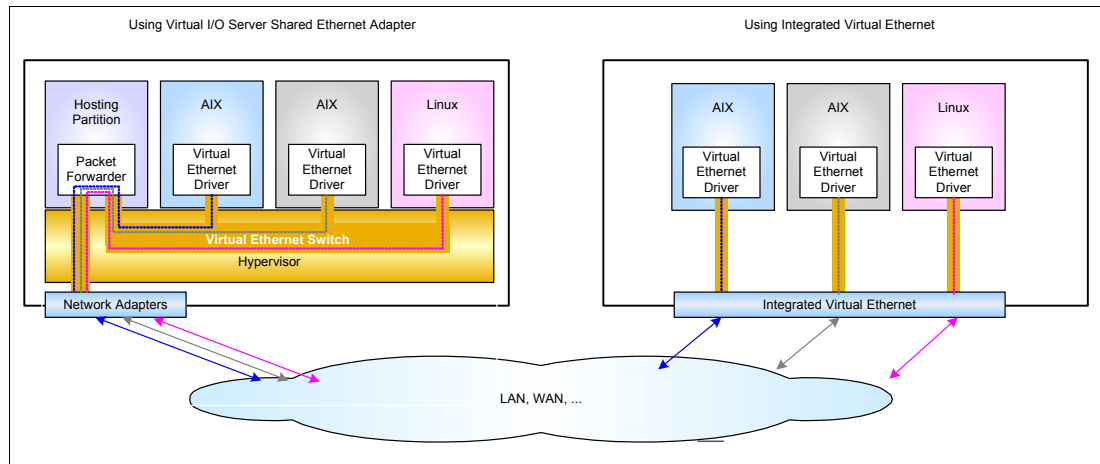


Figure 2-9 Integrated Virtual Ethernet compared to Virtual I/O Server Shared Ethernet Adapter

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

- ▶ Either four 1 Gbps (#5624) or two 10 Gbps ports (#5613)
- ▶ 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 read *Integrated Virtual Ethernet Adapter Technical Overview and Introduction*, REDP-4340.

2.8.2 IVE subsystem

Figure 2-10 shows a high level-logical diagram of the IVE available in the Power 750 server.

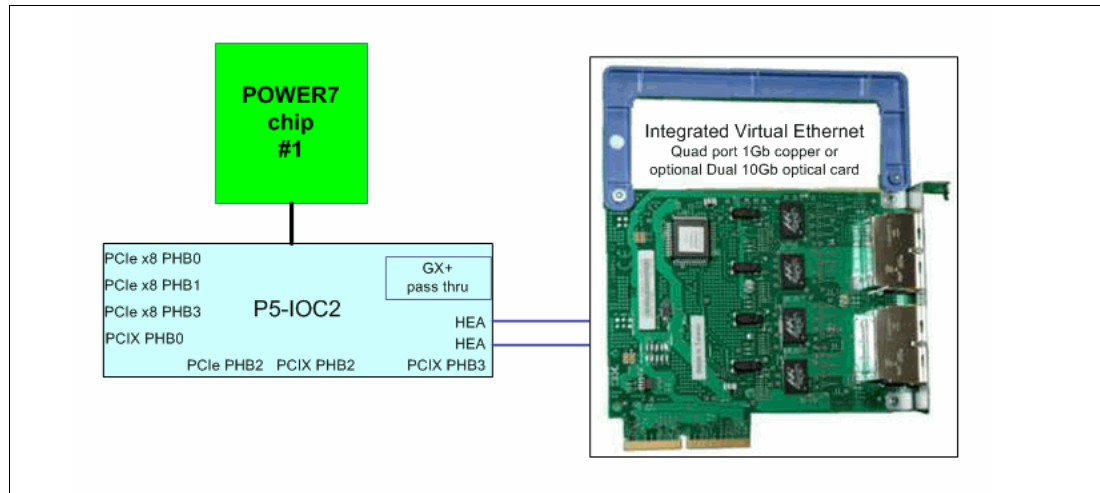


Figure 2-10 IVE system placement

One of the key design goals of the IVE is the capability to integrate up to two 10 Gbps Ethernet ports or four 1 Gbps Ethernet ports into the P5IOC2 chip, with the effect of a low cost Ethernet solution for low-end and mid-range server platforms. Any 10 Gbps, 1 Gbps, 100 Mbps or 10 Mbps speeds share the same I/O pins and do not require additional hardware or feature on top of the IVE card assembly itself.

Two IVE feature code #5624 physical ports are associated to a port group and the other two physical ports are associated to another port group. Any IVE #5613 physical port is associated with a port group. Any port group can address up to 16 logical ports, then each IVE feature code can address up to 32 logical ports. A maximum of sixteen MAC addresses are assigned to any physical port or port group and it is allowed to assign a maximum of one logical port from any physical port to a given LPAR. If the IVE card is replaced, then new IVE card provides a new set of MAC address.

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 using two pairs of wires, one pair for transmit and one pair for receive, and can only transmit 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 is 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 PCIe adapter options for the 750 and 755, as well as PCI and PCI-extended (PCI-X) adapters. All adapters support Extended Error Handling (EEH). 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 may damage the adapter or the slot. A PCI adapter can be installed in a PCI-X slot, and a PCI-X adapter can be installed in a PCI adapter slot. A PCIe adapter cannot be installed in a PCI or PCI-X adapter slot, and a PCI or PCI-X adapter cannot be installed in a PCIe slot.

IBM i IOPs are not supported, which means:

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

SNA applications still can run when encapsulated inside TCP/IP, but the physical device attachment cannot be SNA. It means the earlier Fibre Channel and SCSI controllers which depended upon 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:

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

If you install a new feature, ensure that you have the required software to support the new feature and determine whether any PTF prerequisites must be installed first by checking the IBM Prerequisite Web site:

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) on Power operating systems.

2.9.1 LAN adapters

To connect a Power 750 and Power 755 local area network (LAN), you can use Integrated Virtual Ethernet. Other LAN adapters are supported in the system enclosure PCI slots or in I/O enclosures that are attached to the system using a 12X technology loop.

Table 2-10 lists the additional LAN adapters that are available.

Table 2-10 Available LAN adapters

| Feature code | Adapter description | Slot | Size | OS support |
|--------------------|---|-------|-------|-----------------------|
| 1954 ^{ab} | 4-Port 10/100/1000 Base-TX PCI-X Adapter | PCI-X | Short | A, L |
| 1978 ^{ab} | IBM Gigabit Ethernet-SX PCI-X Adapter | PCI-X | Short | A, L |
| 1979 ^{ab} | IBM 10/100/1000 Base-TX Ethernet PCI-X Adapter | PCI-X | Short | A, L |
| 1983 ^{ab} | IBM 2-Port 10/100/1000 Base-TX Ethernet PCI-X Adapter | PCI-X | Short | A, L |
| 5700 ^{ab} | IBM Gigabit Ethernet-SX PCI-X Adapter | PCI-X | Short | A, i, L |
| 5701 ^{ab} | IBM 10/100/1000 Base-TX Ethernet PCI-X Adapter | PCI-X | Short | A, i, L |
| 5706 | IBM 2-Port 10/100/1000 Base-TX Ethernet PCI-X Adapter | PCI-X | Short | A, i ^c , L |
| 5717 | 4-Port 10/100/1000 Base-TX PCI Express Adapter | PCIe | Short | A, L |
| 5721 ^{ab} | 10 Gb Ethernet-SR PCI-X 2.0 DDR Adapter | PCI-X | Short | A, i, L |
| 5722 ^{ab} | 10 Gb Ethernet-LR PCI-X 2.0 DDR Adapter | PCI-X | Short | A, i, L |
| 5732 | 10 Gigabit Ethernet-CX4 PCI Express Adapter | PCIe | Short | A, L |
| 5740 ^{ab} | 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 ^c , L |
| 5768 | 2-Port Gigabit Ethernet-SX PCI Express Adapter | PCIe | Short | A, i ^c , 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 ^c , L |

a. Supported, but no longer orderable

b. Supported in Power 750 only

c. IBM i operating system is supported in Power 750 only.

2.9.2 Graphics accelerators

The Power 750 support up to eight graphics adapters, and Power 755 support up to three adapters. Table 2-11 on page 46 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 are not support hot pluggable.

Table 2-11 Available Graphics accelerators

| Feature code | Adapter description | Slot | Size | OS support |
|--------------------|---|-------|-------|------------|
| 1980 ^{ab} | POWER GXT135P Graphics Accelerator with Digital Support | PCI-X | Short | A, L |
| 2849 ^{ab} | 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 no longer orderable

b. Supported in Power 750 only

2.9.3 SCSI and SAS adapters

To connect to external SCSI or SAS devices, the adapters listed in Table 2-12 are available to be configured. (Note, in the table, *adjct* is adjacent controller.)

Table 2-12 Available SCSI and SAS adapters

| Feature code | Adapter description | Slot | Size | OS support |
|--------------------|--|-------|-------------|-----------------------|
| 1912 ^{ab} | PCI-X DDR Dual Channel Ultra320 SCSI Adapter | PCI-X | Short | A, L |
| 5736 ^b | PCI-X DDR Dual Channel Ultra320 SCSI Adapter | PCI-X | Short | A, i, L |
| 5778 ^{ab} | PCI-X EXP24 Ctl-1.5 GB No IOP | PCI-X | 2adjct Long | i |
| 5782 ^{ab} | PCI-X EXP24 Ctl-1.5 GB No IOP | PCI-X | 2adjct Long | i |
| 5900 ^{ab} | PCI-X DDR Dual -x4 SAS Adapter | PCI-X | Short | A, L |
| 5901 | PCIe Dual-x4 SAS Adapter | PCIe | Short | A, i ^c , L |
| 5902 ^{ab} | PCI-X DDR Dual - x4 3 Gb SAS RAID Adapter | PCI-X | Long | A, L |
| 5903 | PCIe 380MB Cache Dual - x4 3 Gb SAS RAID Adapter | PCIe | Short | A, i, L |
| 5904 ^b | PCI-X DDR 1.5 GB Cache SAS RAID Adapter | PCI-X | Long | A, i, L |
| 5908 ^b | PCI-X DDR 1.5 GB Cache SAS RAID Adapter (BSC) | PCI-X | Long | A, i, L |
| 5912 ^{ab} | PCI-X DDR Dual - x4 SAS Adapter | PCI-X | Short | A, i, L |

a. Supported, but no longer orderable

b. Supported in Power 750 only

c. IBM i operating system is supported in Power 750 only.

Table 2-13 compares Parallel SCSI to SAS.

Table 2-13 Comparison of 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 degraded as devices added to shared bus | 3 Gbps, scalable to 12 Gbps, performance maintained as more devices added |
| Scalability | 15 drives | Over 16,000 drives |
| Compatibility | Incompatible with all other drive interfaces | Compatible with Serial ATA (SATA) |
| Maximum 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 2-14 provides the orderable iSCSI adapters.

Table 2-14 Available iSCSI adapters

| Feature code | Adapter description | Slot | Size | OS support |
|--------------------|--|-------|-------|-----------------------|
| 1986 ^{ab} | 1 Gigabit iSCSI TOE PCI-X on Copper Media Adapter | PCI-X | Short | A, L |
| 1987 ^{ab} | 1 Gigabit iSCSI TOE PCI-X on Optical Media Adapter | PCI-X | Short | A, L |
| 5714 ^{ab} | 1 Gigabit iSCSI TOE PCI-X on Optical Media Adapter | PCI-X | Short | A, i, L |
| 5713 | 1 Gigabit iSCSI TOE PCI-X on Copper Media Adapter | PCI-X | Short | A, i ^c , L |

a. Supported, but is no longer orderable.

b. Supported in Power 750 only

c. IBM i operating system is supported in Power 750 only.

2.9.5 Fibre Channel adapter

The Power 750 and Power 755 support direct or SAN connection to devices that use Fibre Channel adapters. Table 2-15 on page 49 provides a summary of the available Fibre Channel adapters.

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

Table 2-15 Available Fibre Channel adapters

| Feature code | Adapter description | Slot | Size | OS support |
|--------------------|---|-------|-------|-----------------------|
| 1905 ^{ab} | 4 Gigabit Single-Port Fibre Channel PCI-X 2.0 DDR Adapter | PCI-X | Short | A, L |
| 1910 ^{ab} | 4 Gigabit Dual-Port Fibre Channel PCI-X 2.0 DDR Adapter | PCI-X | Short | A, L |
| 1977 ^{ab} | 2 Gigabit Fibre Channel PCI-X Adapter | PCI-X | Short | A, L |
| 5716 ^{ab} | 2 Gigabit Fibre Channel PCI-X Adapter | PCI-X | Short | A, i, L |
| 5735 ^c | 8 Gigabit PCI Express Dual Port Fibre Channel Adapter | PCIe | Short | A, i ^d , L |
| 5749 ^b | 4 Gigabit Fibre Channel (2-Port) | PCI-X | Short | i |
| 5758 ^{ab} | 4 Gigabit Single-Port Fibre Channel PCI-X 2.0 DDR Adapter | PCI-X | Short | A, L |
| 5759 | 4 Gigabit Dual-Port Fibre Channel PCI-X 2.0 DDR Adapter | PCI-X | Short | A, L |
| 5773 ^{ab} | 4 Gigabit PCI Express Single Port Fibre Channel Adapter | PCIe | Short | A, L |
| 5774 | 4 Gigabit PCI Express Dual Port Fibre Channel Adapter | PCIe | Short | A, i ^d , L |

a. Supported, but no longer orderable

b. Supported in Power 750 only

c. N_Port ID Virtualization (NPIV) capability is supported through the Virtual I/O Server.

d. IBM i operating system is supported in Power 750 only.

2.9.6 Fibre Channel over Ethernet (FCoE)

A new emerging protocol emerging, Fibre Channel over Ethernet (FCoE), is being developed within T11 as part of the Fibre Channel Backbone 5 (FC-BB-5) project; it 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 continue to maintain the resiliency, efficiency, and seamlessness of the existing FC-based data center.

Figure 2-11 on page 50 shows a comparison between existing FC and network connection and FCoE connection.

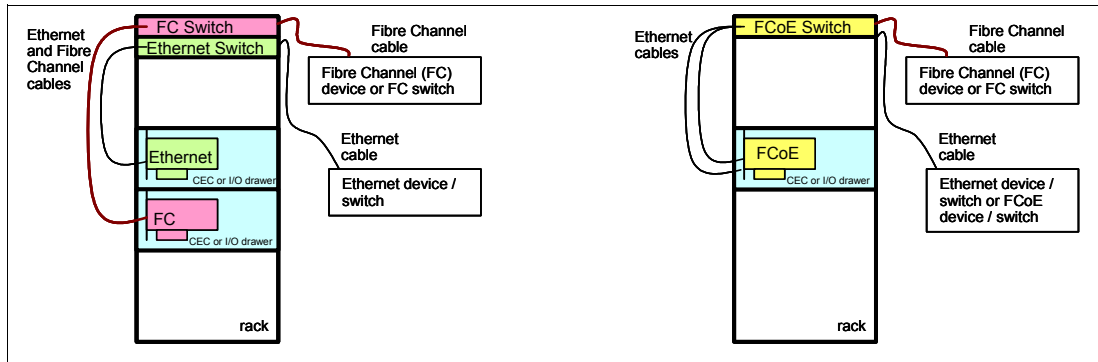


Figure 2-11 comparison between existing FC and network connection and FCoE connection

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

IBM offers 10 Gb FCoE PCIe Dual Port Adapter (#5708) that is a high performance, 10 Gb, dual port, PCIe Converged Network Adapter (CNA), using 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 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:

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

InfiniBand is based on a switched fabric architecture of serial point-to-point links. These IB links can be connected to either host channel adapters (HCAs), used primarily in servers, or 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 bi-directional connection. Combinations of link width and byte lane speed allow for overall link speeds of 2.5 - 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 Dual-port 12X Channel Attach (#5609) that plugs into a GX++ slot in the system enclosure and The GX Dual-port 12x Channel Attach (#5616) that plug into a GX bus slot in a system enclosure. The Power 755 support only #5609 adapter that provides up to twice the data rate capability as the #5616 adapter. Connection to supported infiniband switches is accomplished by using the 12X to 4X Channel Conversion Cables.

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 2-16.

Table 2-16 Available Synchronous adapter

| Feature code | Adapter description | Slot | Size | OS support |
|--------------------|---|-------|-------|------------|
| 2943 ^{ab} | 8-Port Asynchronous Adapter EIA-232/RS-422 | PCI-X | Short | A |
| 5723 ^{ab} | 2-Port Asynchronous EIA-232 PCI Adapter | PCI-X | Short | A, L |
| 5785 | 4 Port Asynchronous EIA-232 PCIe Adapter | PCIe | Short | A, L |

a. Supported, but no longer orderable

b. Supported in Power 750 only

2.10 Internal storage

Power 750 uses an integrated SAS/SATA controller to support PCI-X 2.0 64-bit operation; however, it is connected to a 133 MHz PCI-X bus on the P5-IOC2 chip (see Figure 2-12 on page 52). The SAS/SATA controller used in a Power 750 enclosure has two sets of four SAS/SATA channels. Each channel can support either SAS or SATA operation. SAS controller is then connected to a DASD backplane and supports eight Small Form Factor (SFF) disk drive bays. The disk drives supported in Power 750 engage the connection with the DASD backplane, directly dock and are hot-pluggable. Enclosure Services hot plug control function is performed by the port expander chip on the DASD backplane.

The DASD backplane supports connections for a slim SATA DVD and an SAS 5.25 inch or 3.5 inch half-high tape drive. It also provides four of the SAS controller ports going to a rear bulkhead connector for support of an external storage drawer.

Without a split backplane, SSDs and HDDs may be mixed in any combination.

