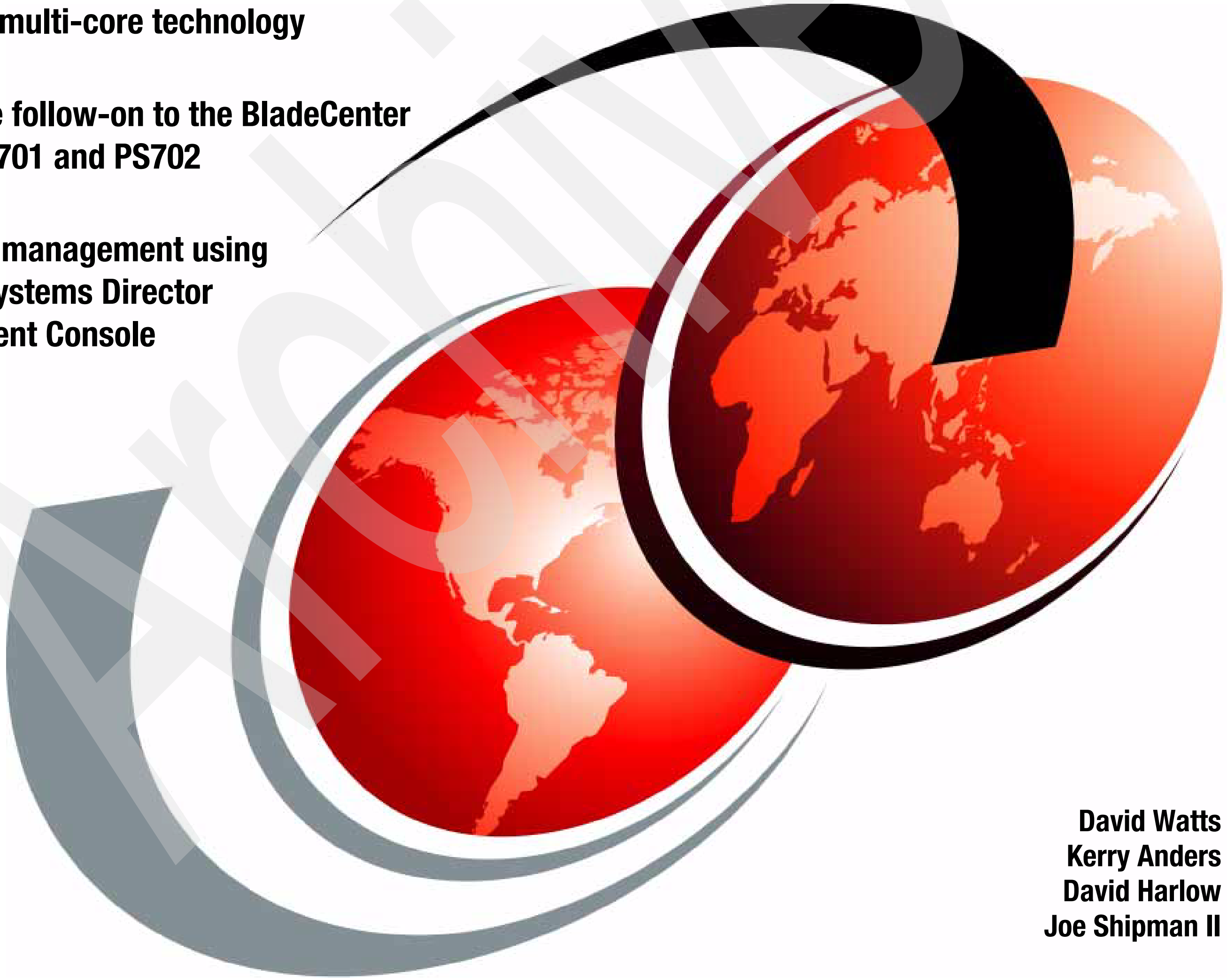
IBM BladeCenter PS703 and PS704 Technical Overview and Introduction

Features the POWER7 processor providing advanced multi-core technology

Details the follow-on to the BladeCenter PS700, PS701 and PS702

Describes management using the new Systems Director Management Console



David Watts
Kerry Anders
David Harlow
Joe Shipman II

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM BladeCenter PS703 and PS704 Technical Overview and Introduction

May 2011

REDP-4744-00

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

First Edition (May 2011)

This edition applies to:

IBM BladeCenter PS703, 7891-73X
IBM BladeCenter PS704, 7891-74X

Contents

Notices

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

Active Memory™
AIX 5L™
AIX®
AS/400®
BladeCenter®
DS4000®
DS8000®
Electronic Service Agent™
EnergyScale™
FlashCopy®
Focal Point™
IBM Systems Director Active Energy Manager™
IBM®
iSeries®
Micro-Partitioning™
POWER Hypervisor™
Power Systems™
Power Systems Software™
POWER4™
POWER5™
POWER6+™
POWER6®
POWER7™
PowerVM™
POWER®
pSeries®
Redbooks®
Redpaper™
Redbooks (logo) ®
ServerProven®
Solid®
System i®
System p5®
System Storage®
System x®
System z®
Tivoli®
Workload Partitions Manager™
XIV®

The following terms are trademarks of other companies:

BNT, and Server Mobility are trademarks or registered trademarks of Blade Network Technologies, Inc., an IBM Company.

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.

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

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

The IBM® BladeCenter® PS703 and PS704 are premier blades for 64-bit applications. They are designed to minimize complexity, improve efficiency, automate processes, reduce energy consumption, and scale easily. These blade servers are based on the IBM POWER7™ processor and support AIX®, IBM i, and Linux® operating systems. Their ability to coexist in the same chassis with other IBM BladeCenter blade servers enhances the ability to deliver the rapid return on investment demanded by clients and businesses.

This IBM Redpaper™ doocument is a comprehensive guide covering the IBM BladeCenter PS703 and PS704 servers. The goal of this paper is to introduce the offerings and their prominent features and functions.

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, Raleigh Center.

David Watts is a Consulting IT Specialist at the IBM ITSO Center in Raleigh. He manages residencies and produces IBM Redbooks® publications for hardware and software topics that are related to IBM System x® and IBM BladeCenter servers, and associated client platforms. He has authored over 80 books, papers, and web documents. He holds a Bachelor of Engineering degree from the University of Queensland (Australia) and has worked for IBM both in the U.S. and Australia since 1989. David is an IBM Certified IT Specialist and a member of the IT Specialist Certification Review Board.

Kerry Anders is a Consultant for POWER® systems and PowerVM™ in Lab Services for the IBM Systems and Technology Group, based in Austin, Texas. He supports clients in implementing IBM Power Systems™ blades using Virtual I/O Server, Integrated Virtualization Manager, and AIX. Kerry's prior IBM Redbooks publication projects include *IBM BladeCenter JS12 and JS22 Implementation Guide*, SG24-7655, *IBM BladeCenter JS23 and JS43 Implementation Guide*, SG24-7740, and *IBM BladeCenter PS700, PS701, and PS702 Technical Overview and Introduction*, REDP-4655. Previously, he was the Systems Integration Test Team Lead for the IBM BladeCenter JS21blade with IBM SAN storage using AIX and Linux. His prior work includes test experience with the JS20 blade, also using AIX and Linux in SAN environments. Kerry began his career with IBM in the Federal Systems Division supporting NASA at the Johnson Space Center as a Systems Engineer. He transferred to Austin in 1993.