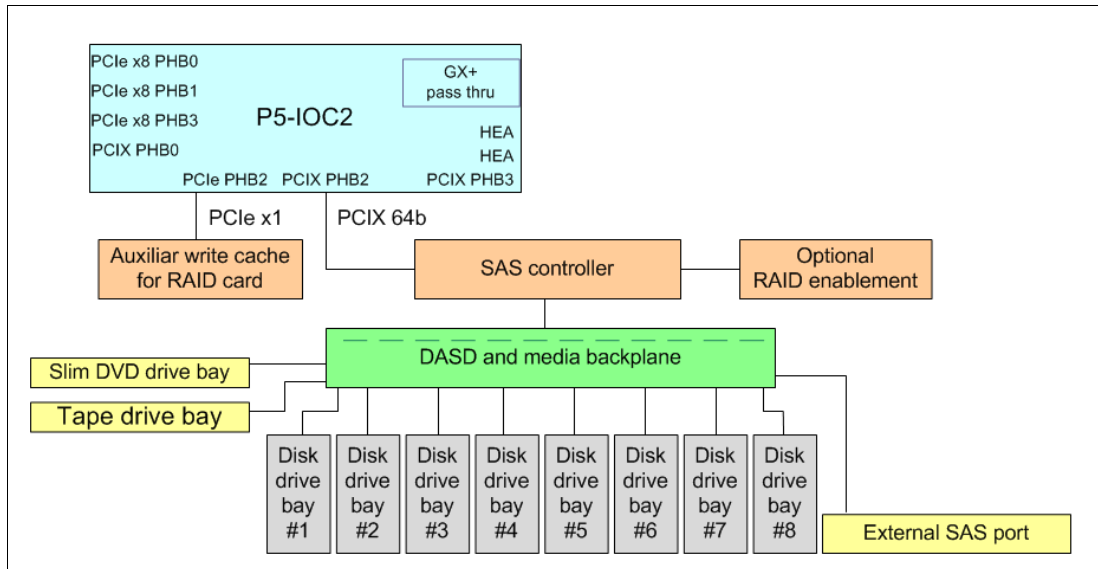


Figure 2-12 Internal storage topology overview

2.10.1 Dual-write cache RAID feature

Power 750 supports a dual-write cache RAID feature, which consists of an auxiliary write-cache for RAID card and the optional RAID enablement. The auxiliary-write cache talks to the operating system, using the PCIe 1x interface for setup and error reporting. The RAID enablement feature is connected to the SAS controller. In a normal operation scenario, all data in the primary-write cache is mirrored in the secondary cache, from PCI-X bus to PCIe bus through two SAS lane available with the P5-IOC2 chip connection.

Supported RAID functions

The RAID enablement feature and the auxiliary-write cache are an optional feature. When features are configured, Power 750 supports hardware RAID 0, 1, 5, 6, and 10, as follows:

- ▶ RAID-0 provides striping for performance, but does not offer any fault tolerance.

The failure of a single drive results in the loss of all data on the array. This process increases I/O bandwidth by simultaneously accessing multiple data paths.
- ▶ RAID-5 uses block-level data striping with distributed parity.

RAID-5 stripes both data and parity information across three or more drives. Fault tolerance is maintained by ensuring that the parity information for any given block of data is placed on a drive separate from those used to store the data itself. The performance of a RAID-5 array can be adjusted by trying various stripe sizes until one is found that is well-matched to the application being used.
- ▶ RAID-6 uses block-level data striping with dual distributed parity.

RAID-6 is the same as RAID-5 except that it uses a second level of independently calculated and distributed parity information for additional fault tolerance. RAID-6 configuration requires N+2 drives to accommodate the additional parity data, which makes it less cost effective than RAID-5 for equivalent storage capacity
- ▶ RAID-10 is also known as a stripe set of mirrored arrays.

RAID-10 is a combination of RAID-0 and RAID-1. A RAID-0 stripe set of the data is created across a 2-disk array for performance benefits. A duplicate of the first stripe set is then mirrored on another 2-disk array for fault tolerance

2.10.2 External SAS port

Power 750 DASD backplane offers the connection to an external SAS port. It can be used to connect external SAS devices or I/O drawer but also to enable the split DASD backplane option.

2.10.3 Split DASD backplane feature

Power 750 DASD backplane supports split mode. If #3669 internal SAS cable is configured, the four small form factor (SFF) disk drives on the left (in the front view) are assigned to the integrated SAS controller, and the four SFF disk drives on the right are assigned to the external rear SAS port. Figure 2-13 helps to show how a PCIe or PCI-X SAS adapter (#5901) can access the right four SFF disk drives through an external SAS cable.

Without a split backplane, SSDs and HDDs may be mixed in any combination.

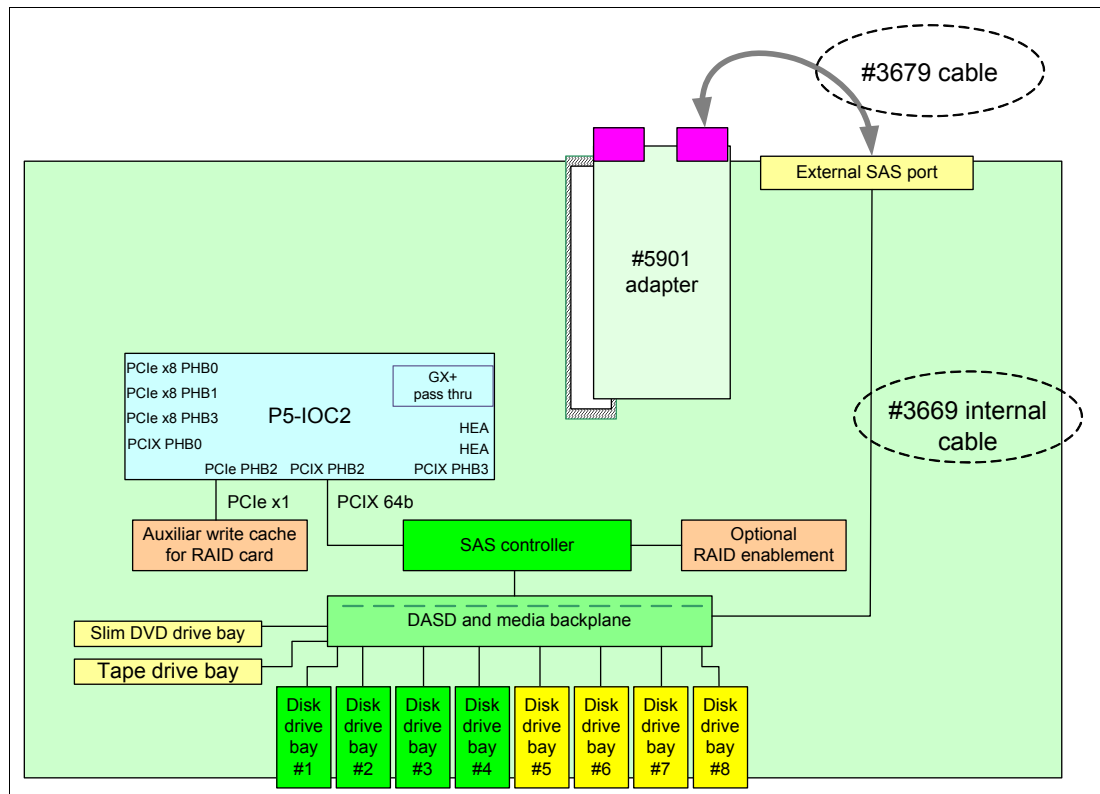


Figure 2-13 Split DASD backplane feature

Note: When a split-DASD-mode SAS cable #3669 is configured, it replaces internal cable #3668.

2.10.4 Media bays

Power 750 offers a slim media bay to support slim SATA DVD device. Direct dock and hot-plug of DVD media device is supported. Also, the half-high bay is available to support an optional SAS tape drive that cannot be hot-pluggable.

2.11 External I/O subsystems

This section describes the external I/O subsystems, which include the I/O drawers, the TotalStorage® EXP24 Disk Drawer (#5786), as well as 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). The #5886 is the only external I/O drawer option for Power 755.

Table 2-17 lists all the supported I/O drawers for Power 750.

Table 2-17 I/O drawer capabilities

| Drawer FC | DASD | PCI Slots | Requirements for a 750 |
|-------------------|---------------------------|-----------|--------------------------------|
| 5796 | - | 6 x PCI-X | GX adapter card #5609 or #5616 |
| 5802 | 18 x SAS disk drive bays | 10 x PCIe | GX adapter card #5609 or #5616 |
| 5877 | - | 10 x PCIe | GX adapter card #5609 or #5616 |
| 5786 ^a | 24 x SCSI disk drive bays | - | Any supported SCSI adapter |
| 5886 ^b | 12 x SAS disk drive bays | - | Any supported SAS adapter |

a. Supported, but no longer orderable

b. Supported in both Power 750 and Power 755

Both GX+ and GX++ slots are shared with the first two PCIe slots. Optional GX Dual-port 12X Channel Attach (#5609) that plugs into only GX++ slot or The GX Dual-port 12x Channel Attach (#5616) that plug into only GX+ slot are used for I/O Drawer expansion for Power 750 only. The GX++ slot is not active unless the second processor card is installed.

The PCI-DDR 12X Expansion Drawer (#5796) is operated with SDR speed, no matter which GX adapter is used. However, the 12X I/O Drawer PCIe (#5802 and #5877) is operated with higher capacity bandwidth (DDR) speed, in case it is attached to GX Dual-port 12X Channel Attache (#5609).

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

The PCI-DDR 12X Expansion Drawer (#5796) is a 4U-high (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-tall (EIA units) 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 that use blind swap cassettes and support hot plugging of adapter cards. The drawer includes redundant hot-plug power and cooling.

There are two available interface adapters for use in the #5796 drawer. The Dual-Port 12X Channel Attach Adapter Long Run (#6457) or the 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 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.

Figure 2-14 shows the back view of the expansion unit.

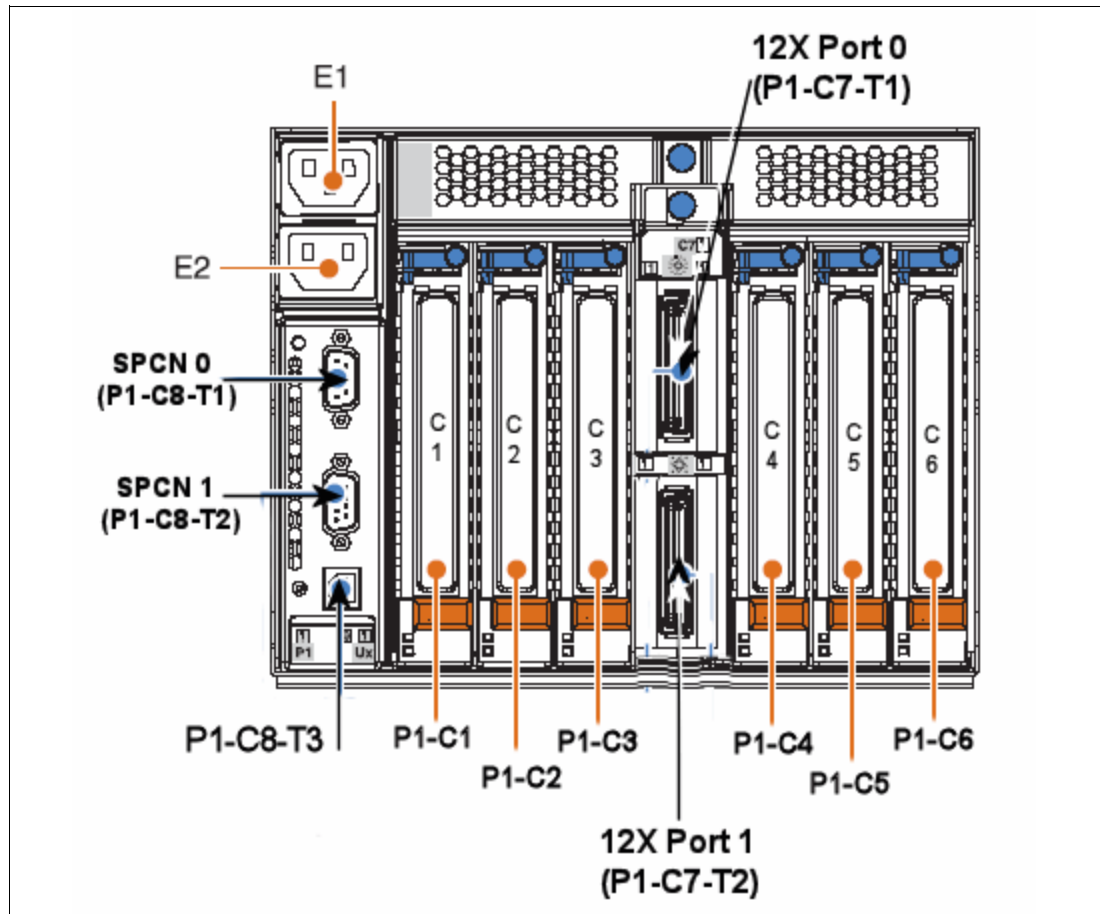


Figure 2-14 PCI-X DDR 12X Expansion Drawer rear side

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 (SFF) 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 cannot 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 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 2-15 shows the front view of the 12X I/O Drawer PCIe (#5802).

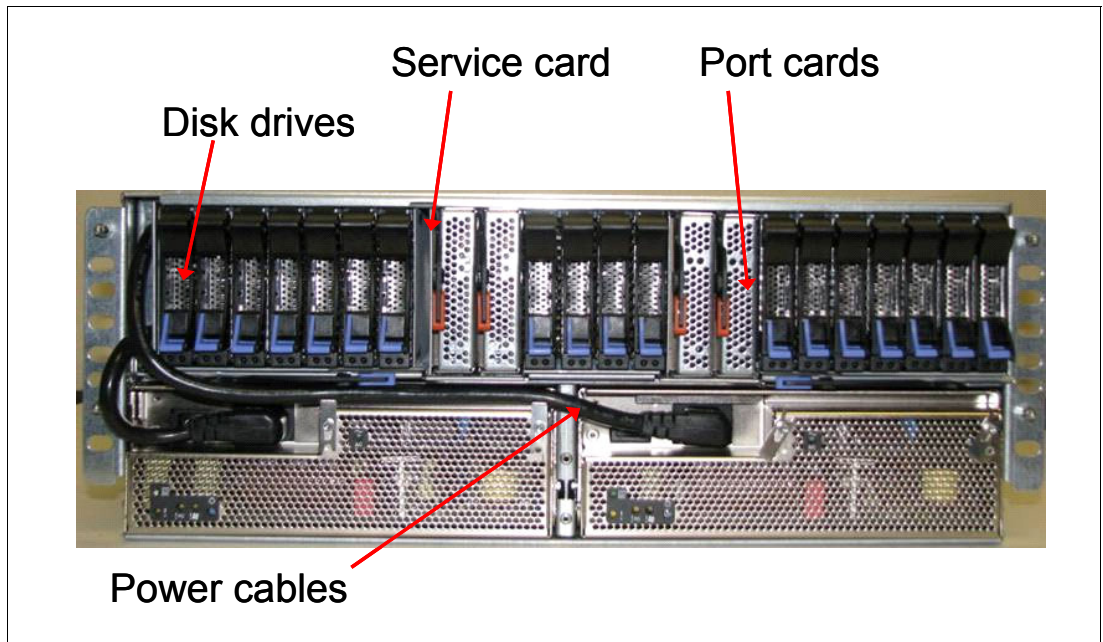


Figure 2-15 Front view of the 12X I/O Drawer PCIe

Figure 2-16 shows the rear view of the 12X I/O Drawer PCIe (#5802)

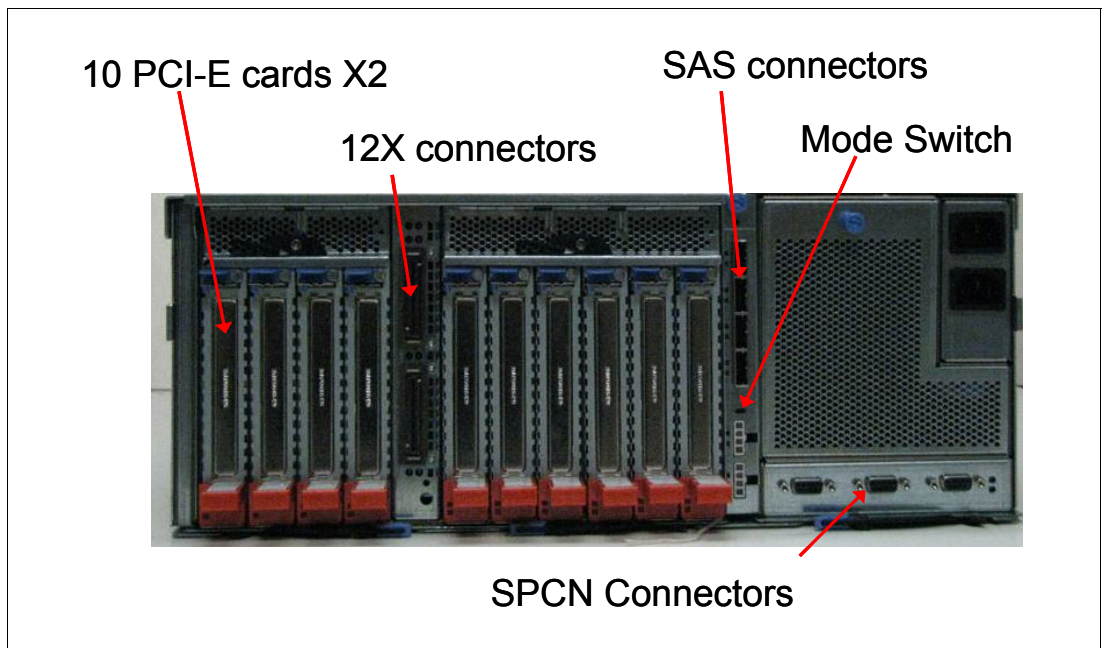


Figure 2-16 rear view of the 12X I/O Drawer PCIe

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 2-16 on page 56 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 2-17 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 Planing Tool supports disk bay partitioning. Also, the IBM configuration tool accepts this configuration from IBM System Planing Tool and passes it through IBM manufacturing using Customer Specified Placement (CSP) option.