David Harlow is a Senior Systems Engineer with business partner Mainline Information Systems, Inc. located in Tallahassee, Florida and he is based in Raleigh, North Carolina. His area of expertise includes Power Systems and Power Blade Servers using the IBM i operating system. He has 19 years of experience with the AS/400®, iSeries®, System i®, IBM i architecture, and IBM i operating systems. He has worked with the Power blade servers with VIOS hosting IBM i partitions since the POWER6® JS12 and JS22 entered marketing. He currently has several IBM certifications including the IBM Certified Technical Sales Expert - Power Systems with POWER7 and the IBM Certified Sales Expert - Power Systems with POWER7.

Joe Shipman II is a BladeCenter and System x Subject Matter Expert for the IBM Technical Support Center in Atlanta, Georgia. He has 7 years of experience working with servers and has worked at IBM for 5 years. His areas of expertise include IBM BladeCenter, System x, BladeCenter Fibre Channel fabrics, BladeCenter Networking, and Power Blade Servers. Previously he worked as an Electrical and Environmental Systems Specialist for the US Air Force for 10 years.

The team (l-r): Joe, David Harlow, Kerry, and David Watts

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

From IBM Power Systems development:

- Chris Austen
- Larry Cook
- John DeHart
- Kaena Freitas
- Bob Galbraith
- Jim Gallagher
- Seth Lewis
- Hoa Nguyen
- Amartey Pearson

From IBM Power Systems marketing:

- John Biebelhausen

From IBM Linux Technology Center:

- Jeff Scheel

This paper is based in part on *IBM BladeCenter PS700, PS701, and PS702 Technical Overview and Introduction*, REDP-4655. Thanks to the authors of that document:

- David Watts
- Kerry Anders
- Berjis Patel

Portions of this paper are from the book *Systems Director Management Console Introduction and Overview*, SG24-7860. Thanks to the authors of that document:

- Thomas Libor
- Allen Oh
- Lakshmikanthan Selvarajan
- Peter Wuestefeld

Now you can become a published author, too!

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

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

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

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

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

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

1

Introduction and general description

This chapter introduces and provides a general description of the new IBM BladeCenter POWER7 processor-based blade servers. These new blades offer processor scalability from 16 cores to 32 cores:

- IBM BladeCenter PS703: single-wide blade with two 8-core processors
- IBM BladeCenter PS704: double-wide blade with four 8-core processors

The new PS703 and PS704 blades are premier blades for 64-bit applications. They are designed to minimize complexity, improve efficiency, automate processes, reduce energy consumption, and scale easily.

The POWER7 processor-based PS703 and PS704 blades support AIX, IBM i, and Linux operating systems. Their ability to coexist in the same chassis with other IBM BladeCenter blade servers enhances the ability to deliver the rapid return on investment demanded by clients and businesses.

This chapter covers the following topics:

1.1 Overview of PS703 and PS704 blade servers

Figure 1-1 shows the IBM BladeCenter PS703 and PS704 blade servers.

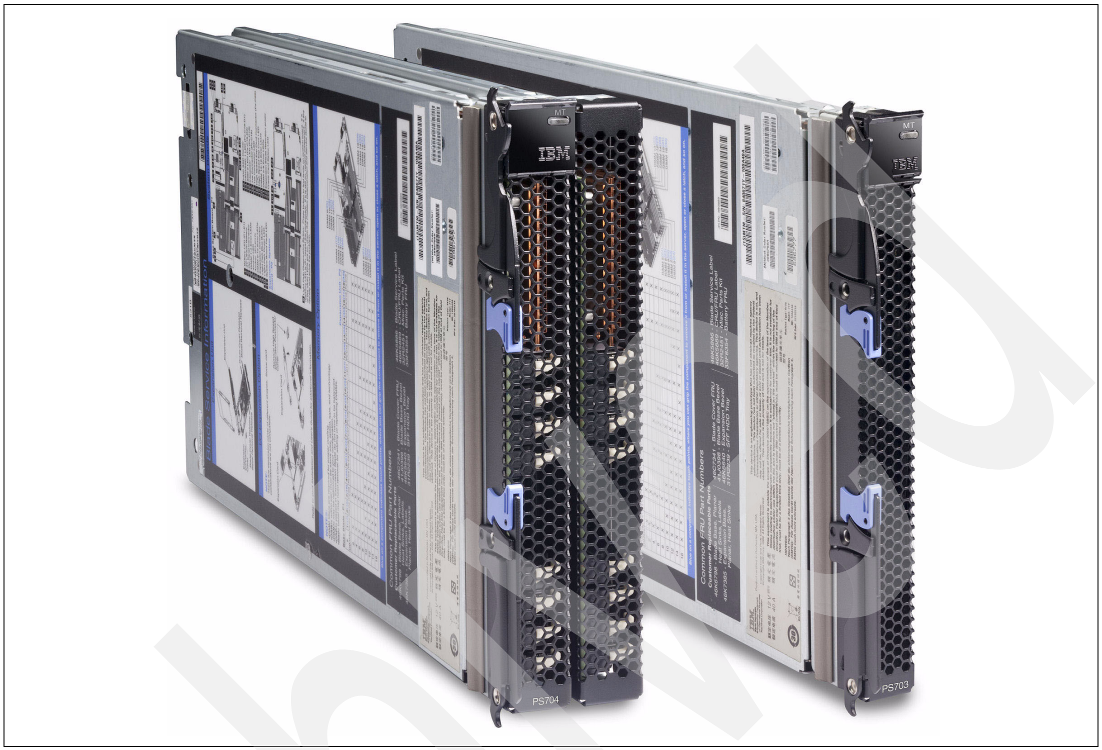


Figure 1-1 The IBM BladeCenter PS703 (right) and BladeCenter PS704 (left)

The PS703 blade server

The IBM BladeCenter PS703 (7891-73X) is a single-wide blade server with two eight-core POWER7 processors with a total of 16 cores. The processors are 64-bit 8-core 2.4 GHz processors with 256 KB L2 cache per core and 4 MB L3 cache per core.

The PS703 blade server has 16 DDR3 memory DIMM slots. The industry standard VLP DDR3 memory DIMMs are either 4 GB or 8 GB or 16 GB running at 1066 MHz. The minimum memory required for a PS703 blade server is 16 GB. The maximum memory that can be supported is 256 GB (16 x 16 GB DIMMs).

The PS703 blade server supports optional Active Memory™ Expansion, which is a POWER7 technology that allows the effective maximum memory capacity to be much larger than the true physical memory. Innovative compression/decompression of memory content using processor cycles can allow memory expansion up to 100%. This can allow an AIX 6.1 or later partition to do significantly more work with the same physical amount of memory, or a server to run more partitions and do more work with the same physical amount of memory.

The PS703 blade server has two onboard 1 Gb integrated Ethernet ports that are connected to the BladeCenter chassis fabric (midplane). The PS703 also has an integrated SAS controller that supports local (on-board) storage, integrated USB controller and Serial over LAN console access through the service processor, and the BladeCenter Advance Management Module.

The PS703 has one on-board disk drive bay. The on-board storage can be one 2.5-inch SAS HDD or two 1.8-inch SATA SSD drives (with the addition of an SSD interposer tray). The

PS703 also supports one PCIe CIOv expansion card slot and one PCIe CFFh expansion card slot. See 1.6.7, “I/O features” on page 21 for supported I/O expansion cards.

The PS704 blade server

The IBM BladeCenter PS704 (7891-74X) is a double-wide blade server with four eight-core POWER7 processors with a total of 32 cores. The processors are 64-bit 8-core 2.4 GHz processors with 256 KB L2 cache per core and 4 MB L3 cache per core.

The PS704 is a double-wide blade, meaning that it occupies two adjacent slots in the IBM BladeCenter chassis.

The PS704 blade server has 32 DDR3 memory DIMM slots. The industry standard VLP DDR3 memory DIMMs are either 4 GB or 8 GB running at 1066 MHz. The minimum memory required for PS704 blade server is 32 GB. The maximum memory that can be supported is 256 GB (32x 8 GB DIMMs).

The PS704 blade server supports optional Active Memory Expansion, which is a POWER7 technology that allows the effective maximum memory capacity to be much larger than the true physical memory. Innovative compression/decompression of memory content using processor cycles can allow memory expansion up to 100%. This can allow an AIX 6.1 or later partition to do significantly more work with the same physical amount of memory, or a server to run more partitions and do more work with the same physical amount of memory.

The PS704 blade server has four onboard 1 Gb integrated Ethernet ports that are connected to the BladeCenter chassis fabric (midplane). The PS704 also has an integrated SAS controller that supports local (on-board) storage, integrated USB controller and Serial over LAN console access through the service processor, and the BladeCenter Advance Management Module.

The PS704 blade server has two disk drive bays, one on the base blade and one on the expansion unit. The on-board storage can be one or two 2.5-inch SAS HDD or up to four 1.8-inch SSD drives. The integrated SAS controller supports RAID 0, 10, 5, or 6 depending on the numbers of HDDs or SSDs installed.

The PS704 supports two PCIe CIOv expansion card slots and two PCIe CFFh expansion card slots. See 1.6.7, “I/O features” on page 21 for supported I/O expansion cards.

Note: For the PS704 blade server, the service processor (FSP or just SP) in the expansion blade is set to IO mode, which provides control busses from IOs, but does not provide redundancy and backup operational support to the SP in the base blade.

1.2 Comparison between the PS70x blade servers

This section describes the difference between the five POWER7 blade servers:

- The PS700 is a single-wide blade with one 4-core 64-bit POWER7 3.0 GHz processor.
- The PS701 is a single-wide blade with one 8-core 64-bit POWER7 3.0 GHz processor.
- The PS702 is a double-wide blade with two 8-core 64-bit POWER7 3.0 GHz processors.
- The PS703 is a single-wide blade with two 8-core 64-bit POWER7 2.4 GHz processors.
- The PS704 is a double-wide blade with four 8-core 64-bit POWER7 2.4 GHz processors.

The POWER7 processor has 4 MB L3 cache per core and 256 KB L2 cache per core.

Table 1-1 compares the processor core options and frequencies, and L3 cache sizes of the POWER7 blade servers.

Table 1-1 Comparison of POWER7 blade servers

System	Number of processors	Cores per processor	Core frequency	L3 cache per processor	Minimum / Maximum memory	Form factor
PS700 blade	1	4	3.0 GHz	16 MB	4 GB / 64 GB	Single-wide
PS701 blade	1	8	3.0 GHz	32 MB	16 GB / 128 GB	Single-wide
PS702 blade	2	8	3.0 GHz	32 MB	32 GB / 256 GB	Double-wide
PS703 blade	2	8	2.4 GHz	32 MB	16 GB / 128 GB	Single-wide
PS704 blade	4	8	2.4 GHz	32 MB	32 GB / 256 GB	Double-wide

For a detailed comparison, see 2.6, “Technical comparison” on page 54.

Full details about the PS700, PS701, and PS702 can be found in the IBM Redpaper, *IBM BladeCenter PS700, PS701, and PS702 Technical Overview and Introduction*, REDP-4655 available from:

http://www.redbooks.ibm.com/abstracts/redp4655.html

1.3 IBM BladeCenter chassis support

Blade servers are thin servers that insert into a single rack-mounted chassis that supplies shared power, cooling, and networking infrastructure. Each server is an independent server with its own processors, memory, storage, network controllers, operating system, and applications. The IBM BladeCenter chassis is the container for the blade servers and shared infrastructure devices.

The IBM BladeCenter chassis can contain a mix of POWER, Intel®, Cell, and AMD processor-based blades. Depending on the IBM BladeCenter chassis selected, combinations of Ethernet, SAS, Fibre Channel, and FCoE I/O fabrics can also be shared within the same chassis.

All chassis can offer full redundancy for all shared infrastructure, network, and I/O fabrics. Having multiple power supplies, network switches, and I/O switches contained within a BladeCenter chassis eliminates single points of failure in these areas.

The following sections describe the BladeCenter chassis that support the PS703 and PS704 blades. For a comprehensive look at all aspects of BladeCenter products see the IBM Redbooks publication, *IBM BladeCenter Products and Technology*, SG24-7523, available from the following web page:

http://www.redbooks.ibm.com/abstracts/sg247523.html

Refer to the *BladeCenter Interoperability Guide* for complete coverage of the compatibility information. The latest version can be downloaded from the following address:

http://ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5073016

1.3.1 Supported BladeCenter chassis

The PS703 and PS704 blades are supported in the IBM BladeCenter chassis as listed in Table 1-2.

Table 1-2 The blade servers supported in each BladeCenter chassis

Blade	Machine type-model	Blade width	BC S 8886	BC E 8677	BC T 8720	BC T 8730	BC H 8852	BC HT 8740	BC HT 8750
PS703	7891-73X	1 slot	Yes	No	No	No	Yes	Yes	Yes
PS704	7891-74X	2 slot	Yes	No	No	No	Yes	Yes	Yes

IBM BladeCenter H delivers high performance, extreme reliability, and ultimate flexibility for the most demanding IT environments. See “BladeCenter H” on this page.

IBM BladeCenter HT models are designed for high-performance flexible telecommunications environments by supporting high-speed networking technologies (such as 10G Ethernet). They provide a robust platform for NGNs. See “BladeCenter HT” on page 7.

IBM BladeCenter S combines the power of blade servers with integrated storage, all in an easy-to-use package designed specifically for the office and distributed enterprise environments. See “BladeCenter S” on page 10.

Note: The number of blade servers that can be installed into chassis is dependent on the power supply configuration, power supply input (110V/208V BladeCenter S only) and power domain configuration options. See 1.3.2, “Number of PS703 and PS704 blades in a chassis” on page 12 for more information.

BladeCenter H

IBM BladeCenter H delivers high performance, extreme reliability, and ultimate flexibility to even the most demanding IT environments. In 9 U of rack space, the BladeCenter H chassis can contain up to 14 blade servers, 10 switch modules, and four power supplies to provide the necessary I/O network switching, power, cooling, and control panel information to support the individual servers.

The chassis supports up to four traditional fabrics using networking switches, storage switches, or pass-through devices. The chassis also supports up to four high-speed fabrics for support of protocols such as 4X InfiniBand or 10 Gigabit Ethernet. The built-in media tray includes light path diagnostics, two front USB 2.0 inputs, and an optical drive.

Figure 1-2 displays the front view of an IBM BladeCenter H and Figure 1-3 displays the rear view.