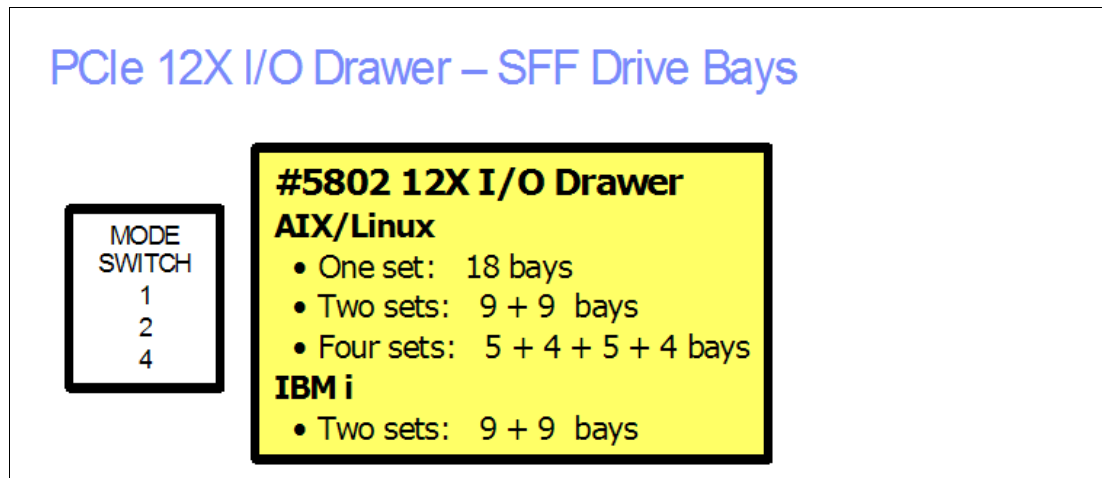


Figure 2-17 Disk Bay Partitioning in #5802 PCIe 12X I/O drawer

The SAS ports as associated with the mode selector switch map to the disk bays have the mappings shown in Table 2-18.

Table 2-18 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 2-18 on page 58 and Figure 2-19 on page 58.

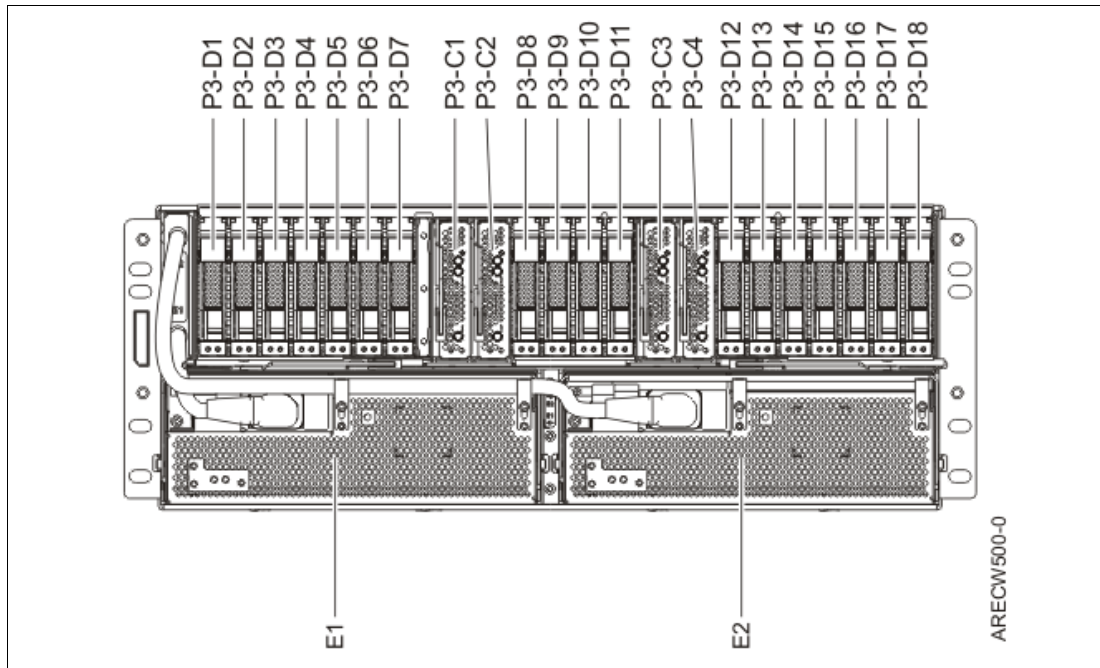


Figure 2-18 5802 I/O drawer from view location codes

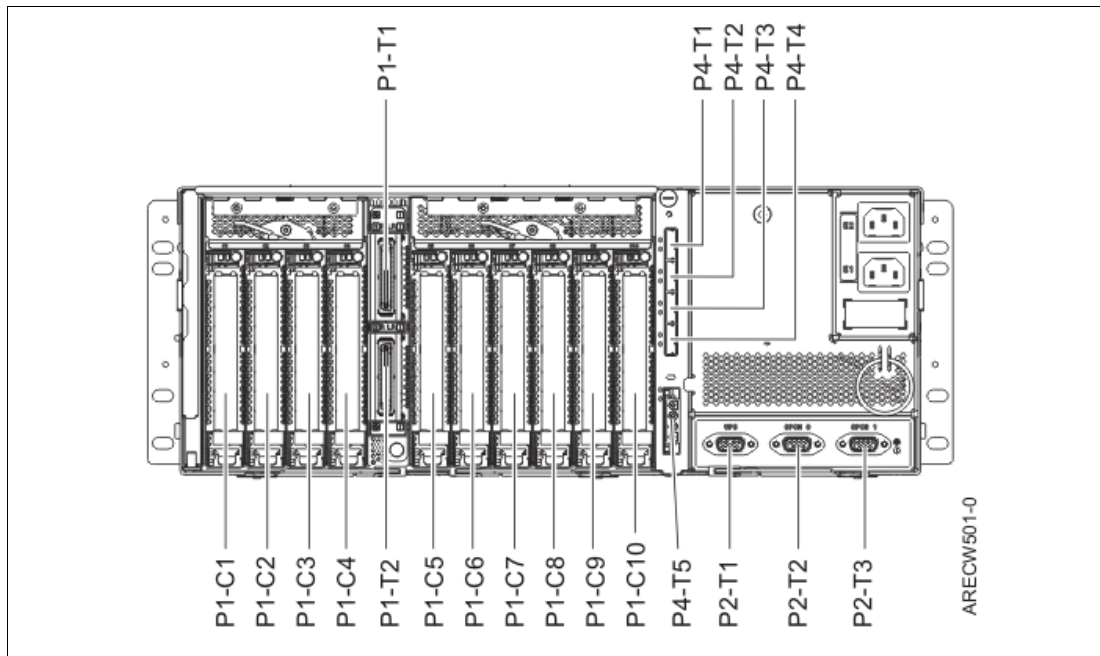


Figure 2-19 5802 I/O drawer rear view location codes

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 different 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. The #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 details about how to configure, see IBM Power Systems Hardware Information Center:
<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 or DDR cables (or 12X SDR and DDR cables) for the #5796 I/O drawers. The first 12X I/O Drawer 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 2-20 shows typical 12X I/O loop port connections.

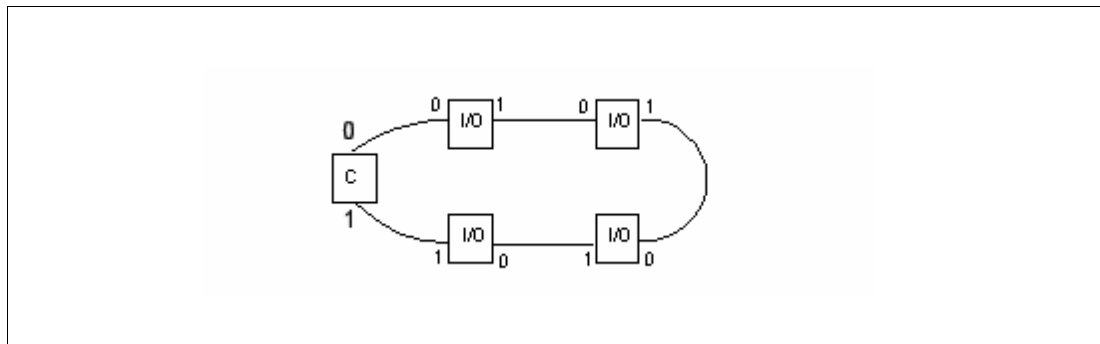


Figure 2-20 Typical 12X I/O loop port connections

Table 2-19 shows various 12X cables to satisfy various length requirements:

Table 2-19 12X connection cables

| Feature code | Description |
|-------------------|--------------------------|
| 1829 ^a | 0.6 Meter 12X Cable, SDR |
| 1830 ^a | 1.5 Meter 12X cable, SDR |
| 1840 ^a | 3.0 Meter 12X Cable, SDR |
| 1834 ^a | 8.0 Meter 12X Cable, SDR |
| 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 |

a. Supported, but no longer orderable

General rules for 12X I/O Drawer configurations

If you have only the 12X I/O Drawer PCIe (#5802 and #5877), populate the GX++ bus first. Figure 2-21 on page 60 shows PCIe I/O Drawer configurations.

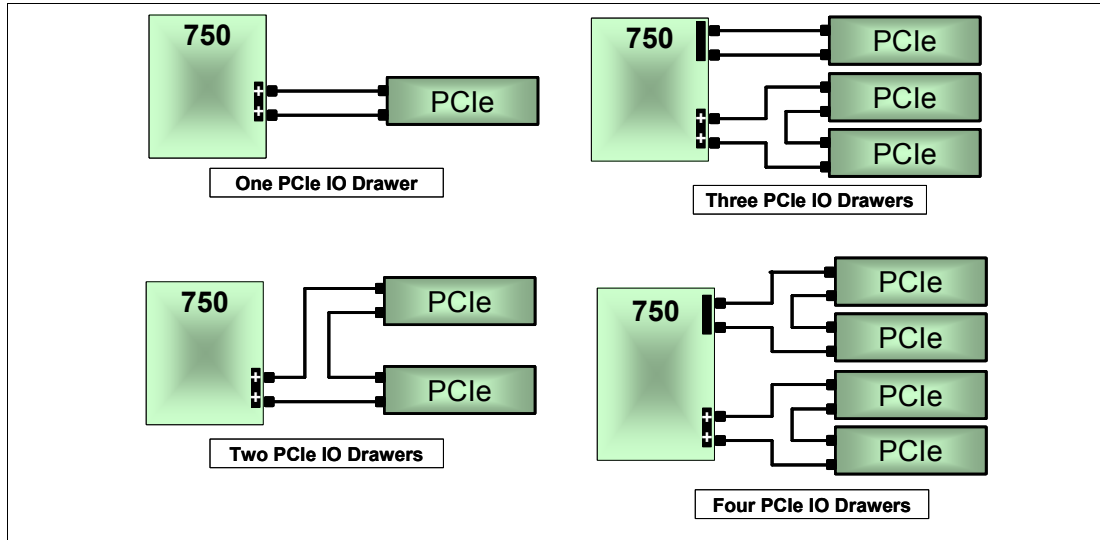


Figure 2-21 PCIe I/O Drawer configurations

If you have both PCI-DDR 12X Expansion drawer (#5796) and 12X I/O drawer PCIe (#5802 and #5877), populate the GX++ bus with PCIe drawers and GX+ with PCI-X drawers. Figure 2-22 shows mixed I/O drawer configurations.

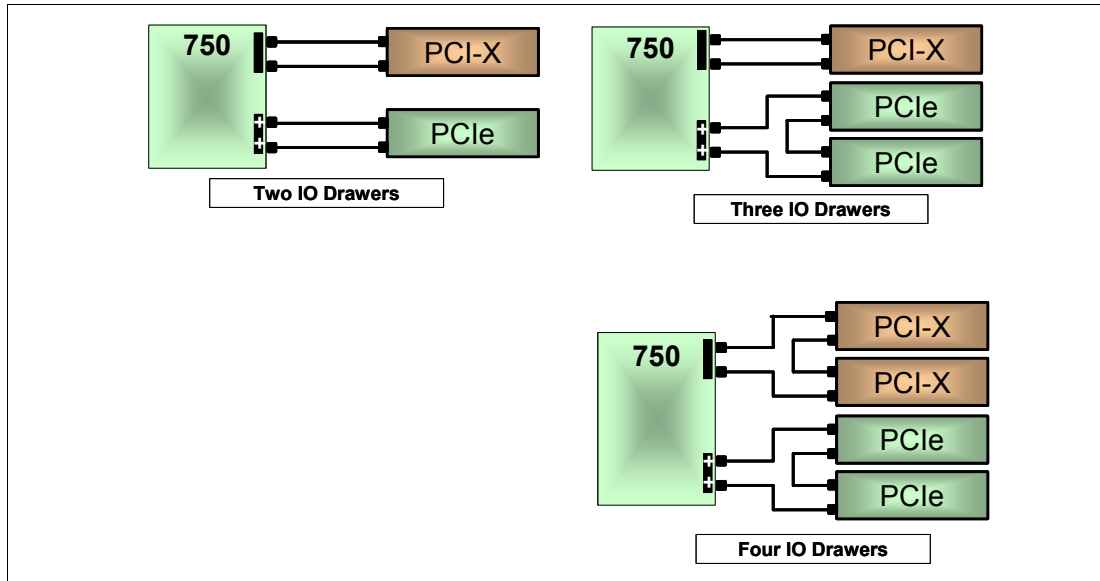


Figure 2-22 Mixed I/O Drawer configurations

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 2-20 on page 61 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 the table, Yes indicates that the particular 12X cable option can be used to connect the drawer configuration identified to the left; No indicates the option cannot be used.

Table 2-20 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 | No |

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 following example diagram, Figure 2-23. In the example:

- ▶ Start at SPCN 0 (T1) of the 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.
- ▶ To complete the cabling loop, from J16 (T2) of the final expansion unit, connect to the CEC, SPCN 1 (T2).
- ▶ Ensure a complete loop exists from the CEC, through all attached expansions and back to the CEC drawer.

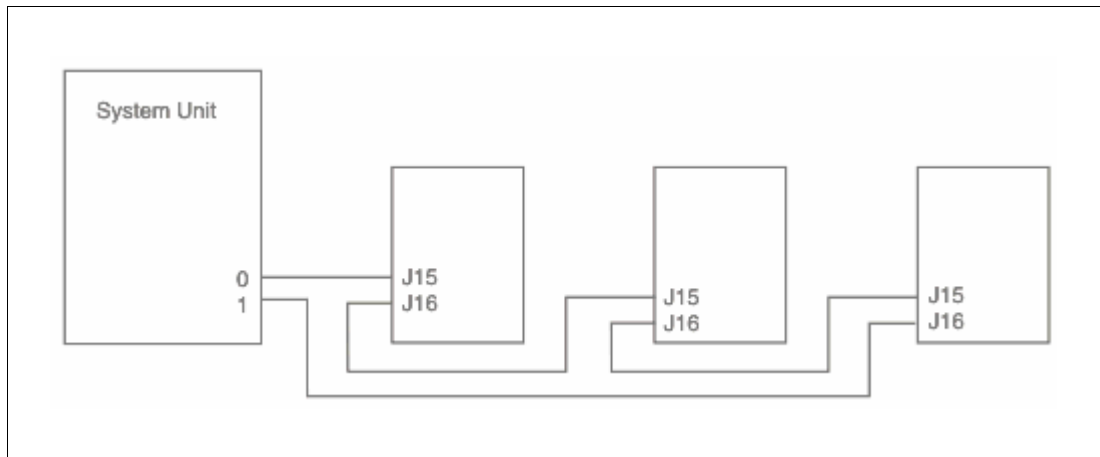


Figure 2-23 SPCN cabling examples

Table 2-21 shows various SPCN cables that satisfy various length requirements:

Table 2-21 SPCN cables

| Feature code | Description |
|-------------------|----------------------------------|
| 6001 ^a | SPCN cable drawer-to-drawer, 2 m |
| 6006 | SPCN cable drawer-to-drawer, 2 m |
| 6008 ^a | SPCN cable rack-to-rack, 6 m |

| Feature code | Description |
|-------------------|-------------------------------|
| 6007 | SPCN cable rack-to-rack, 15 m |
| 6029 ^a | SPCN cable rack-to-rack, 30 m |

a. Supported, but no longer orderable

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 EXP 12S 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 an SAS Y cable when a single port is running the EXP12S. An SAS X cable is used when a pair of adapters are used for controller redundancy.
- ▶ The medium cache PCI-X DDR Dual - x4 3Gb SAS RAID Adapter (#5902) and PCIe 380MB Cache Dual - x4 3 Gb SAS RAID Adapter (#5903) adapters are 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 and Linux environments when a pair of adapters are 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 using two SAS EE cables, providing 24 SAS bays instead of 12 bays for the same SAS controller port. This is called cascading. In this configuration, all 24 SAS bays are controlled by a single controller or a single pair of controllers.

The feature 5886 can also be directly attached to the SAS port on the rear of the Power 750, providing a very low cost disk storage solution. The rear SAS port is provided by the Enhanced DASD or Media Backplane for 2.5-inch DASD, SATA, DVD, or tape with External SAS Port (#8340). When used this way, the embedded SAS controller, augmented by the 175 MB write cache SAS RAID Enablement (#5679), in the system unit controls the disk drives in

EXP12S. A second unit cannot be cascaded to a feature 5886 attached in this way and one of the following features are required:

- ▶ SAS Cable, DASD Backplane to Rear Bulkhead (#3668)
- ▶ SAS Cable, DASD Backplane (Split) to Rear Bulkhead (#3669)

Figure shows an EXP 12S SAS drawer (#5886) connected to a Power 750 through the rear SAS port using a Y cable.

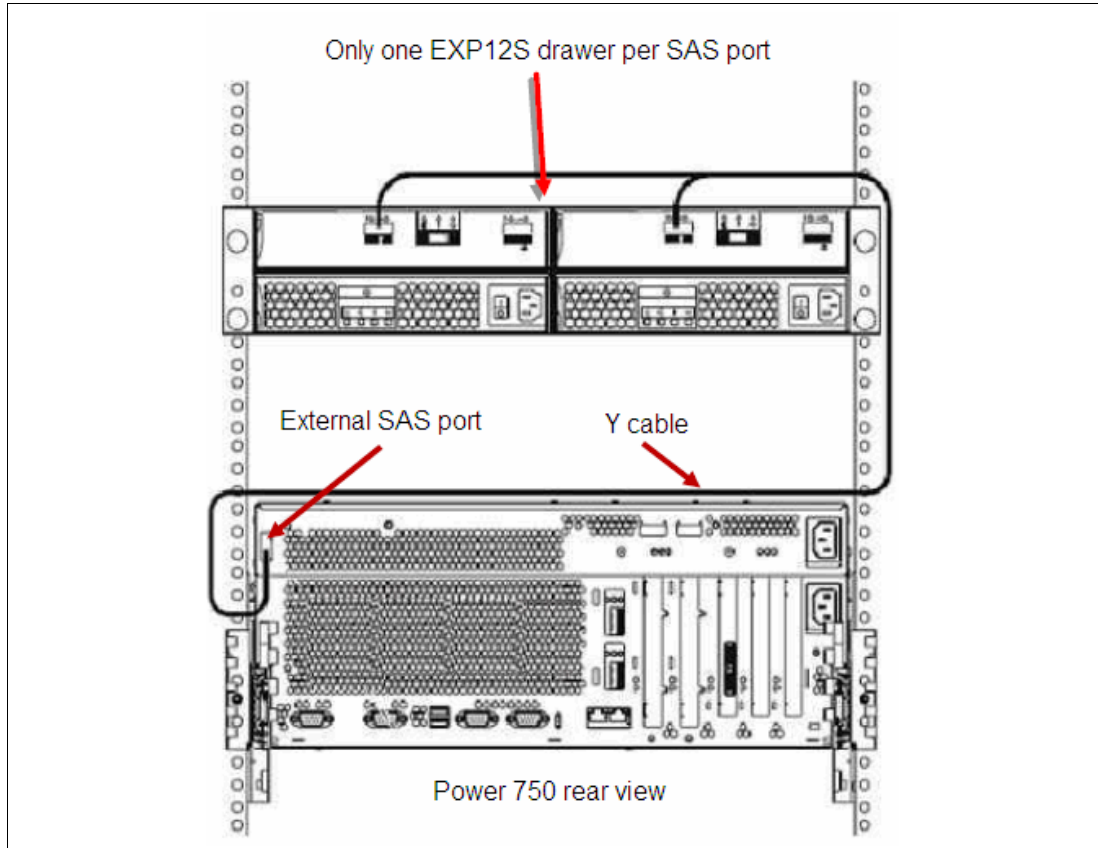


Figure 2-24 EXP 12S connection diagram to a Power 750

Note: EXP 12S SAS Drawer (#5886) is the only drawer supported on the Power 755

2.12.2 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

- ▶ The new N series Software Packs help you 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 take advantage of current Fibre Channel installations.

For more information, see the following Web site:

<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, see the following Web site:

<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, see the following Web site:

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

IBM System Storage DS5000

New 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, you can see up to a one-third reduction in floor space over standard enclosures. With the addition of 1 Gbps iSCSI host-attach, you can reduce cost for less demanding applications and continue to provide high performance where necessary by using the 8 Gbps FC host ports. With DS5000, you get consistent performance from a smarter design, that simplifies your infrastructure, improves your total cost of ownership (TCO), and reduces your cost.

For more information, see the following Web site:

<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-sized capacity requirements can take advantage of latest technology from IBM for

their most demanding applications with as little as 27 TB of usable capacity and incremental upgrades.

For more information, see the following Web site:

<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 dual IBM POWER6 based controllers that usher in a new level of performance for the company's flagship enterprise disk platform. The new DS8700 supports 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 be reliably available around the clock and with minimal delay, can you really afford not to run your business on the DS8000 series? With its tremendous scalability, flexible tiered storage options, broad server support, and support for advanced IBM duplication technology, the DS8000 can help simplify the storage environment by consolidating multiple storage systems onto a single system, and provide the availability and performance you have come to trust for your most important business applications.

For more information, see the following Web site:

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

2.13 Hardware Management Console (HMC)

The 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 (as well as 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, 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, 755, 770, or 780 servers. For updates of the machine code and HMC functions and hardware prerequisites, see the following Web site:

<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 / 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, and so on)
- ▶ Custom Groups

HMC management

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

- ▶ Set-up wizard
- ▶ User Management
 - User IDs
 - Authorization levels
 - Customizable authorization
- ▶ Disconnect and reconnect
- ▶ Network Security
 - Remote operation enable and disable
 - User definable SSL certificates
- ▶ Console logging
- ▶ HMC Redundancy
- ▶ Scheduled Operations
- ▶ Back-up and Restore
- ▶ Updates, Upgrades
- ▶ Customizable Message of the day

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

- ▶ Server Management
 - Support for Power 750 and 755 Server
 - Support for Power 770 and 780 Server (V7R710 SP1)

- ▶ Virtualization Management
 - Remove limit of 128 PowerVM 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 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 command line interface is also available by using the SSH secure shell connection to the HMC. It can be used by an external management system or a partition to perform HMC operations remotely.

2.13.2 HMC connectivity to the POWER7 processor based systems

POWER5, POWER5+, POWER6, POWER6+, and POWER7 processor technology based servers 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 4 Ethernet ports. The deskside 7042-C07 HMC standard configuration offers only one Ethernet port; be sure to order an optional PCIe 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 2-25 on page 68 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.

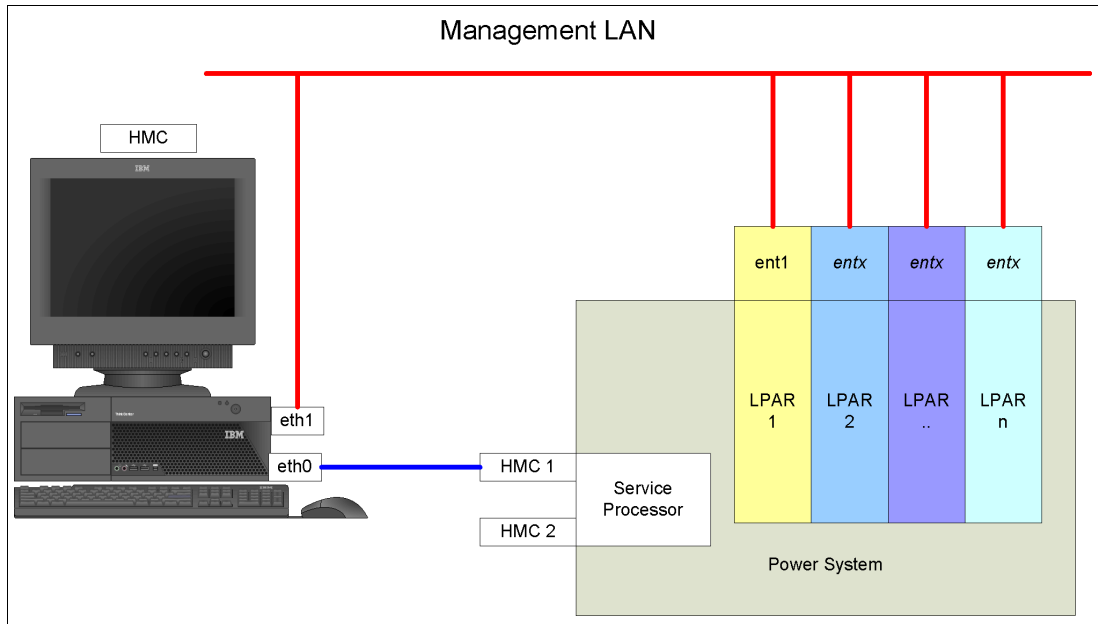


Figure 2-25 HMC to service processor and LPARs network connection

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 service processor is allocated an 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 DHCP reply before time-out, predefined IP addresses will be setup 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 only visible to the service processor and can be used to attach the server to an HMC or to access the Advanced System Management Interface (ASMI) options from a client Web browser, using the HTTP server integrated into the service processor internal operating system.
- ▶ When not configured otherwise (DHCP or from a previous ASMI setting) both Ethernet ports of the first service processor 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

More information about the Service Processor is in “Service processor” on page 122.

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, some 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 can 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 (#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 2-26 shows one possible high available HMC configuration managing two servers. These servers have only one CEC and therefore only one service processor. Each HMC is connected to one service processor port of all managed servers.

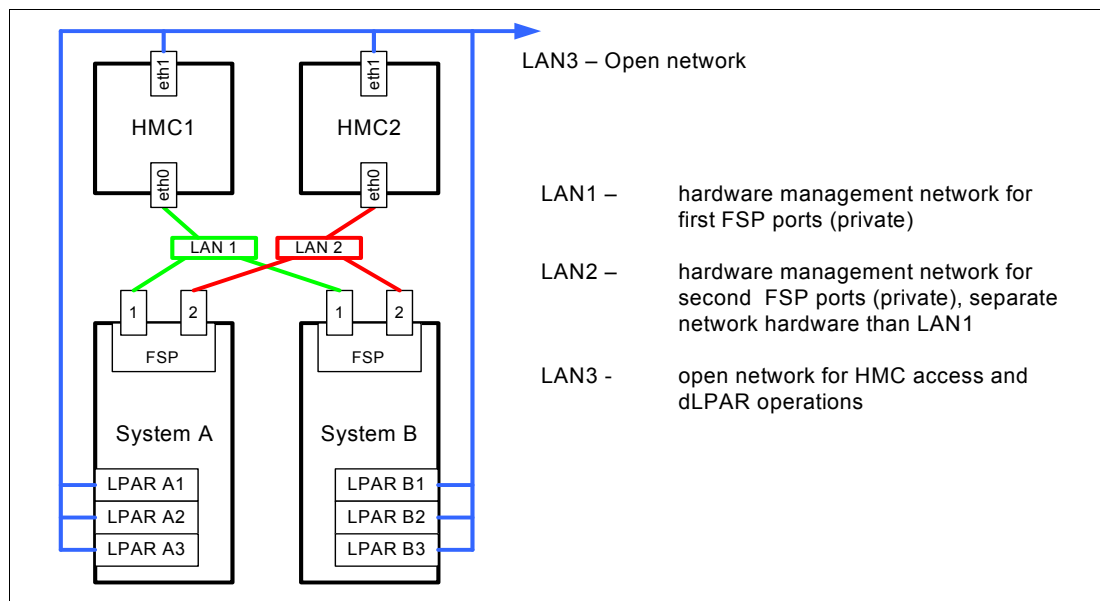


Figure 2-26 Highly available HMC and network architecture

Note, in Figure 2-26, that only hardware management networks (LAN1 and LAN2) are highly available in order to keep 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 on a DHCP server for its VLAN.

In a configuration with multiple systems or HMC's, the customer is required to provide switches or hubs to connect each HMC to the server service processor 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 provides redundancy both for the HMCs and the service processors.

For more 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 Power 750 and 755 systems
- ▶ V7R710 SP1 to support the Power 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 an HMC in a dual HMC configuration, a good practice is to disconnect one HMC and avoid both HMCs being connected to the same server with various code levels. If no profile or partition changes take place during the upgrade, both HMCs can stay connected. If the HMCs are at different code levels and a profile change is performed from the HMC at the higher code level, the format of the data that is stored in the server could change, and the HMC at the lower code level could go into a recovery state if it does not understand the new data format.

There are compatibility rules between the software that is executing within a POWER7 processor-based server environment: HMC, Virtual I/O Server, system firmware or partition operating systems. To check what combinations are supported, and to identify required upgrades, use the Fix Level Recommendation Tool:

<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, which manages the destination server, is at version 7, release 7.1 or later.

2.14 IVM

The HMC has been designed to be the comprehensive solution for hardware management that can be used either for a small configuration or for a multiserver environment. Although complexity has been kept low by design and many recent software revisions support this, the HMC solution might not fit in small and simple environments where only a few servers are deployed or not all HMC functions are required.

Integrated Virtualization Manager (IVM) is a simplified hardware management solution that inherits most of the HMC features. It manages a single server, avoiding the need of an independent appliance. It is designed to provide a solution that enables the administrator to reduce system setup time and to make hardware management easier, at a lower cost.

IVM provides a management model for a single system. Although it does not offer all of the HMC capabilities, it enables the exploitation of PowerVM technology. IVM targets the small

and medium systems environment. At the time of writing, in the family of POWER7 processor based servers, IVM can only manage the Power 750 model.

IVM is an addition to the Virtual I/O Server (VIOS), the product that enables I/O virtualization in the family of POWER processor-based systems. The IVM functions are provided by software executing within the Virtual I/O Server partition installed on the server to manage. See Table 2-22. For a complete description of the possibilities offered by IVM, see *Integrated Virtualization Manager on IBM System p5*, REDP-4061.

Table 2-22 Comparison of IVM and HMC

| Characteristics, functions | IVM | HMC |
|-----------------------------------|--|---|
| General characteristics | | |
| Delivery vehicle | Integrated into the server | A desktop or rack-mounted appliance |
| Footprint | Runs in 60 MB memory and requires minimal CPU as it runs stateless. | 2-Core x86, 2 GB RAM, 80 GB HD |
| Installation | Installed with the Virtual I/O Server (optical or network). Preinstall option available on some systems. | Appliance is preinstalled. Reinstall through optical media or network is supported. |
| Multiple system support | One IVM per server | One HMC can manage multiple servers (48 CECs / 1024 LPARS) |
| User interface | Web browser (no local graphical display) and telnet session | Web browser (local or remote) |
| Scripting and automation | VIOS command-line interface (CLI) and HMC compatible CLI. | HMC CLI |
| RAS characteristics | | |
| Redundancy and HA of manager | Only one IVM per server | Multiple HMCs can manage the same system for HMC redundancy. |
| Multiple VIOS | No, single VIOS | Yes |
| Fix or update process for manager | VIOS fixes and updates | HMC e-fixes and release updates |
| Adapter microcode updates | Inventory scout through RMC | Inventory scout through RMC |
| Firmware updates | Inband through OS; not concurrent | Service Focal Point™ with concurrent firmware updates |
| I/O concurrent maintenance | VIOS support for slot and device level concurrent maintenance through the diag hot plug support | Guided support in the “Repair and Verify” function on the HMC. |
| Serviceable event management | Service Focal Point Light: Consolidated management of firmware- and management partition-detected errors | Service Focal Point support for consolidated management of operating system- and firmware-detected errors |

| Characteristics, functions | IVM | HMC |
|---|-----------------------------|--------------------|
| PowerVM functions | | |
| Full PowerVM Capability | Partial | Full |
| Capacity on Demand | Entry of PowerVM codes only | Full support |
| I/O Support for i5/OS® | Virtual only | Virtual and direct |
| Multiple Shared Processor Pool | No, default pool only | Yes |
| Workload Management (WLM) Groups Supported | One | 254 |
| Support for multiple profiles per partition | No | Yes |
| SysPlan Deploy & mksysplan | Yes | Yes |

2.15 Operating system support

The IBM POWER7 processor based systems supports three families of operating systems: AIX, IBM i, and 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, see Power Systems Software™ Web site:

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

Virtual I/O Server

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

Power 750 VIOS 2.1.2.11 with Fix Pack 22.1 and Service Pack 1

Power 755 VIOS feature is not available on this model.

Power 770 and 780 VIOS 2.1.2.12 with Fix Pack 22.1 and Service Pack 2

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

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

IBM AIX Version 5.3

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

The minimum level of AIX Version 5.3 to support the POWER7 processor-based server is:

- ▶ AIX Version 5.3 with the 5300-11 Technology Level and Service Pack 2 or later

The Power 750, 755, 770, and 780 also support other existing technology levels, which are available at the end of May 2010:

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

A partition using AIX Version 5.3 executes in POWER6 or POWER6+ compatibility mode.

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

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

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 functionality, see the following Web site:

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

AIX Version 6.1

If you are installing AIX 6.1 on a POWER7 processor technology based server, the minimum level requirements depends on the target server model:

Power 750 and 755 AIX Version 6.1 with the 6100-04 Technology Level, Service Pack 2

Power 770 and 780 AIX Version 6.1 with the 6100-04 Technology Level, Service Pack 3

IBM supports other Technology Levels on these models:

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

A partition that uses AIX 6.1 with TL2, TL3, or TL4 up to SP2 executes in POWER6 or POWER6+ compatibility mode. Starting from TL4 SP3, AIX 6.1 fully supports POWER7 mode.

A partition using IBM AIX Version 5.3 executes in POWER6 or POWER6+ compatibility mode. IBM is making available a new version of AIX, AIX V6.1 which will include significant new capabilities for virtualization, security features, continuous availability features and manageability. AIX V6.1 is the first generally available version of AIX V6.

For information regarding AIX V6.1 maintenance and support, go to the Fix Central Web site:

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

IBM i

The IBM i operating system is supported on IBM Power 750, 770 and 780, at this version:

- ▶ IBM i 6.1 with i 6.1.1 machine code, or later

IBM i Standard Edition, Enterprise Edition, and Application Server Express Edition options are available for these three server models:

- ▶ IBM i Express Edition offers IBM i without DB2® for application and infrastructure serving.
- ▶ IBM i Standard Edition offers an integrated operating environment for business processing.
- ▶ IBM i Enterprise Edition offers IBM i plus Enterprise Enablement which provides 5250 transaction processing support.

IBM i is not supported on Power 755 model.

Table 2-23 summarizes the POWER family servers that are supported by IBM i.

Table 2-23 IBM i support for POWER servers.

| Servers | IBM i 5.4 | IBM i 6.1 |
|---|-----------|-----------|
| POWER7 750, 770, 780 | No | Yes |
| POWER6 JS12, JS22, JS23, JS43 550 560 | No | Yes |
| POWER6 520, 550, 570, 595 | Yes | Yes |
| POWER5+ 515, 525 | Yes | Yes |
| POWER5 520, 550, 570, 595 800, 810, 825, 870, 890 | Yes | Yes |
| POWER4™ 270, 820, 830, 840 | Yes | No |

Version 6.1.1 of IBM i on POWER7 processor-based servers introduce several enhancements:

- ▶ Performance enhancements
 - Support for 32 nodes
 - Support for 2 levels of affinity
- ▶ Support for 1000 partitions with PowerVM
 - IBM i i 6.1.1 can run on a system with 1000 partitions but it must have partition IDs in the range of 1 - 254.
- ▶ POWER7 SMT4 IBM i enhancements
 - IBM i 6.1.1 supports SMT2 (POWER6 mode only) and SMT4.
 - IBM i 6.1.1 is limited to 32 way-SMT4 configuration. A 64 way-SMT2 configuration is supported in POWER6 mode).

Linux

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

This section discusses two brands of Linux to be run in partitions. At the time of this writing, the supported versions of Linux on POWER7 processor technology 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

In addition, other distribution of Linux are expected to support POWER7 processor technology based server.