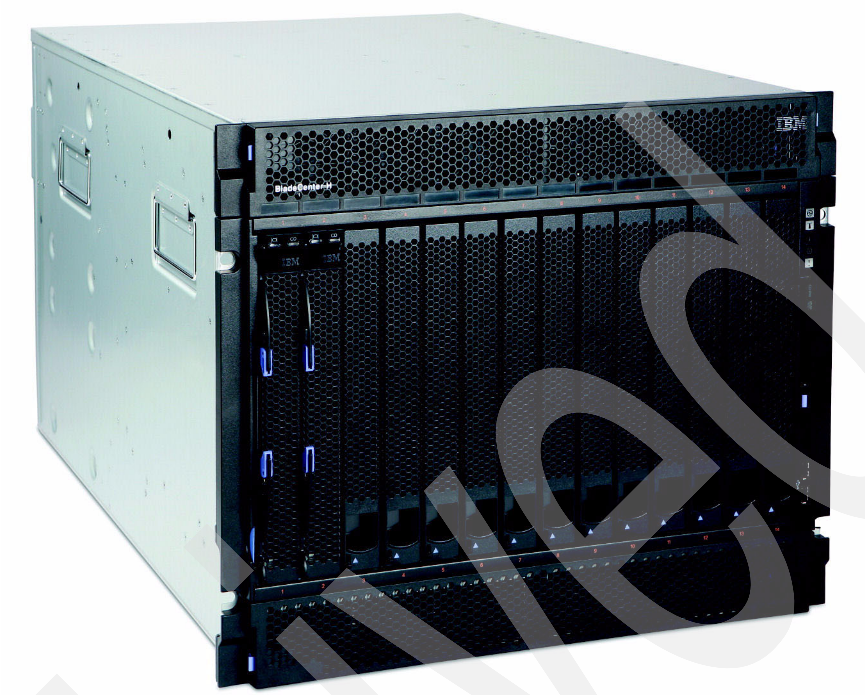


Figure 1-2   BladeCenter H front view

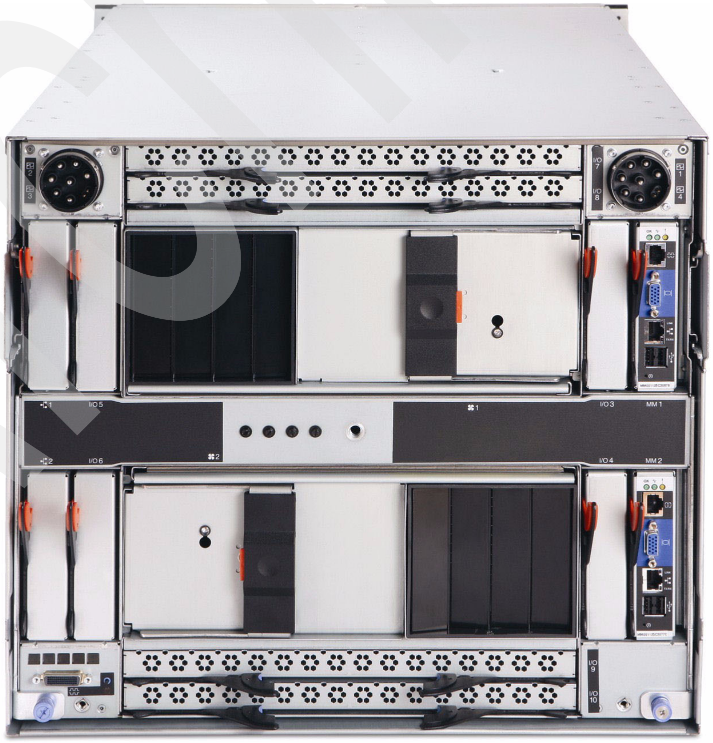


Figure 1-3   BladeCenter H rear view

The key features of the IBM BladeCenter H chassis are as follows:

- A rack-optimized, 9 U modular design enclosure for up to 14 hot-swap blades.
- A high-availability mid-plane that supports hot-swap of individual blades.
- Two 2,900 watt or 2,980 watt hot-swap power modules and support for two optional 2,900 watt or 2,980 watt power modules, offering redundancy and power for robust configurations (cannot mix power module types).

> **Power supply requirements:** BladeCenter H model 8852-4TX has 2,980 watt power supplies. Other models have 2,900 W powers supplies and the 2,980 W supplies are optional.
>
> The PS703 and PS704 do not require the 2,980 watt power supply. They are designed to fully function with both the 2,900 watt and 2,980 watt power supplies.

- Two hot-swap redundant blowers. Two additional hot-swap fan modules are included with additional power module option.

> **Blower requirements:** BladeCenter H model 8852-4TX has enhanced blowers compared with standard blowers in model 8852-4SX and earlier models. The enhanced blowers are optional in the model 8852-4SX and earlier models.
>
> The PS700, PS701, PS702, PS703, and PS704 do not require the enhanced blowers. They are designed to fully function with both the standard and the enhanced blowers.

- An Advanced Management Module that provides chassis-level solutions, simplifying deployment and management of your installation.
- Support for up to four network or storage switches or pass-through modules.
- Support for up to four bridge modules.
- A light path diagnostic panel, and two USB 2.0 ports.
- Serial port breakout connector.
- Support for UltraSlim Enhanced SATA DVD-ROM and multi-burner drives.
- IBM Systems Director and Tivoli® Provisioning Manager for OS Deployments for easy installation and management.
- Energy-efficient design and innovative features to maximize productivity and reduce power usage.
- Density and integration to ease data center space constraints.
- Help in protecting your IT investment through IBM BladeCenter family longevity, compatibility, and innovation leadership in blades.
- Support for the latest generation of IBM BladeCenter blades, helping provide investment protection.

BladeCenter HT

The IBM BladeCenter HT is a 12-server blade chassis designed for high-density server installations, typically for telecommunications use. It offers high performance with the support of 10 Gb Ethernet installations. This 12 U high chassis with DC or AC power supplies provides a cost-effective, high performance, high availability solution for telecommunication networks and other rugged non-telecommunications environments. The IBM BladeCenter HT chassis is positioned for expansion, capacity, redundancy, and carrier-grade NEBS level 3/ETSI compliance in DC models.

BladeCenter HT provides a solid foundation for next-generation networks (NGN), enabling service providers to become on demand providers. IBM's technological expertise in the enterprise data center, coupled with the industry know-how of key business partners, delivers added value within service provider networks.

Figure 1-4 shows the front view of the BladeCenter HT.

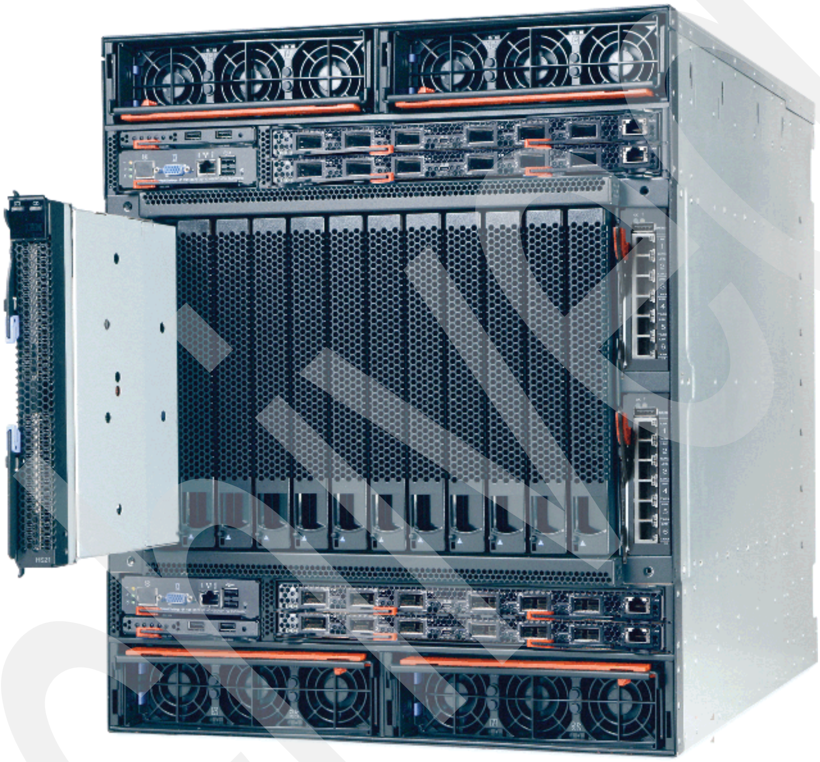


Figure 1-4 BladeCenter HT front view

Figure 1-5 shows the rear view of the BladeCenter HT.

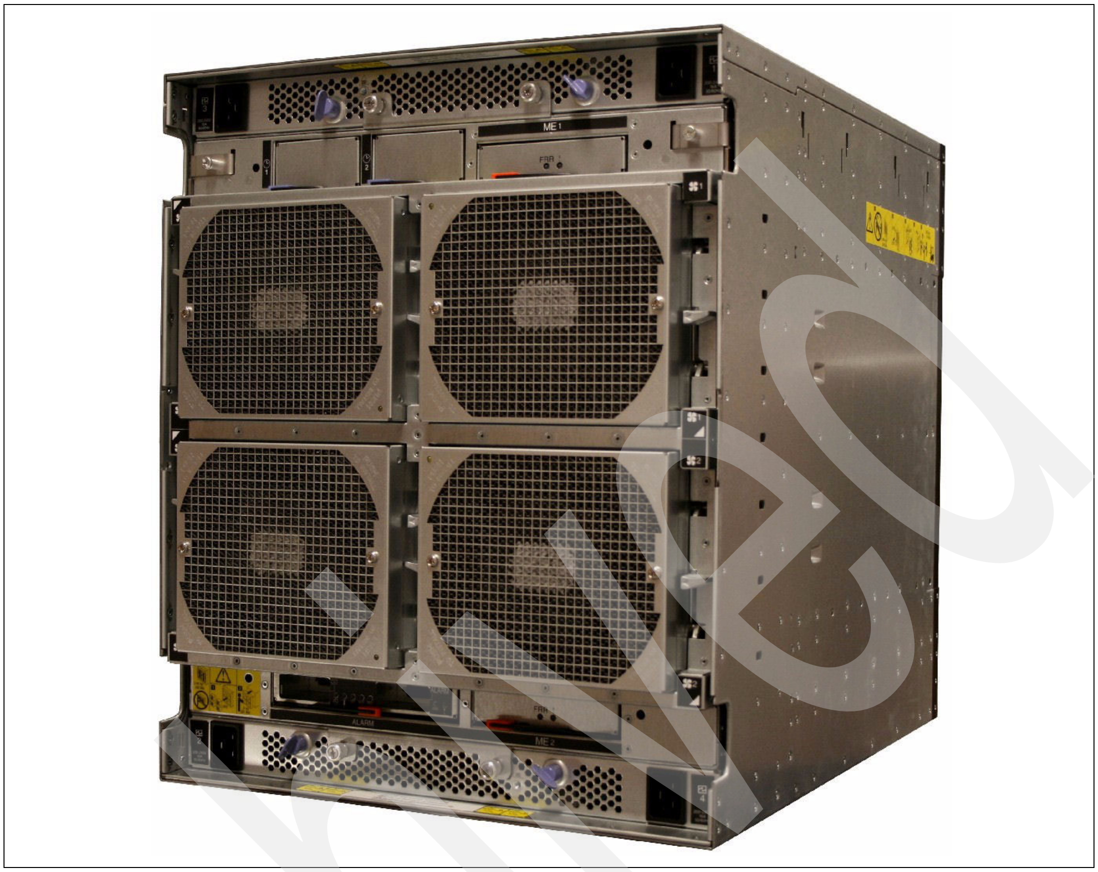


Figure 1-5 BladeCenter HT rear view

BladeCenter HT delivers rich telecommunications features and functionality, including integrated servers, storage and networking, fault-tolerant features, optional hot-swappable redundant DC or AC power supplies and cooling, and built-in system management resources. The result is a Network Equipment Building Systems (NEBS-3) and ETSI-compliant server platform optimized for next-generation networks.

The following BladeCenter HT applications are well suited for these servers:

- Network management and security
  - Network management engine
  - Internet cache engine
  - RSA encryption
  - Gateways
  - Intrusion detection
- Network infrastructure
  - Softswitch
  - Unified messaging
  - Gateway/Gatekeeper/SS7 solutions
  - VOIP services and processing
  - Voice portals
  - IP translation database

The key features of the BladeCenter HT are as follows:

- Support for up to 12 blade servers, compatible with the other chassis in the BladeCenter family
- Four standard and four high-speed I/O module bays, compatible with the other chassis in the BladeCenter family
- A media tray at the front with light path diagnostics, two USB 2.0 ports, and optional compact flash memory module support
- Two hot-swap management-module bays (one management module standard)
- Four hot-swap power-module bays (two power modules standard)
- New serial port for direct serial connection to installed blades
- Compliance with the NEBS 3 and ETSI core network specifications

BladeCenter S

The BladeCenter S chassis can hold up to six blade servers, and up to 12 hot-swap 3.5-inch SAS or SATA disk drives in just 7 U of rack space. It can also include up to four C14 950-watt/1450-watt power supplies. The BladeCenter S offers the necessary I/O network switching, power, cooling, and control panel information to support the individual servers.

The IBM BladeCenter S is one of five chassis in the BladeCenter family. The BladeCenter S provides an easy IT solution to the small and medium office and to the distributed enterprise. Figure 1-6 shows the front view of the IBM BladeCenter S.

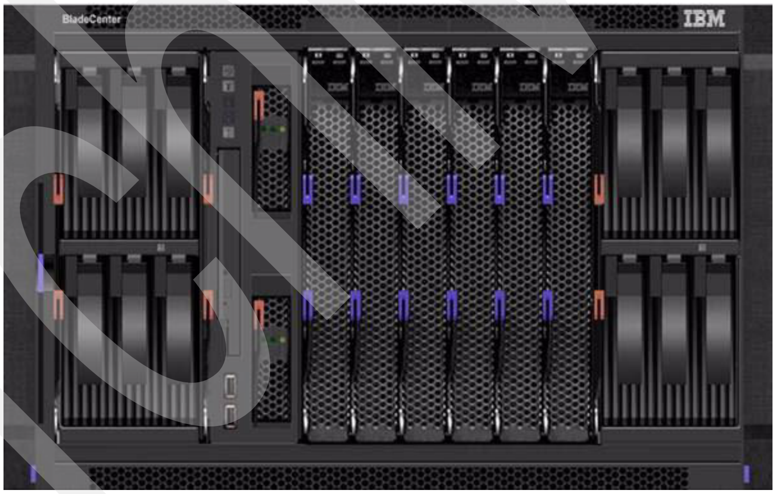


Figure 1-6 The front of the BladeCenter S chassis

Figure 1-7 shows the rear view of the chassis.

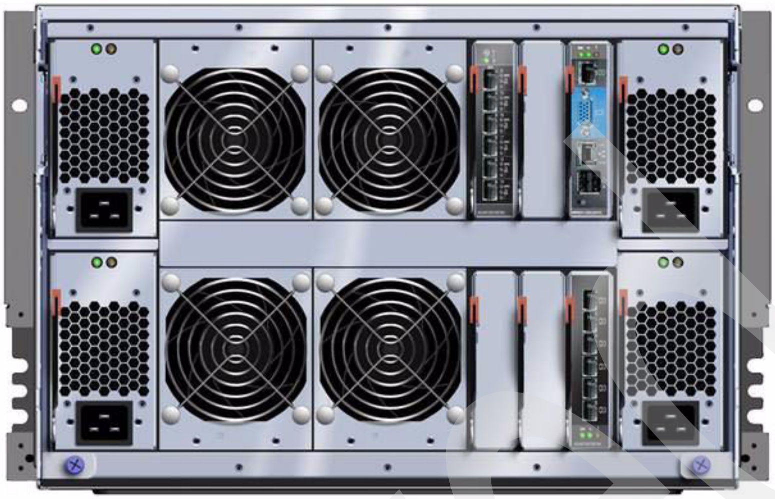


Figure 1-7 The rear of the BladeCenter S chassis

The key features of IBM BladeCenter S chassis are as follows:

- A rack-optimized, 7 U modular design enclosure for up to six hot-swap blades
- Two optional Disk Storage Modules for HDDs, six 3.5-inch SAS/SATA drives each
- High-availability mid-plane that supports hot-swap of individual blades
- Two 950-watt/1450-watt, hot-swap power modules and support for two optional 950/1450-watt power modules, offering redundancy and power for robust configurations
- Four hot-swap redundant blowers, plus one fan in each power supply
- An Advanced Management Module that provides chassis-level solutions, simplifying deployment and management of your installation
- Support for up to four network or storage switches or pass-through modules
- A light path diagnostic panel, and two USB 2.0 ports
- Support for optional UltraSlim Enhanced SATA DVD-ROM and Multi-Burner Drives
- Support for SAS RAID Controller Module to make it easy for clients to buy the all-in-one BladeCenter S solution
- IBM Systems Director, Storage Configuration Manager (SCM), Start Now Advisor, and Tivoli Provisioning Manager for OS Deployments support for easy installation and management
- Energy-efficient design and innovative features to maximize productivity and reduce power usage
- Help in protecting your IT investment through IBM BladeCenter family longevity, compatibility, and innovation leadership in blades
- Support for the latest generation of IBM BladeCenter blades, helping provide investment protection

1.3.2 Number of PS703 and PS704 blades in a chassis

The number of POWER7 processor-based blades that can be installed in a BladeCenter chassis depends on several factors:

- BladeCenter chassis type
- Number of power supplies installed
- Power supply voltage option (BladeCenter S only)
- BladeCenter power domain configuration

Table 1-3 shows the maximum number of PS703 and PS704 blades running in a maximum configuration (memory, disk, expansion cards) for each supported BladeCenter chassis that can be installed with fully redundant power and without performance reduction. IBM blades that are based on processor types other than POWER7 might reduce these numbers.

> **Tip:** As shown in Table 1-3, there is no restriction to the number of POWER7 blade servers that you can install in a BladeCenter chassis other than the number of power supplies installed in the chassis.

Table 1-3 PS703 and PS704 blades per chassis type

	BladeCenter H		**BladeCenter HT**		**BladeCenter S**			
	14 Slots Total		12 Slots Total		6 Slots Total			
					110VAC		208VAC	
Server	2 PS	4 PS	2 PS	4 PS	2 PS	4 PS	2 PS	4 PS
PS703	7	14	6	12	2	6	2	6
PS704	3	7	3	6	1	3	1	3

When mixing blades of different processor types in the same BladeCenter, the BladeCenter Power Configurator tool helps determine whether the combination desired is valid. It is expected that this tool will be updated to include the PS703 and PS704 blade configurations. For more information about this update, see the following web page:

http://www.ibm.com/systems/bladecenter/powerconfig

1.4 Operating environment

In this section, we list the operating environment specifications for the PS703 and PS704 blade servers and BladeCenter H and S.

PS703 and PS704

- Operating temperature
  - 10°C - 35°C (50°F - 95°F) at 0 - 914 meters altitude (0 - 3000 feet)
  - 10°C - 32°C (50°F - 90°F) at 914 - 2133 meters altitude (3000 - 7000 feet)
- Relative Humidity 8% - 80%
- Maximum Altitude 2133 meters (7000 ft.)

IBM BladeCenter H

- Operating temperature
  - 10.0°C - 35 °C (50°F - 95 °F) at 0 - 914 m (0 - 3000 ft.)
  - 10.0°C - 32 °C (50°F - 90 °F) at 914 - 2133 m (3000 - 7000 ft.)
- Relative humidity 8% - 80%
- Maximum altitude: 2133 meters (7000 ft.)

IBM BladeCenter S

- – Operating Temperature:
  - 10°C - 35°C (50°F - 95°F) at 0 - 914 m (0 - 3000 ft.)
  - 10°C - 32°C (50°F - 90°F) at 914 - 2133 m (3000 - 7000 ft.)
- Relative humidity: 8% - 80%
- Maximum altitude: 2133 meters (7000 ft.)

BladeCenter HT

- Operating temperature
  - 5°C - 40°C (41°F - 104 °F) at -60 - 1800 m (-197 - 6000 ft.)
  - 5°C - 30°C (41°F - 86 °F) at 1800 - 4000 m (6000 - 13000 ft.)
- Relative humidity 5% - 85%
- Maximum altitude: 4000 meters (13000 ft.)

1.5 Physical package

The PS703 and PS704 blade servers are supported in BladeCenter H, HT, and S.

This section describes the physical dimensions of the POWER7 blade servers and the supported BladeCenter chassis only. Table 1-4 shows the physical dimensions of the PS703 and PS704 blade servers.

Table 1-4 Physical dimensions of PS703 and PS704 blade servers

Dimension	PS703 blade server	PS704 blade server
Height	9.65 inch (245 mm)	9.65 inch (245 mm)
Width	1.14 inch (29 mm) Single-wide blade	2.32 inch (59 mm) Double-wide blade
Depth	17.55 inch (445 mm)	17.55 inch (445 mm)
Weight	9.6 lbs (4.35 kg)	19.2 lbs (8.7 kg)

Table 1-5 shows the physical dimension of the BladeCenter chassis that supports the POWER7 processor-based blade servers.

Table 1-5 Physical dimension of Supported BladeCenter chassis

Dimension	BladeCenter H	BladeCenter S	BladeCenter HT
Height	15.75 inch (400 mm)	12 inch (305 mm)	21 inch (528 mm)
Width	17.4 inch (442 mm)	17.5 inch (445 mm)	17.4 inch (442 mm)
Depth	28 inch (711 mm)	28.9 inch (734 mm)	27.8 inch (706 mm)

1.6 System features

The PS703 and PS704 blade servers are 16-core and 32-core POWER7 processor-based blade servers.This section describes the features on each of the POWER7 blade servers. The following topics are covered:

- 1.6.1, “PS703 system features” on page 14
- 1.6.2, “PS704 system features” on page 16
- 1.6.3, “Minimum features for the POWER7 processor-based blade servers” on page 18
- 1.6.4, “Power supply features” on page 19
- 1.6.5, “Processor” on page 20
- 1.6.6, “Memory features” on page 20
- 1.6.7, “I/O features” on page 21
- 1.6.8, “Disk features” on page 26
- 1.6.9, “Standard onboard features” on page 26

1.6.1 PS703 system features

The BladeCenter PS703 is shown in Figure 1-8.