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 targeted for availability during the first half of 2010. For additional questions on the availability of this release, contact Red Hat.

To configure Linux partitions in virtualized IBM Power Systems, consider the following information:

- ▶ Not all devices and features supported by the AIX operating system are supported in logical partitions running the Linux operating system.
- ▶ Linux operating system licenses are ordered separately from the hardware. You may obtain Linux operating system licenses from IBM, to be included with your POWER7 processor technology-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” on page 105.

2.16 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.

2.17 Energy management

The Power 750 and 755 are ENERGY STAR-qualified, designed with features to help clients become more energy efficient. ENERGY STAR-qualified products use less energy and reduce greenhouse gas emissions by meeting strict energy efficiency guidelines. 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.

2.17.1 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 Director plug-in)

IBM EnergyScale functions include:

- ▶ Energy trending

EnergyScale provides continuous collection of real-time server energy consumption. This function enables administrators to predict power consumption across 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 reduces 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 re-boot operations although it is a persistent condition that is sustained after the boot when the system starts executing instructions.

- ▶ Dynamic Power Saver Mode

Dynamic Power Saver Mode varies processor frequency and voltage based on the utilization of the POWER7 processors. The user must 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 utilization. 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 varies, 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 can reduce the maximum frequency to 95% of nominal values. If performance is favored over energy consumption, the maximum frequency will 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 datacenter. 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 an 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) such 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 mode. 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 power savings settings are enforced (either Power Energy Saver Mode or Dynamic Power Saver Mode), fan speed varies 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.

Note: When the DC power supply is installed, Power 750 and 755 will not support the IBM Director Active Energy Management functions

2.17.2 Thermal power management device card (TPMD)

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, and 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.



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

Note: This chapter is relevant to users of the IBM Power 750 server. The IBM Power 755 server is intended for HPC application and supports a single partition configuration. PowerVM and virtualization features are not available on the IBM Power 755.

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, and 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 93).
- ▶ Saves and restores all processor state information during a logical processor context switch.
- ▶ Controls hardware I/O interrupt management facilities for logical partitions.
- ▶ Provides virtual LAN channels between logical partitions that help to reduce the need for physical Ethernet adapters for inter-partition communication.
- ▶ Monitors the Service Processor and performs a reset or reload if it detects the loss of the Service Processor, notifying the operating system if the problem is not corrected.

The POWER Hypervisor is always active, regardless of the system configuration and also when not connected to the 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 98. 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 AIX Version 5.3, or appropriate level of Linux supporting virtual Ethernet devices (see 3.4.9, “Operating system support for PowerVM” on page 105). 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 98 for more details about shared Ethernet). This is 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 84 depicts the connections between the client partition virtual Fibre Channel adapters and the external storage. For additional information, refer to 3.4.8, “NPIV” on page 105.

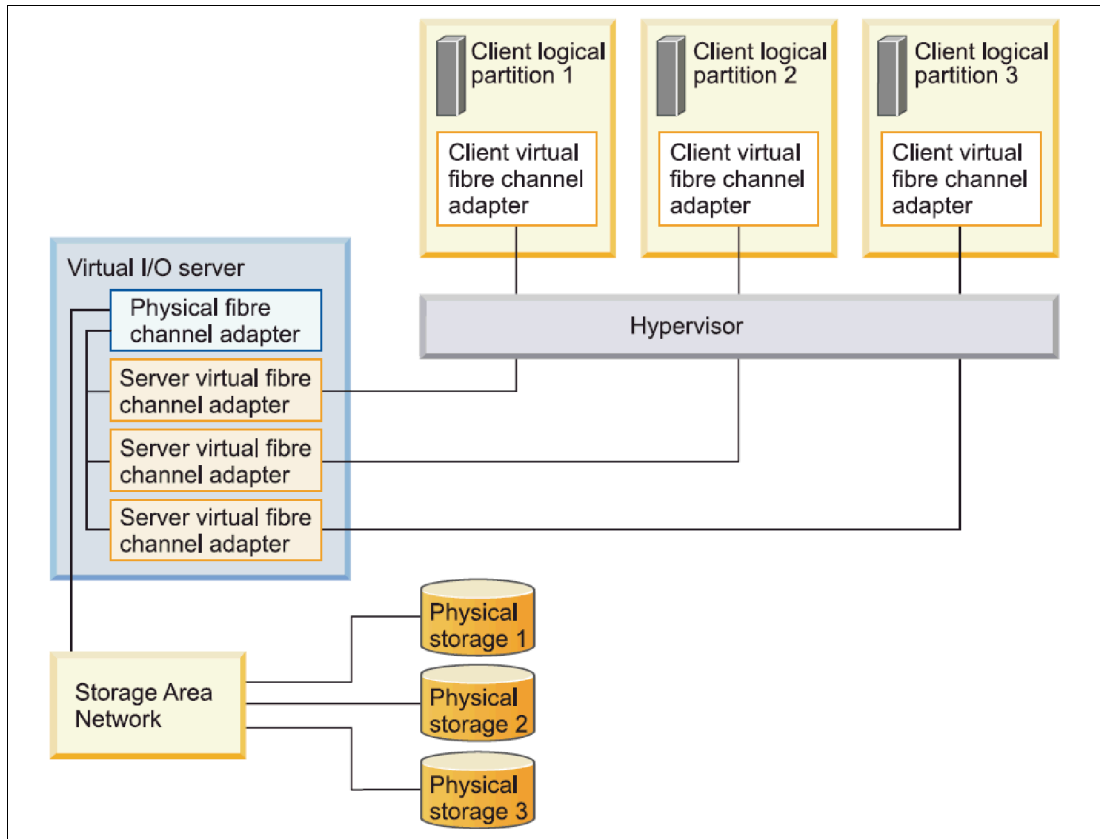


Figure 3-1 Connectivity between virtual Fibre Channels adapters and external SAN devices.

Virtual (TTY) console

Each partition needs to have access to a system console. Tasks such as operating system installation, network setup, and some problem analysis activities require a dedicated system console. The POWER Hypervisor provides the virtual console using a virtual TTY or serial adapter and a set of Hypervisor calls to operate on them. Virtual TTY does not require the purchase of any additional features or software such as the 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, POWER modes are described in this section because they affect certain virtualization features.

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

- ▶ POWER6 compatibility mode

This execution mode is compatible with v2.05 of the Power Instruction Set Architecture (ISA). For more information, see the following Web file:

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, see the following file:

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.

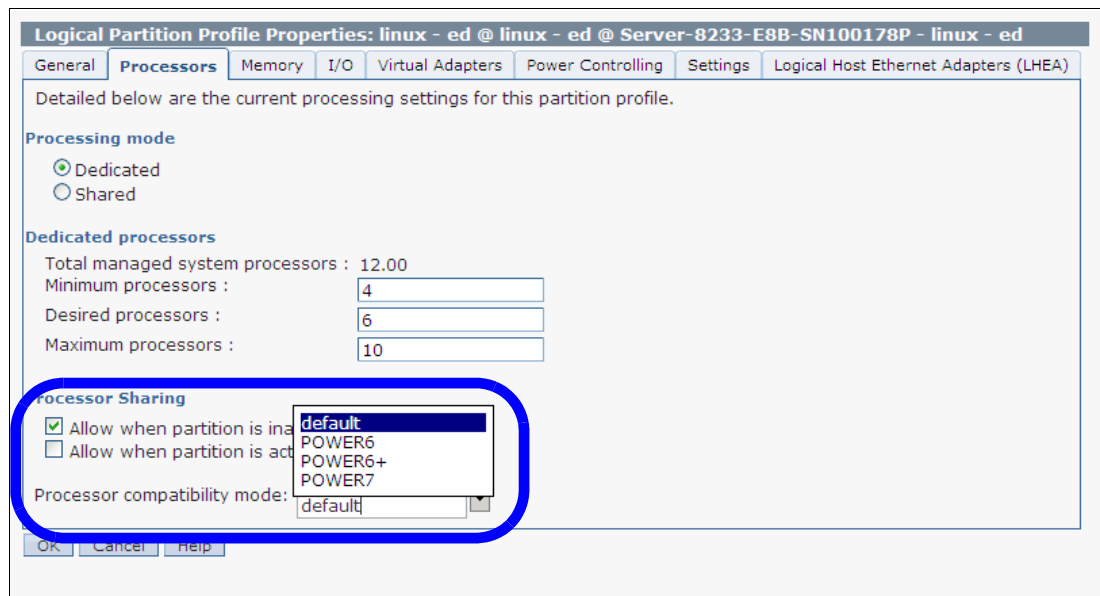


Figure 3-2 Configuring partition profile compatibility mode from the HMC

Table 3-2 on page 86 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 or 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 |
| 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, which 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%. This can allow a partition to do 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 used. Other test workloads had more modest results.

Clients have a great deal of 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 IPL is required for the specific partition that is turning memory expansion on or off. When 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 used to compress memory for two partitions with different profile. The green curve corresponds to a partition that has spare processing power capacity; the blue curve corresponds to a partition constrained in processing power.

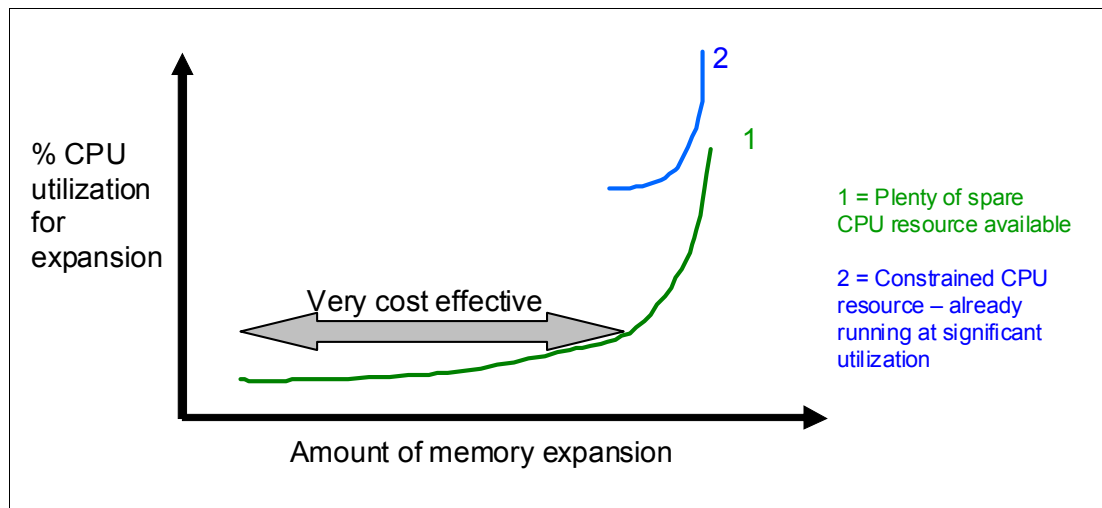


Figure 3-3 CPU usage versus memory expansion effectiveness

Both cases shows 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 required

The knee varies depending on how compressible memory contents are. This demonstrates the need for a case by case study of whether memory expansion can provide a positive return on investment.

To help you performing this study, a planning tool is included with AIX 6.1 Technology Level 4 allowing you to sample actual workloads and estimate both 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 88 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 to allocate 58% of a processor, to benefit from 45% extra memory capacity.

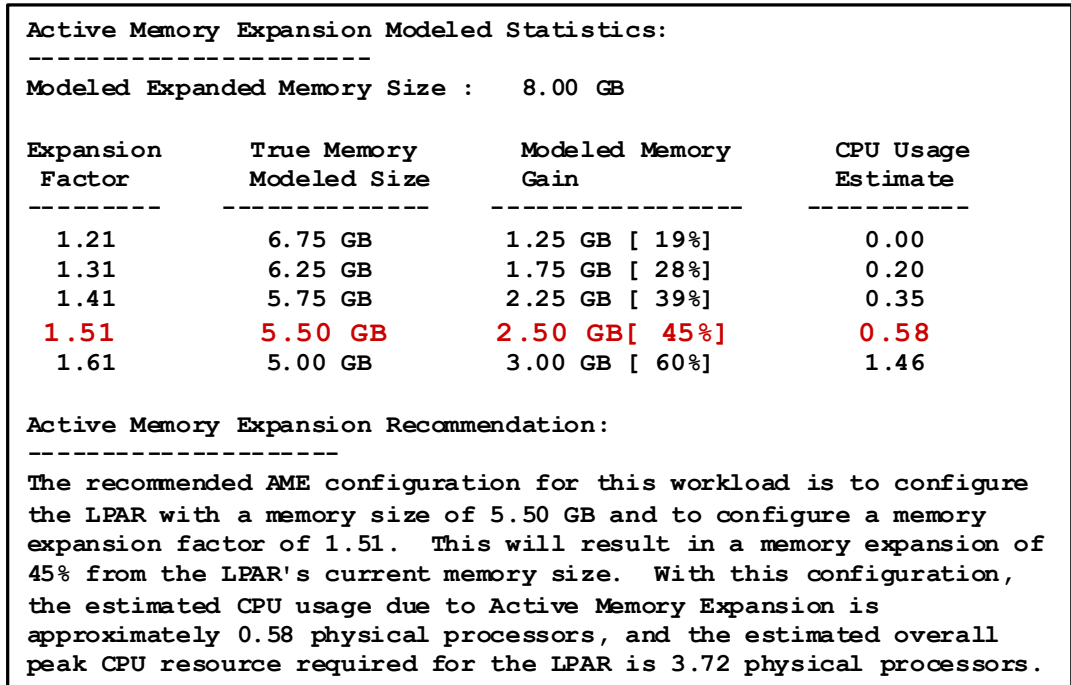


Figure 3-4 Output from Active Memory Expansion planning tool

When you have selected 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

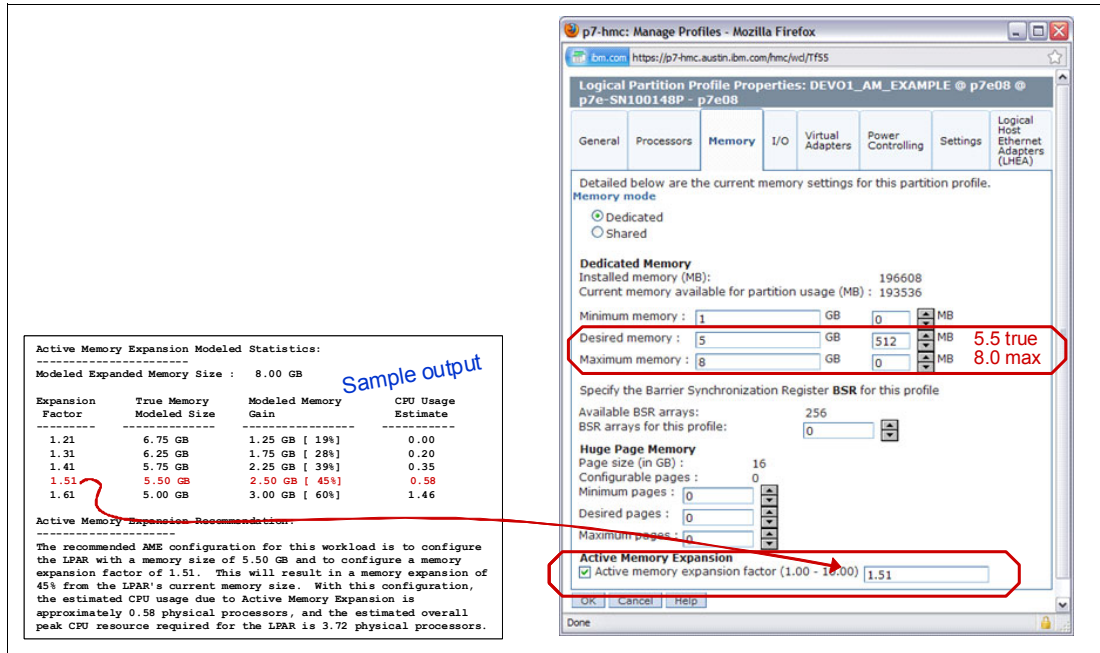


Figure 3-5 Using the planning tool result to configure the partition

On the HMC menu describing the partition, check the **Active Memory Expansion** box and enter the true and maximum memory, as well as 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 a miscellaneous equipment specification (MES) order. A software key is provided when the enablement feature is ordered that is applied to the server. A reboot 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 different server. This feature is ordered per server, independently of the number of partitions using memory expansion.

You may view whether the Active Memory Expansion feature has been activated on a server from the HMC, as shown in Figure 3-6

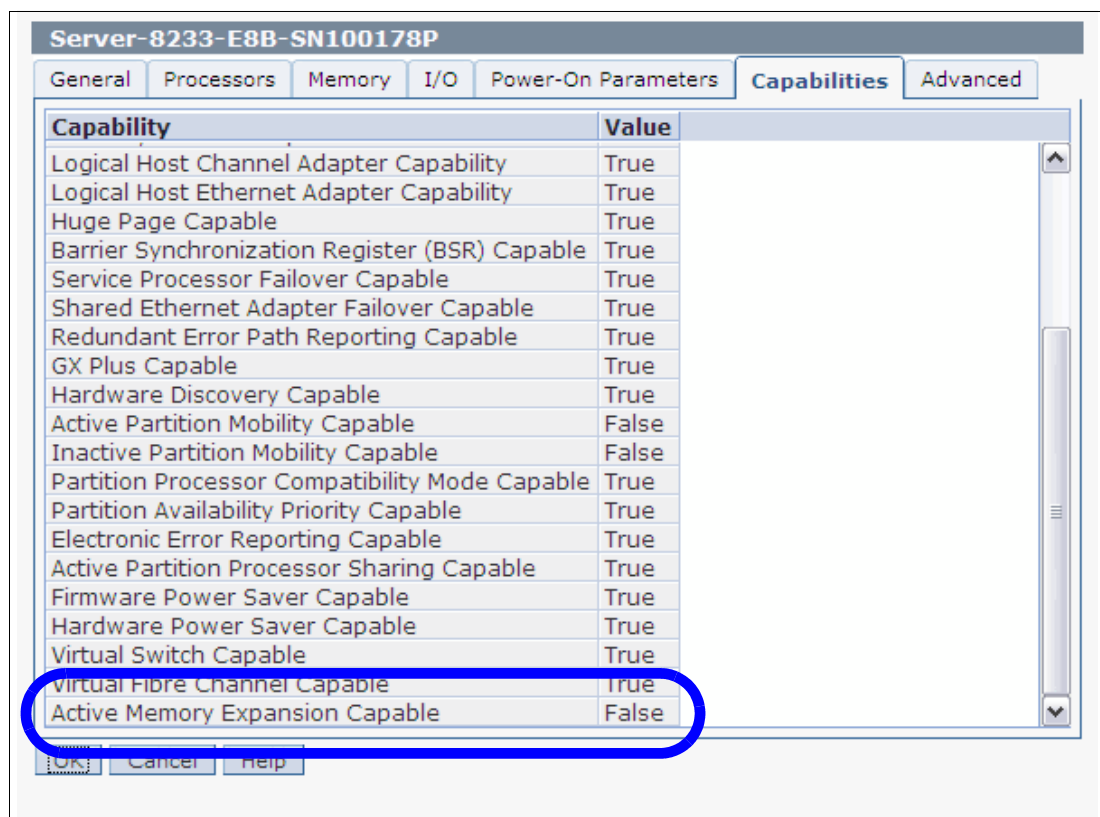


Figure 3-6 Server capabilities listed from the HMC.

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, download the document *Active Memory Expansion: Overview and Usage Guide*:

http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=SA&subtype=WH&appname=STGE_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 PowerVM (Logical Partitioning, Micro-Partitioning™, Power Hypervisor, Virtual I/O Server, Live Partition Mobility, Workload Partitions, and so on). As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and value-added software. The licensed features of each of the three editions of PowerVM are discussed next, in 3.4.1, “PowerVM editions” on page 90.

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 which are available on each model of POWER7 processor technology based servers:

Table 3-3 Availability of PowerVM per POWER7 processor-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 is possible. Table 3-4 on page 91, lists 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 or 2 per server | 10 per core | 10 per core |
| Management | VMcontrol IVM | VMcontrol IVM, HMC | VMcontrol IVM, HMC |
| Virtual I/O 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

Logical partitions (LPARs) and virtualization increase utilization of system resources and add a new level of configuration possibilities. This section provides details and configuration specifications about this topic.

Dynamic logical partitioning

Logical partitioning (LPAR) was introduced with the POWER4 processor-based product line and the IBM 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. IBM 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 two-core server has two physical processors.

For 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 that the partition is dispatched on physical processors. Processing units define the capacity entitlement of the partition.
- ▶ The shared processor pool: Select one from the list with the names of each configured shared processor pool. The list also displays the pool ID, in parentheses, of each configured shared processor pool. If the name of the desired shared processor pool is not available here, you must first configure the desired shared processor pool by using the Shared Processor Pool Management window. Shared processor partitions use the default shared processor pool called DefaultPool by default.
- ▶ Cap or uncap partition: Select whether or not the partition will be able to access extra processing power to “fill up” its virtual processors above its capacity entitlement, selecting either to cap or uncap your partition. If spare processing power is available in the shared processor pool or other partitions are not using their entitlement, an uncapped partition can use additional processing units if its entitlement is not enough to satisfy its application processing demand.
- ▶ Weight: The weight (preference) in the case of an uncapped partition.
- ▶ Virtual processors: The minimum, desired, and maximum number of virtual processors.

The POWER Hypervisor calculates a partition’s processing power based on minimum, desired, and maximum values, processing mode, and on the 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 0.1 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:

- ▶ Respectively 32 and 64 dedicated partitions
- ▶ Up to 160 micro-partitions

It is important to point out 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:

- ▶ 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. 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 processor units from a shared processor pool. This setting will define the processing mode of the logical partition. Figure 3-7 shows a diagram of the concepts discussed in the remaining sections.

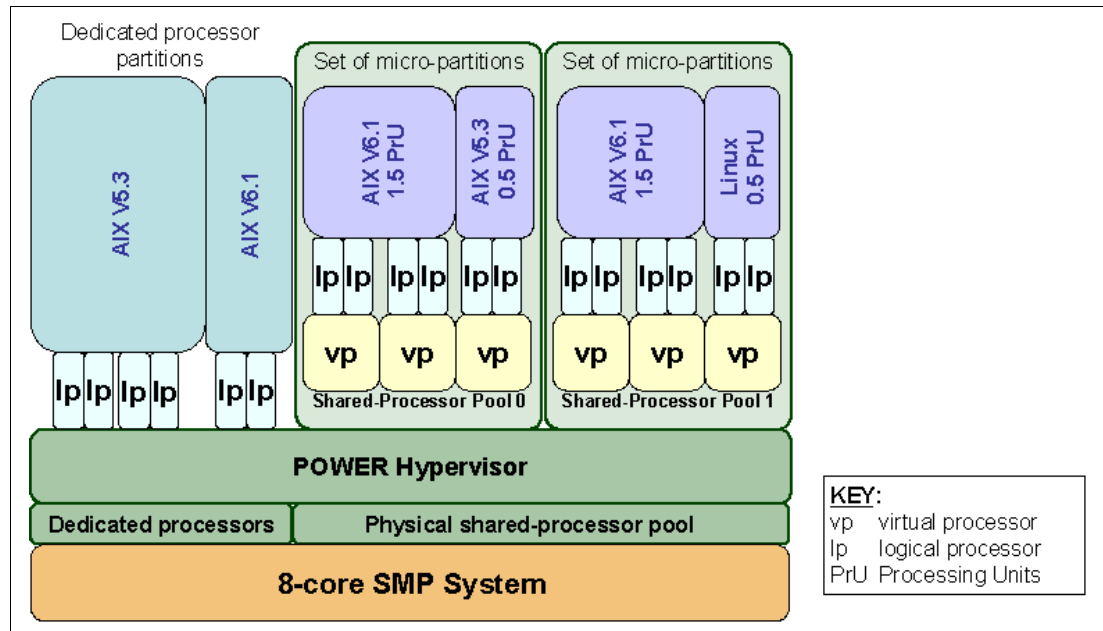


Figure 3-7 Logical partitioning concepts

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 can help to increase system utilization, without compromising the computing power for critical workloads in a dedicated processor.

Shared mode

In shared mode, logical partitions use virtual processors to access fractions of physical processors. Shared partitions can define any number of virtual processors (maximum number is 10 times the number of processing units assigned to the partition). From the POWER Hypervisor perspective, 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 receive total CPU time equal to their processing units entitlement. The logical processors are defined on top of virtual processors. Therefore, 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 is shown in Figure 3-8.

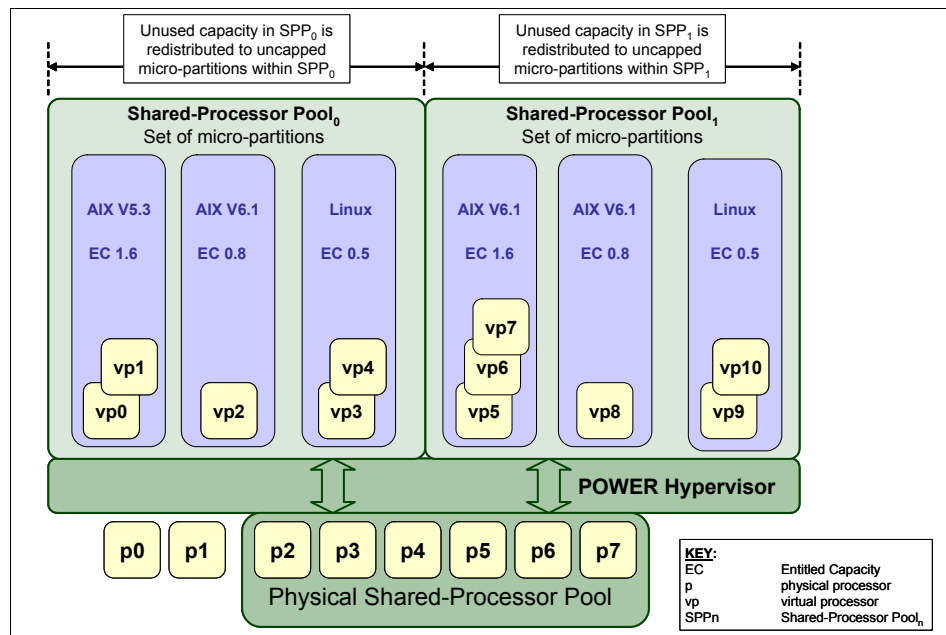


Figure 3-8 Overview of the architecture of Multiple Shared Processor Pools

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 may 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).

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

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. This is the second level of processor capacity resolution.

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

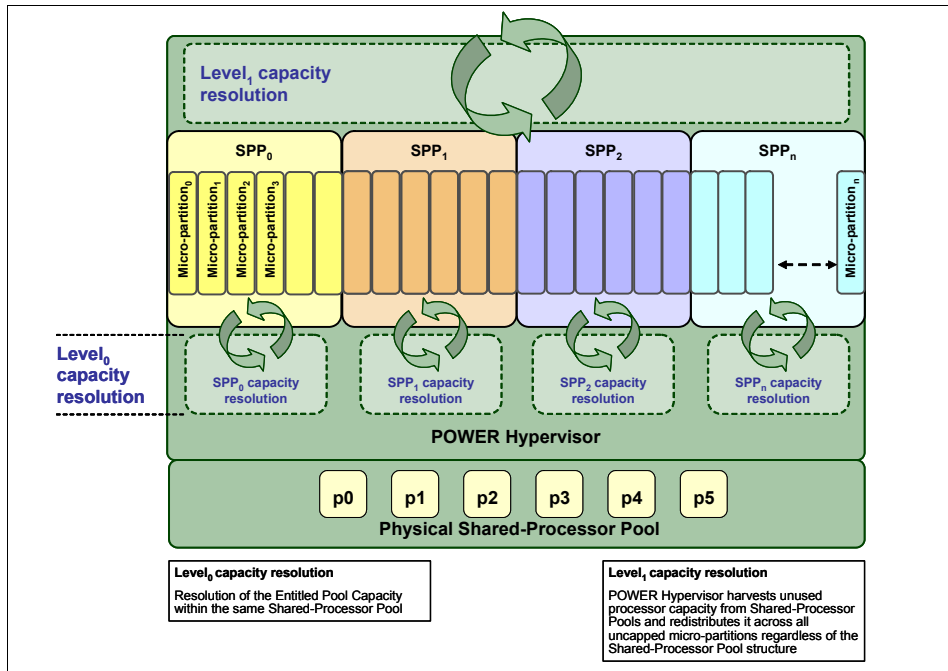


Figure 3-9 The two levels of unused capacity redistribution

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 some 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 items 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 approach 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 CLI interface.

Dynamic adjustment of Reserved 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 by using the HMC with 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 increases.

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.

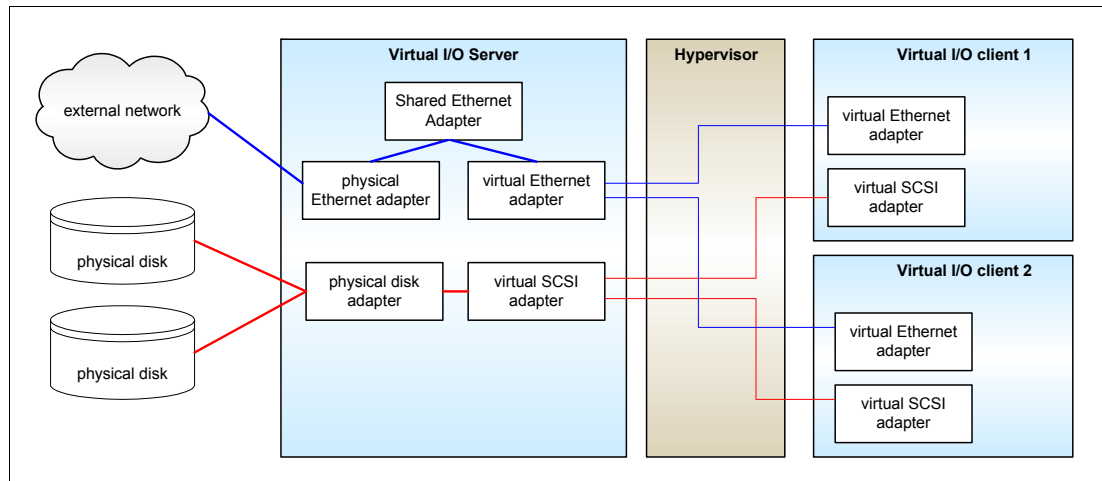


Figure 3-10 Architectural view of the Virtual I/O Server

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 105.

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. Using 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 as well, by using the `brct1` 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.

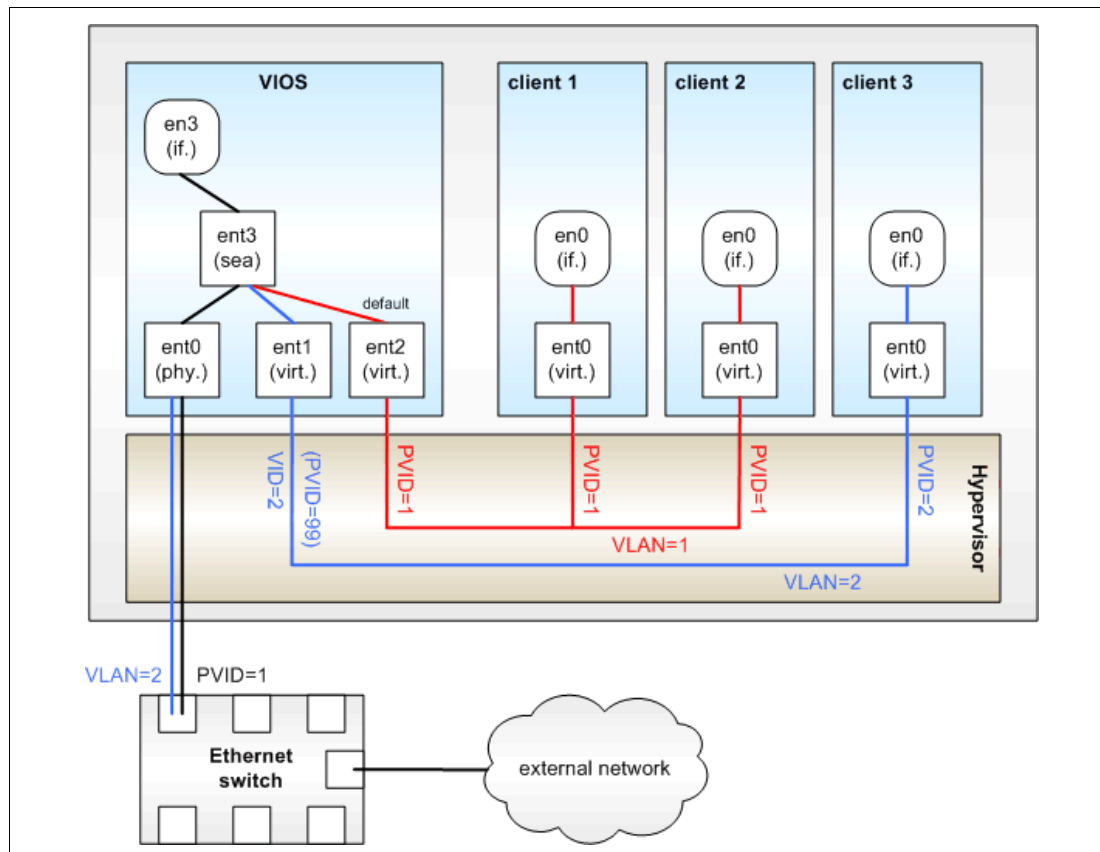


Figure 3-11 Architectural view of a Shared Ethernet Adapter

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, it is possible for a single physical Ethernet to be shared between 320 internal VLAN networks. The number of

shared Ethernet adapters that can be set up in a Virtual I/O server partition is limited only by the resource availability as there are no configuration limits.

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 across 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 very 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 approach 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 the following address:

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 different clients.
- ▶ As of Virtual I/O Server 1.5, files can be created on these disks and file backed storage devices can be created.

The Logical volumes or files can be assigned to separate partitions. Therefore, virtual SCSI enables sharing of adapters as well as disk devices.

Figure 3-12 on page 101 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.

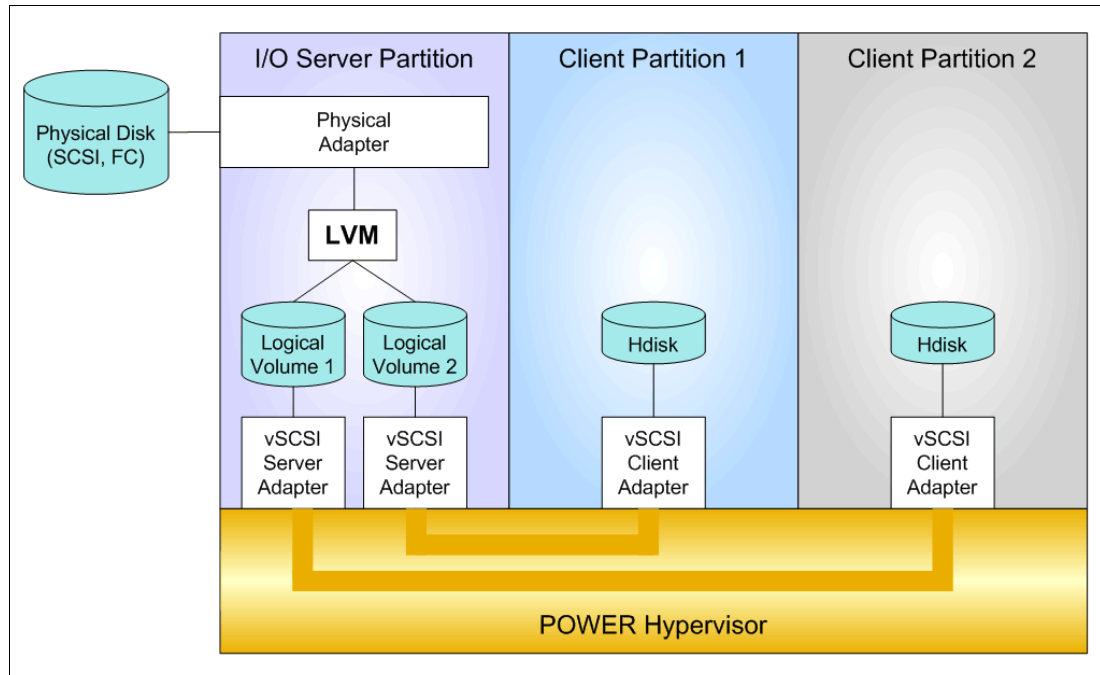


Figure 3-12 Architectural view of virtual SCSI

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 the following Web site:

<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, as follows:

- ▶ 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 102.
- ▶ Support for virtual SCSI devices backed by a file. These are then accessed as standard SCSI-compliant LUNs.
- ▶ Support for virtual Fibre Channel devices 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 are designed to ease the deployment of a virtualized infrastructure. For more information about the System Planning Tool, see 3.5, “System Planning Tool” on page 107.
- ▶ IBM Systems Director and a number of pre-installed Tivoli® agents are included, such as Tivoli Identity Manager, in order to allow easy integration into an existing Tivoli Systems Management infrastructure, and 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 command-line interface (CLI) statistics in **svmon**, **vmstat**, **fcstat** and **topas** commands
- ▶ 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.
- ▶ 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 technology based servers also support partition mobility.