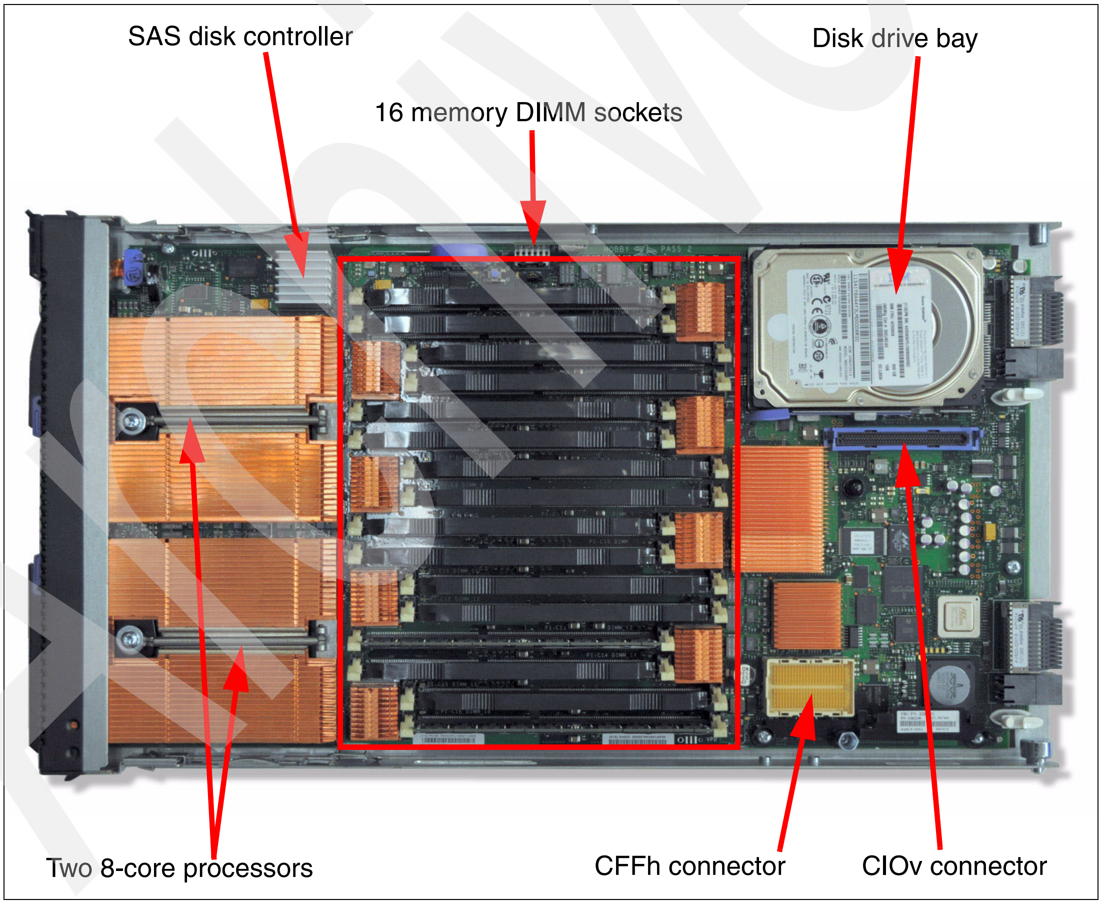


Figure 1-8 Top view of the PS703 blade server

The features of the server are as follows:

- Machine type and model number

  7891-73X

- Form factor

  Single-wide (30 mm) blade
- Processors:
  - Two eight-core 64-bit POWER7 processors operating at a 2.4 GHz clock speed for a total of 16 cores in the blade server
  - Based on CMOS 12S 45 nm SOI (silicon-on-insulator) technology
  - Power consumption is 110 W per socket
  - Single-wide (SW) Blade package
- Memory
  - 16 DIMM slots
  - Minimum 16 GB, maximum capacity 256 GB (using 16 GB DIMMs)
  - Industry standard VLP DDR3 DIMMs
  - Optional Active Memory Expansion
- Disk
  - 3 Gb SAS disk storage controller
  - One disk drive bay which supports one 2.5-inch SAS HDD (hard disk drive) or two 1.8-inch SATA SSD (solid state drive)
  - Hardware mirroring:
    - One HDD: RAID 0
    - One SSD: RAID 0
    - Two SSDs: RAID 0 or RAID 10
- On-board integrated features:
  - Service processor (SP)
  - Two 1 Gb Ethernet ports
  - One SAS Controller
  - USB Controller which routes to the USB 2.0 port on the media tray
  - 1 Serial over LAN (SOL) Console through SP
- Expansion Card I/O Options:
  - One CIOv expansion card slot (PCIe)
  - One CFFh expansion card slot (PCIe)

1.6.2 PS704 system features

The PS704 is a double-wide server. The two halves of the BladeCenter PS704 are shown in Figure 1-9 on this page and Figure 1-10 on page 17.