Source and destination system requirements

The source partition must be one that only has virtual devices. If there are any physical devices in its allocation, they must be removed before the validation or migration is initiated. An 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 different firmware levels, but they must be compatible with each other.

Another possibility is to migrate partitions back and forth between POWER6 and POWER7 processor-based servers. Partition Mobility can take advantage of 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, consider performing the following steps:

1. Set the percolating 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 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 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 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 and shared storage (external hdisk with `reserve_policy=no_reserve`). In addition, systems must not be running on battery power, and all logical partitions should be on the same open network with RMC established to the HMC.

The HMC is used to configure, to validate, and to 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 which will 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. When the POWER7 processor technology 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 for the whole set of POWER6 and POWER7 processor-based servers
- ▶ When performing server maintenance, you get the additional flexibility to utilize 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 it is the system that 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, which 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, for partitions using AIX 5.3, AIX 6.1, IBM i 6.1, SLES 11, and RHEL 5.4.

3.4.9 Operating system support for PowerVM

Table 3-6 summarizes the PowerVM features that are supported by the operating systems and that are 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 V10 SP3 | 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 |

| Feature | AIX V5.3 | AIX V6.1 | IBM i 6.1.1 | RHEL V5.4 | SLES V10 SP3 | SLES V11 |
|-------------------------|----------|------------------|-------------|-----------|--------------|----------|
| 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 supports only two threads, and supports four threads as of TL4 SP3
- c. IBM i 6.1.1 and later support SMT4
- d. Available on selected models
- e. On AIX 6.1 with TL4 SP2 and later

3.4.10 POWER7 and 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

| Feature | 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 is 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 a release that is targeted for availability during the first half of 2010. For additional questions about availability of this release, contact Red Hat.

For information regarding Advanced Toolchain, see “How to use Advance Toolchain for Linux on Power” at the following Web site:

<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 can also plan for and design non-partitioned systems using the SPT. The resulting output of your design is called a *system plan*, which is stored in a `.sysplan` file. This file can contain plans for a single system or multiple systems. The `.sysplan` file can be used for the following 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 IBM Business Partner to use the Customer Specified Placement manufacturing option if you want to automatically deploy your partitioning environment on a new machine. SPT looks for the resource’s allocation to be the same as what is specified in your `.sysplan` file.

You can create an entirely new system configuration, or you can create a system configuration that is based on any of the following items:

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

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

You may use the SPT before ordering 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.

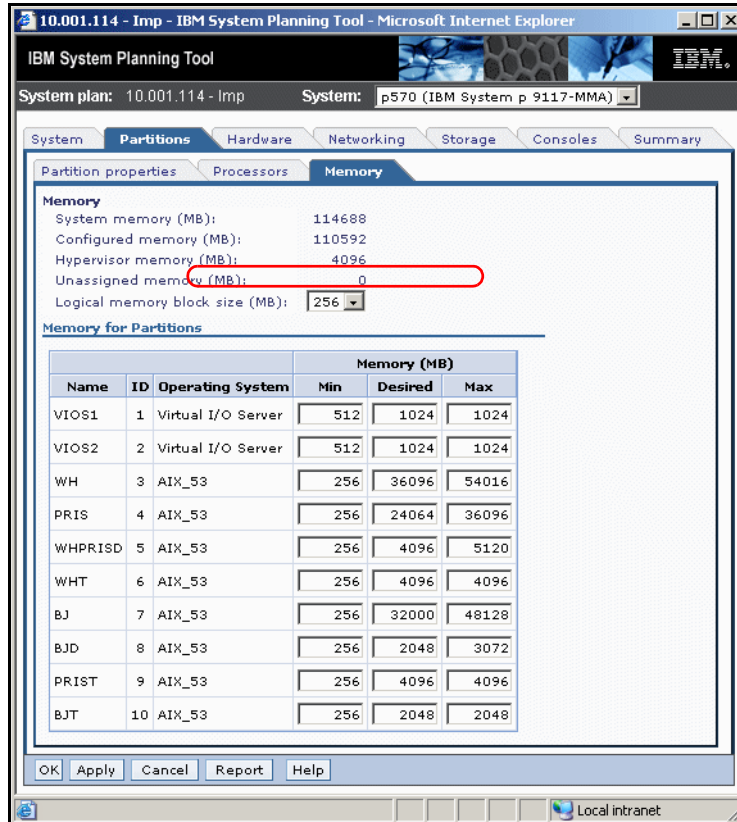


Figure 3-13 IBM System Planning Tool window showing Hypervisor memory requirements

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

<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 effects are communicated to users and services and how efficiently and nondisruptively the faults are repaired.

Each successive generation of IBM servers is designed to be more reliable than the previous server family. POWER7 processor-based servers have new features to support new levels of virtualization, help ease administrative burden and increase system utilization.

Reliability starts with components, devices, and subsystems designed to be fault-tolerant. POWER7 uses lower voltage technology improving reliability with stacked latches to reduce soft error (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 that is 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 is 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.

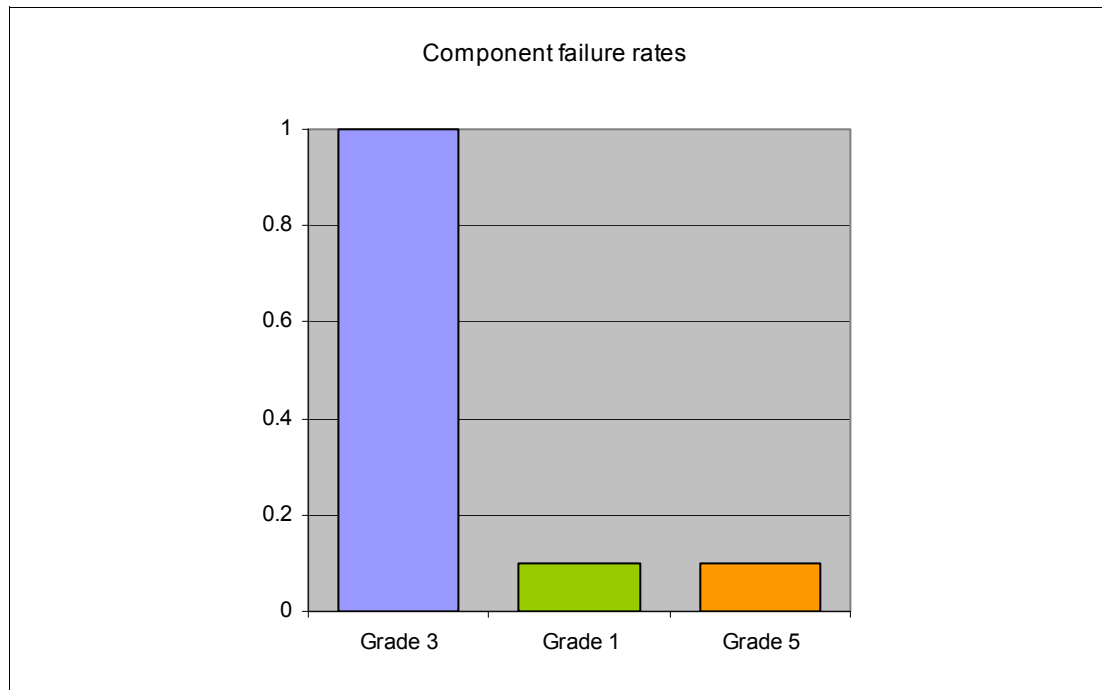


Figure 4-1 Component failure rates

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, and 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 them are:

- ▶ POWER7 cores include redundant bits in L1-I, L1-D, L2 caches, and L2 and L3 directories
- ▶ Power 770 and 780 main memory DIMMs contain an extra DRAM chip for improved redundancy
- ▶ Power 770 and 780 redundant system clock and service processor for configurations with more than two central electronics complex (CEC) drawers
- ▶ 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 offer and introduce significant enhancements that can increase system availability, and to drive towards a high availability objective with hardware components that can:

- ▶ Self-diagnose and self-correct during run time
- ▶ Automatically reconfigure to mitigate potential problems from suspect hardware
- ▶ Self-heal or to 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 122.

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 attempt to claim the needed resource. On a properly configured POWER processor-based server, this approach allows that capacity to be first obtained from a low priority partition instead of a high priority partition.

This capability is relevant to the total system availability because it gives the system an additional stage before an unplanned outage. In the event that insufficient resources exist to maintain full system availability, these servers attempt to maintain partition availability by user-defined priority.

Partition-availability priority is assigned to partitions by using a *weight value* or integer rating. The lowest priority partition rated at 0 (zero) and the highest priority partition valued at 255. The default value is set at 127 for standard partitions and 192 for Virtual I/O Server (VIOS) partitions. 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, 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, 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 approach prevents the same faulty hardware from affecting system operation again, and the repair action is deferred to a more convenient, less critical time.

Persistent deallocation includes:

- ▶ 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 Advanced System Management Interface (ASMI).



Figure 4-2 ASMI Auto Power Restart setting screen

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 approach significantly reduces exposure to both permanent and intermittent errors in the processor core.

Intermittent errors, often as a result 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 some logic functions, the POWER7 processor will first automatically retry 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 would have 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 capacity-on-demand (CoD) processor cores, the system will effectively put a CoD processor into operation after it has been determined that an activated processor is no longer operational. In this way the server will remain 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 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 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 across 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. The code can additionally mark a 2-bit symbol as bad and then correct it. Providing a double-error detect or single error correct ECC or a better level of protection is additional 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 replaced with a spare. 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, as explained in the previous paragraph.

► Hardware scrubbing

Hardware scrubbing is a method for dealing 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 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 would not affect more of what's 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.

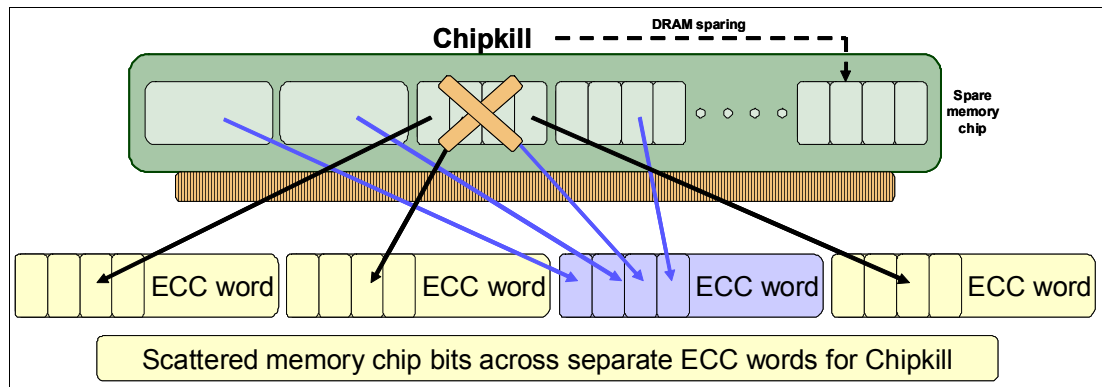


Figure 4-3 Chipkill in action with a spare memory DRAM chip on a Power 750 and 755

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 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 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.

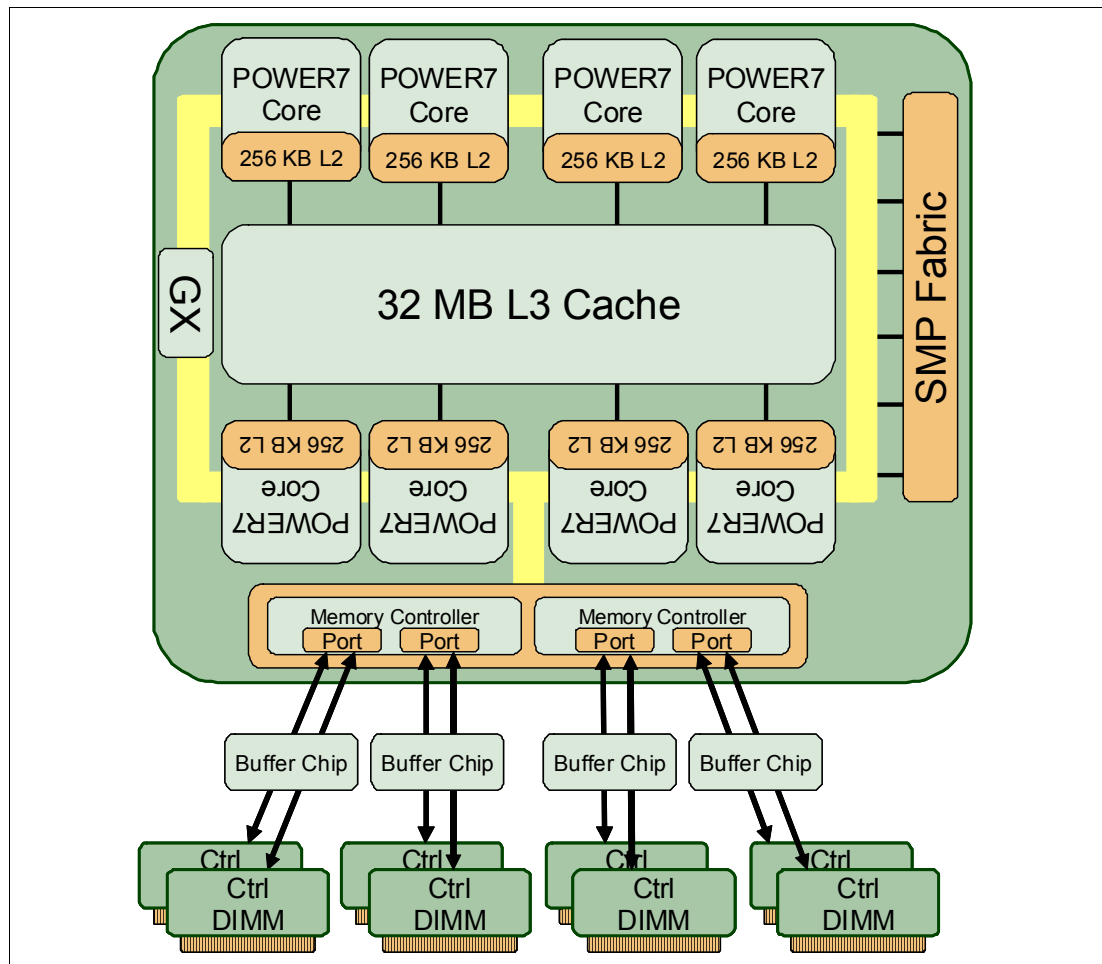


Figure 4-4 POWER7 memory subsystem

Memory page deallocation

Although coincident cell errors in separate memory chips are a statistic rarity, IBM POWER processor-based systems can contain these errors using a memory page deallocation scheme for partitions 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) 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 end users and applications.

The POWER Hypervisor maintains a list of pages marked for deallocation during the current platform IPL. During a partition IPL, the partition receives a list of all the bad pages in its address space. In addition, if memory is dynamically added to a partition (through a dynamic LPAR operation), the POWER Hypervisor warns the operating system when memory pages are included that need to be deallocated.

Finally, if an uncorrectable error in memory is discovered, the logical memory block that is associated with the address with the uncorrectable error is marked for deallocation by the POWER Hypervisor. This deallocation takes effect on a partition reboot if the logical memory block is assigned to an active partition at the time of the fault.

In addition, the system deallocates the entire memory group that is associated with the error on all subsequent system reboot operations until the memory is repaired. This approach is intended to guard against future uncorrectable errors while waiting for parts replacement.

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

Memory persistent deallocation

Defective memory 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, OS Service Focal Point, or 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, as well as L2 and L3 directories.

L1 instruction and data array protection

The POWER7 processor's instruction and data caches are protected against intermittent errors 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 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 that is detected in the cache can also trigger a purge and delete operation of the cache line. This occurrence results in no loss of operation because an unmodified copy of the data can be held in system memory to reload the cache line from main memory; modified data would be 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 would simply trigger 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. Note the following information:

- ▶ If the error is irrelevant, a check stop is not forced.
- ▶ If the data is used, termination can be limited to the program or kernel, or hypervisor owning the data; or 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 in the case where the corrupt data is used by the POWER Hypervisor does the entire system have to 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 may include the entire PHB chip, or simply a single bridge. This results in the loss of all I/O operations that use the frozen hardware until a power-on reset of the PHB. The impact to partitions depends on how the I/O is configured for redundancy. In a server configured for fail-over availability, redundant adapters spanning multiple PHB chips could 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 relative to high reliability. In certain cases, they might be more likely to encounter internal microcode errors, or many of the hardware errors described for the rest of the server.

The traditional means of handling these problems is through adapter internal error reporting and recovery techniques in combination with operating system device driver management and diagnostics. In certain cases, an error in the adapter might cause transmission of bad data on the PCI bus itself, resulting in a hardware-detected parity error and causing a global machine-check interrupt, eventually requiring a system reboot to continue.

PCI enhanced error handling (EEH) 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), allow the device driver to reset it, and continue without a system reboot. For Linux, EEH support extends to the majority of frequently used devices, although certain third-party PCI devices may 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 use ECC.

Figure 4-5 on page 121 shows the location and various mechanisms used throughout the I/O subsystem for PCI enhanced error handling.

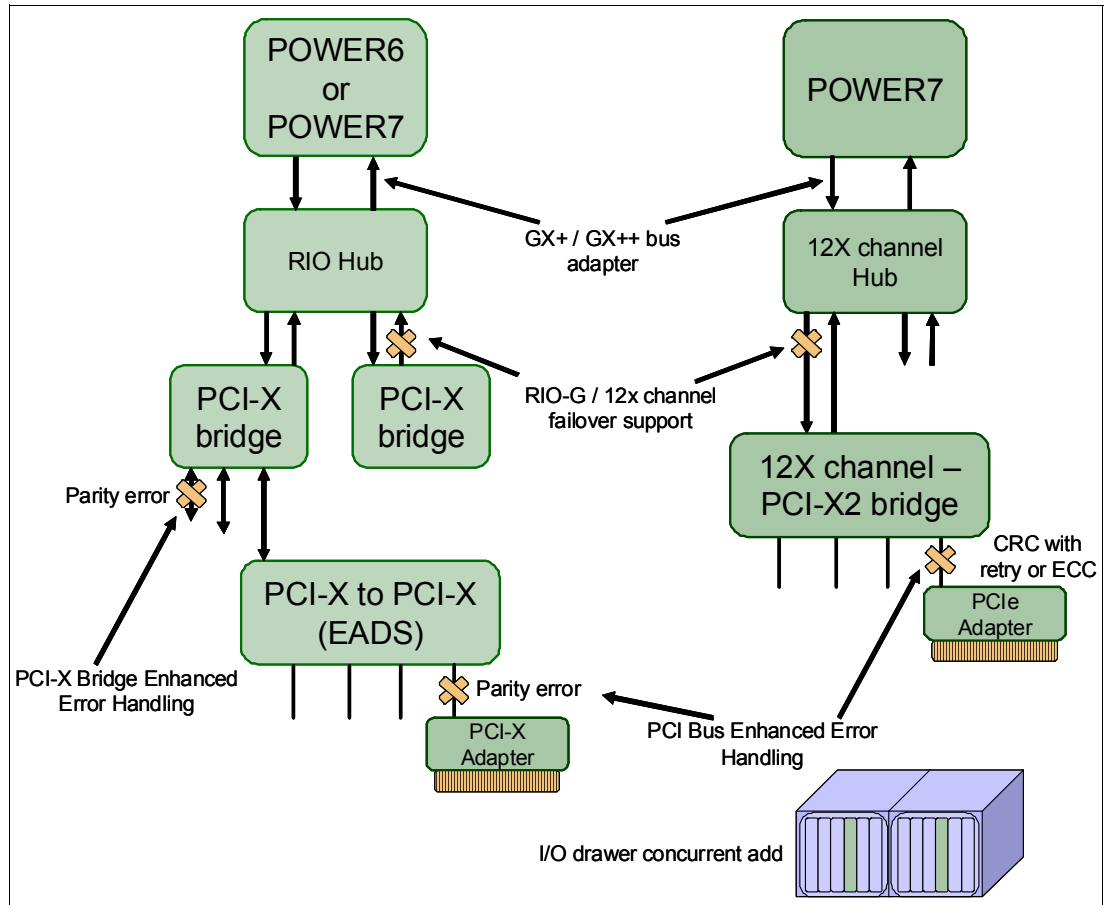


Figure 4-5 PCI enhanced error handling

4.3 Serviceability

IBM Power Systems design considers both IBM and the client's 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 found in all POWER processor-based systems.

4.3.1 Detecting

The first and most crucial component of a solid serviceability strategy is the ability to accurately and effectively detect errors when they occur. Although not all errors are a guaranteed threat to system availability, those that go undetected can cause problems because the system does not have the opportunity to evaluate and act if necessary. POWER processor-based systems employ IBM System z® server-inspired error detection mechanisms that extend from processor cores and memory to power supplies and hard drives.

Service processor

The service processor is a separately powered microprocessor, separate 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 environmental-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 if the temperature remains above the warning level for too long.
- Internal component temperatures reach critical levels.
- Non-redundant fan fails.

- ▶ **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 required for proper booting of the operating system, when the system is powered on at the initial install 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 into the system error log where it is analyzed by error log analysis (ELA) routines. Appropriate actions are taken to report the boot time error for subsequent service if required

- ▶ **Concurrent access to the service processors menus of the Advanced System Management Interface**

This 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, access all service processor functions without having to power-down the system to the standby state. This approach allows the administrator or service representative to dynamically access the menus from any Web browser-enabled console attached to the Ethernet service network concurrent with normal system operation.

- ▶ **Managing the interfaces for connecting uninterruptible power supply 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:

- ▶ Continual monitoring of system operations to detect potential calculation errors
- ▶ Attempt 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)

First-failure data capture (FFDC) is an error isolation technique, which ensures that when a fault is detected in a system through error checkers or other types of detection methods, the root cause of the fault 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 will be 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 an FRU.

This proactive diagnostic strategy is a significant improvement over the classic, less accurate “reboot and diagnose” service approaches.

Figure 4-6 on page 125 shows a schematic of a fault isolation register implementation.

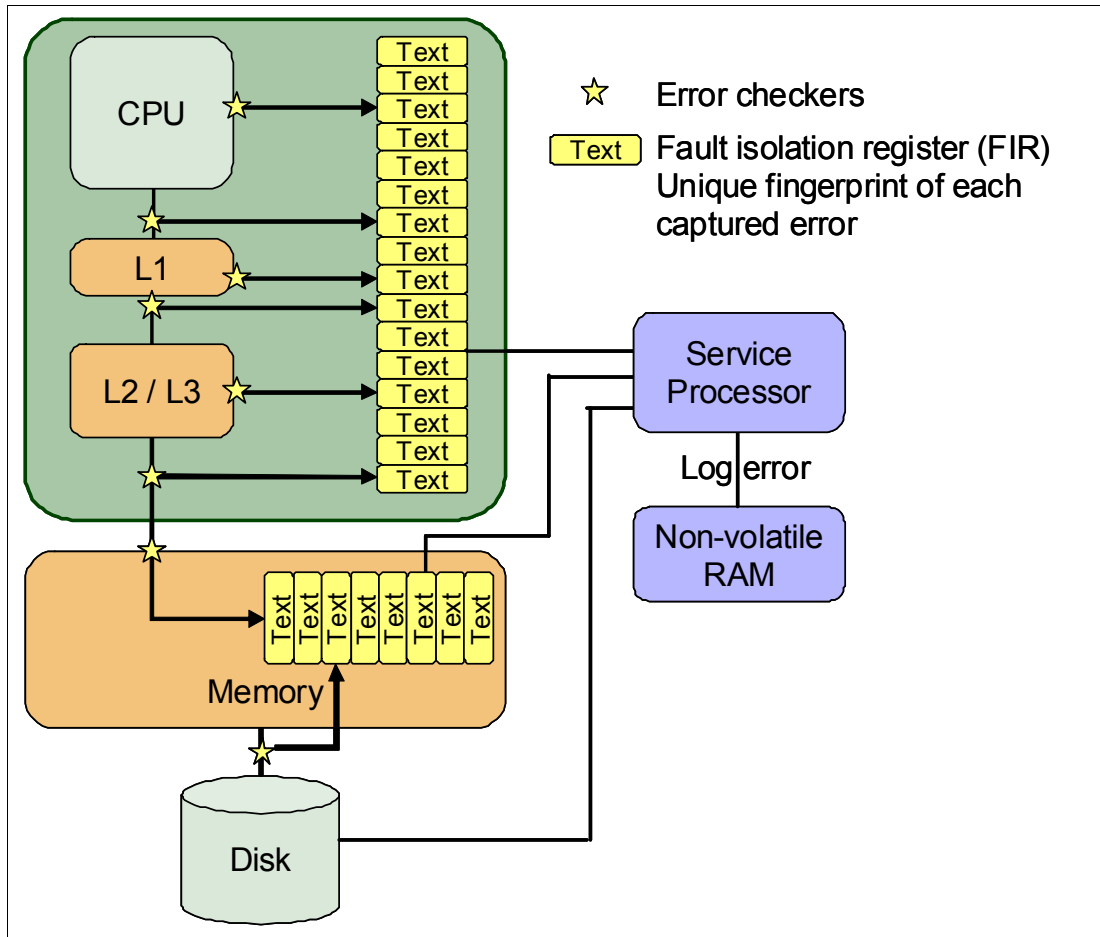


Figure 4-6 Schematic of a FIR implementation

Fault isolation

The service processor interprets error data 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 upon the isolation analysis, recoverable error threshold counts may be incremented. No specific service action could be 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 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 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 system processors are operational. Boot time BISTs might 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 checking 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 re-creation 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 will also illuminate 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 are sent to an IBM Service organization. Customer 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 like engineering and manufacturing levels
- ▶ Return codes
- ▶ Resource identifiers
- ▶ First-failure data capture 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 as well, 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 or 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)

EED is additional data that is collected either automatically at the time of a failure or manually at a later time. The data that is 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 some circumstances, an error may require a dump to be automatically or manually created. In this event, it will be off loaded to the HMC upon 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 which HMC that the dump is located on.

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 like 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, but 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 which requires client awareness in the event they want 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. Call home invokes the service organization in order for the appropriate service action to begin, automatically opening a problem report and in some cases also dispatching field support. This automated reporting provides faster and potentially more accurate transmittal of error information. Although configuring call home is optional, you 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 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 related to the error along with any service actions taken by the service representative are recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

4.3.5 Locating and servicing

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER processor-based systems utilize 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/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 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 like 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 the models 750 and 755, that might 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 will be illuminated, which will be rolled up to the system fault LED. The Light Path system pinpoints the exact part by turning on the amber FRU fault LED 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 (turning 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

For midrange and high-end systems, including models 770 and 780 and up, they usually are repaired by IBM support personnel.

The enclosure and system identify LEDs turn solidly on and 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 waits for the servicer's indication of what failure to attend first and then illuminates the LEDs to the failing component.

Data centers can be complex places, and Guiding Light is designed to do more than identify visible components. When a component might be hidden from view, Guiding Light can flash a sequence of LEDs that extend to the frame exterior, clearly *guiding* the service representative to the correct rack, system, enclosure, drawer, and component.

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:

- ▶ Location diagrams are strategically located on the system hardware, relating information regarding the placement of hardware components. Location diagrams might include location codes, drawings of physical locations, concurrent maintenance status, or other data 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.
- ▶ The remove or replace procedure labels contain procedures often found on a cover of the system or in other spots accessible to the servicer. These labels provide systematic procedures, including diagrams, detailing how to remove/replace certain serviceable hardware components.

- ▶ Numbered arrows are used to indicate the order of operation and serviceability direction of components. Certain serviceable parts such as latches, levers, and touch points must be pulled or pushed in a certain direction and certain order for the mechanical mechanisms to 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 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 allowing 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 might begin to wear from repeated plugging and unplugging. For this reason, these devices are specifically designed to be concurrently maintainable, when properly configured.

In other cases, a client may 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 will 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 fix pack updates when possible. The determination of whether a firmware fix pack release can be updated concurrently is identified in the *readme* file released with the firmware.

Blind-swap cassette

Blind-swap PCIe adapters represent significant service and ease-of-use enhancements in I/O subsystem design, and maintains 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, this is the case of 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 required on previous designs. The adapter is correctly connected by sliding in the cassette.

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 from any of the following sources:

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

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

IBM supports multiple firmware releases in the field, so 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 several 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, will require installation of a new firmware release level. This process is disruptive to server operations because it requires a scheduled outage and full server reboot.

In addition to concurrent and disruptive firmware updates, IBM also offers concurrent patches that include functions which are not activated until a subsequent server reboot. A server with these patches operates normally. The additional concurrent fixes is installed and activated when the system reboots after the next scheduled outage.

Additional capability is added to the firmware to be able to view the status of a system power control network background firmware update. This subsystem will update as necessary as migrated nodes or I/O drawers are added to the configuration. The new firmware 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 are designed to be used both by service representative providers who are familiar with the task at hand 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 on the latest updates available for service-related documentation. The latest version of the documentation is accessible through the Internet, and a CD-ROM-based version is also available.

4.4 Manageability

Several functions and tools help manageability, and can allow 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 130 and “Guiding Light” on page 130.
- ▶ Service Processor
Advanced System Management Interface
- ▶ 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 running its own operating system. It is a component of the service interface card.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface connected to the 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 using the ASMI, and enables complete system and partition management from the HMC.

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

The service processor uses two Ethernet 10/100/1000 Mbps 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.
- ▶ Because of firmware heavy workload, firmware can only support these ports at 10/100 Mbps rate, although the Ethernet MAC is capable of 1 Gbps
- ▶ Both Ethernet ports have a default IP address:
 - Service processor Eth0 or HMC1 port is configured as 169.254.2.147
 - Service processor Eth1 or HMC2 port is configured as 169.254.3.147

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

The Advanced System Management interface (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. Some repair procedures require connection to the ASMI.

The ASMI is accessible through the HMC. It is also accessible 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. Use the ASMI to change the service processor IP addresses or to apply some 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:

- ▶ Accessing the ASMI using an HMC
 - If configured to do so, the HMC connects directly to the ASMI for a selected system. To connect to the Advanced System Management interface from an HMC:
 - 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)**.

- ▶ Accessing the ASMI 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 connected to the service processor. The Web interface is available during all phases of system operation, including the initial program load (IPL) and run time. However, several menu options in the Web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides a Secure Sockets Layer (SSL) Web connection to the service processor. To establish an SSL connection, open your browser using `https://`.

Note: To make the connection through Internet Explorer, click **Tools Internet Options**. Clear the **Use TLS 1.0** box, and click **OK**.

- ▶ Accessing the ASMI using an ASCII terminal

The ASMI on an ASCII terminal supports a subset of the functions provided by the Web interface and is available only when the system is in the platform standby state. The ASMI on an ASCII console is not available during some phases of system operation, such as the IPL and run time.

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 out the panel to access the switches and to 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 operator panel features are as follows:

- ▶ 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 service tools problem log, IBM i history log (QHST), and IBM i problem log. The modes are:

- ▶ 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, some devices might require additional actions by the user or diagnostic application before testing can be done.

- ▶ Maintenance mode

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

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

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

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

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 super-set 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, take several kinds of system dumps, or take other options.

Depending on the operating system, the service level functions you typically see when using the operating system service menus are:

- ▶ 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 it detects and forwards the event to the Service Focal Point (SFP) application 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 functions available through the Service Focal Point on the Hardware Management Console include:

- ▶ 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 update
- ▶ 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 functions available through the Service Focal Point Lite on the Integrated Virtualization Manager include:

- ▶ 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 USB port number one on the rear of a Power 750 and 755 in case 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. Install the new levels of firmware 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 addressing this capability, see the following Web site:

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

For Power Systems, select the **Power** link. Figure 4-8 on page 140 is an example of the Support for IBM Power servers Web page.



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 up to date.

If there is an HMC to manage the server, the HMC interface can be used to view the levels of server firmware and power subsystem firmware that are installed and are available to download and install.

Each IBM Power Systems server has the following levels of server firmware and power subsystem firmware:

► **Installed level**

This is the level of server firmware or power subsystem firmware that has been installed and will be installed into memory after the managed system is powered off and powered on. It is installed on the temporary side of system firmware.

► **Activated level**

This is the level of server firmware or power subsystem firmware that is active and running in memory.

► **Accepted level**

This is the backup level of server or power subsystem firmware. You can return to this level of server or power subsystem firmware if you decide to remove the installed level. It is installed on the 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 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 system firmware file naming convention.

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 are different.

- ▶ 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 are designed to 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 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 collected through the Service Agent is made available to IBM service support representatives when they help answer your questions or diagnose problems. Installation and use of IBM Electronic Service Agent for problem reporting enables IBM to provide better support and service for your IBM server.

To learn how Electronic Services can work for you, visit the following site:

<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 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 Sparring | X | X | X | X | X | X |
| Processor Instruction Retry | X | X | X | X | X | X |
| Alternate Processor Recovery | X | X | X | X | X | X |

| RAS feature | AIX 5.3 | AIX 6.1 | IBM i | RHEL 5 | SLES 10 | SLES 11 |
|---|----------------|----------------|--------------|---------------|----------------|----------------|
| 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 | - | - | - |
| 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 |

| RAS feature | AIX 5.3 | AIX 6.1 | IBM i | RHEL 5 | SLES 10 | SLES 11 |
|--|---------|---------|-------|--------|---------|---------|
| 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 with 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 |
| 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 147. Note that some of the documents referenced here might be available in softcopy only.

- ▶ *IBM Power 770 and 780 Technical Overview and Introduction*, REDP-4639
- ▶ *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

POWER7 server data sheets and other resources are on the following Web pages:

- ▶ Active Memory Expansion: Overview and Usage Guide
http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=SA&subtype=WH&appname=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/>
- ▶ IBM 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&appname=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.
<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
- ▶ Power Instruction Set Architecture (ISA) Version 2.05
http://www.power.org/resources/reading/PowerISA_V2.05.pdf
- ▶ Power Instruction Set Architecture (ISA) Version 2.06
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>
- ▶ Virtual networking on AIX 5L
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



IBM Power 750 and 755 Technical Overview and Introduction



Features the POWER7 processor providing advanced multi-core technology

Discusses Power 755 model for high performance computing

Describes leading midrange performance

This IBM Redpaper publication is a comprehensive guide covering the IBM Power 750 and Power 755 servers supporting AIX, IBM i, and Linux operating systems. The goal of this paper is to introduce the major innovative Power 750 and 755 offerings and their prominent functions, including:

- ▶ The POWER7 processor available at frequencies of 3.0 GHz, 3.3 GHz, and 3.55 GHz
- ▶ The specialized POWER7 Level 3 cache that provides greater bandwidth, capacity, and reliability
- ▶ The 1 Gb or 10 Gb Integrated Virtual Ethernet adapter, included with each server configuration, and providing native hardware virtualization
- ▶ 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
- ▶ EnergyScale technology that provides features such as power trending, power-saving, capping of power, and thermal measurement.

Professionals who want to acquire a better understanding of IBM Power Systems products 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 550 system. This paper does not replace the latest marketing materials and configuration tools. It is intended as an additional source of information that, together with existing sources, 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