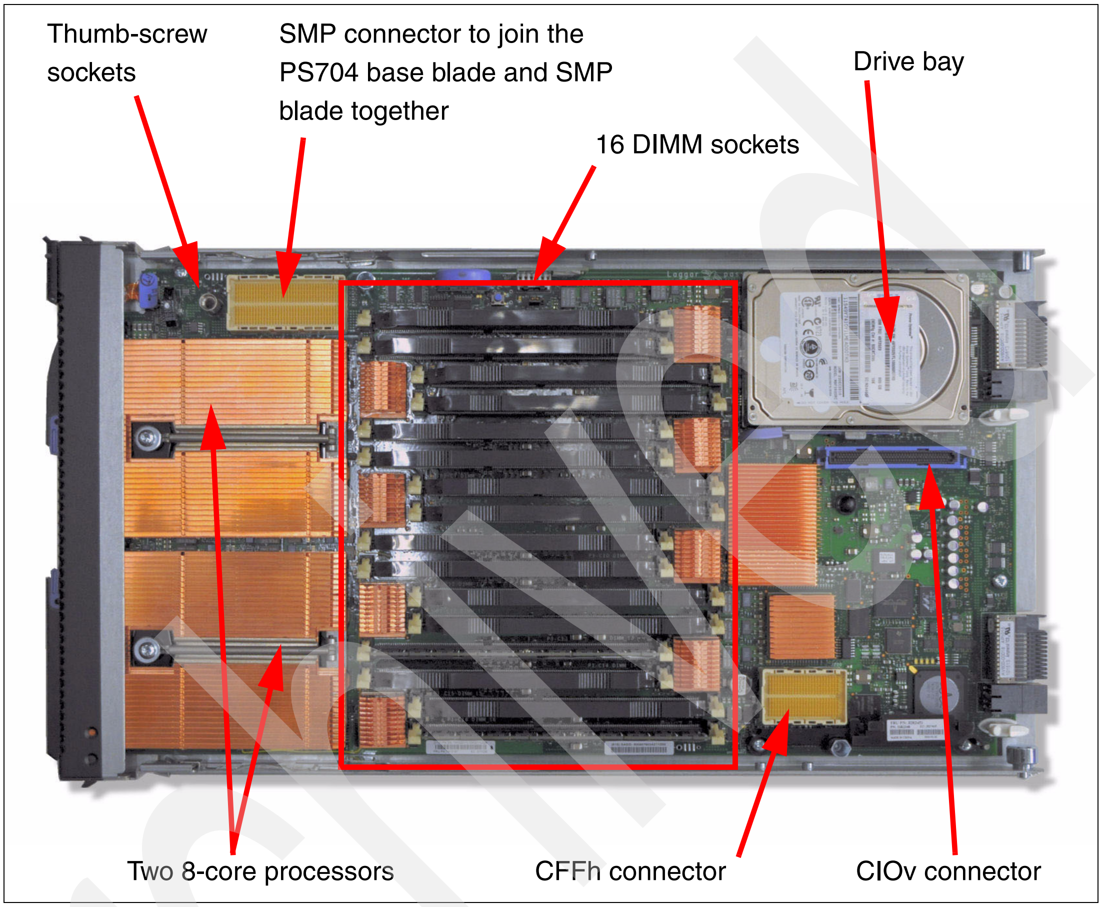


Figure 1-9 Top view of PS704 blade server base unit

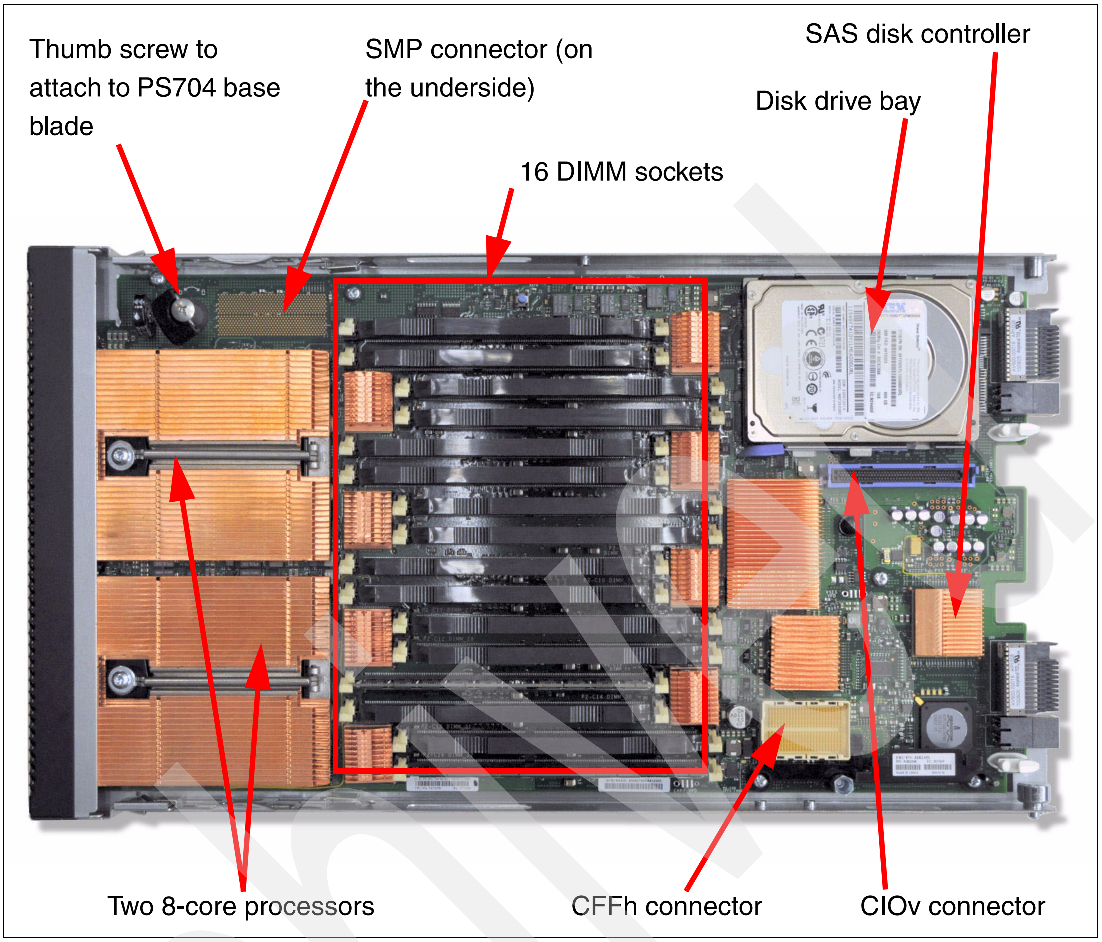


Figure 1-10 Top view of PS704 blade server SMP unit

The features of the server are as follows:

- Machine type and model number

  7891-74X
- Form factor

  Double-wide (60 mm) blade
- Processors:
  - Four eight-core 64-bit POWER7 processors operating at a 2.4 GHz clock speed for a total of 32 cores in the blade server
  - Based on CMOS 12S 45 nm SOI (silicon-on-insulator) technology
  - Power consumption is 110W per socket
- Memory
  - 32 DIMM slots
  - Minimum 32 GB, maximum capacity 512 GB (using 16 GB DIMMs)
  - Industry standard VLP DDR3 DIMMs
  - Optional Active Memory Expansion
- Disk
  - 3 Gb SAS disk storage controller which is located in the SMP unit

  - Two disk drive bays supporting up to two 2.5-inch SAS HDD (hard disk drive) or up to four 1.8-inch SAS SSD (solid state drive)
  - Hardware mirroring:
    - One HDD: RAID 0
    - One SSD: RAID 0
    - Two HDDs: RAID 0 or RAID 10
    - One HDD and one SSD: RAID 0 on each disk; combining HDD and SSD in one RAID configuration is not allowed.
    - Two SSDs: RAID 0 or RAID 10
    - Three SSDs: RAID 0, RAID 5, or RAID 10 (RAID 10 with only two disks)
    - Four SSDs: RAID 0, RAID 5, RAID 6, or RAID 10
- On-board integrated features:
  - Service processor (one on each blade[1])
  - Four 1 Gb Ethernet ports
  - One SAS Controller
  - USB Controller which routes to the USB 2.0 port on the media tray
  - 1 Serial over LAN (SOL) Console through FSP
- Expansion Card I/O Options:
  - Two CIOv expansion card slots (PCIe)
  - Two CFFh expansion card slots (PCIe)

1.6.3 Minimum features for the POWER7 processor-based blade servers

At the minimum a PS703 requires a BladeCenter chassis, two eight-core 2.4 GHz processors, minimum memory of 16 GB, zero or one disks, and a Language Group Specify (mandatory to order voltage nomenclature/language).

At the minimum a PS704 requires a BladeCenter chassis, four eight-core 2.4 GHz processors, minimum memory of 32GB, zero or one disks, and a Language Group Specify (mandatory to order voltage nomenclature/language).

Each system has a minimum feature set to be valid. The minimum system configuration for PS703 and PS704 blade servers is shown in Table 1-6 on page 19.

[1] The service processor (or flexible service processor) on the expansion unit provides control but does not offer redundancy with the SP on the base unit.

Table 1-6 Minimum features for PS703 and PS704 blade server

Category	**Minimum features required**
BladeCenter chassis	Supported BladeCenter chassis Refer to 1.3.1, “Supported BladeCenter chassis” on page 5
Processor	▸ Two 8-core 2.4 GHz Processors in a PS703 Blade (7891-73X) ▸ Four 8-core 2.4 GHz Processors in a PS704 Blade(7891-74X)
Memory	DDR3 Memory DIMM For PS703: ▸ 16GB - two 8 GB (2 x 4 GB DIMMs) DDR3 1066 MHz (#8196) or one 16 GB (2 x 8 GB DIMMs) DDR3 1066 MHz (#8199) For PS704: ▸ 32 GB - four 8 GB (2 x 4 GB DIMMs) DDR3 1066 MHz (#8196) or two 16 GB (2 x 8 GB DIMMs) DDR3 1066 MHz (#8199)
Storage	AIX/Linux/Virtual I/O Server/IBM i (Required VIOS partition): ▸ 300 GB SAS 2.5-inch HDD (#8274) or ▸ 600 GB SAS 2.5 inch HDD (#8276) or ▸ 177 GB SATA SSD (#8207) If Boot from SAN 8 GB Fibre Channel HBA is selected with FC #8240, #8242 or #8271 or Fibre Channel over Ethernet Adapter FC #8275 must be ordered. FC #8207 requires FC #4539 - Interposer for 1.8-inch Solid® State Drives
1x Language Group	Country specific (selected by the customer)
Operating system	1x primary operating system (one of the following) ▸ AIX (#2146) ▸ Linux (#2147) ▸ IBM i (#2145) plus IBM i 6.1.1 (#0566) ▸ IBM i (#2145) plus IBM i 7.1 (#0567)

1.6.4 Power supply features

The peak power consumption is 428 W for the PS703 and 848 W for the PS704 blade server; power is provided by the BladeCenter power supply modules. The maximum measured value is the worst case power consumption expected from a fully populated server under intensive workload. The maximum measured value also takes into account component tolerance and non-ideal operating conditions. Power consumption and heat load vary greatly by server configuration and use.

Use the IBM Systems Energy Estimator to obtain a heat output estimate based on a specific configuration. The Estimator is available from the following web page:

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

For information about power supply requirements for each of the BladeCenter chassis supported by POWER7 blade servers and the number of POWER7 blades supported, see 1.3.2, “Number of PS703 and PS704 blades in a chassis” on page 12.

1.6.5 Processor

The processors used in the PS703 and PS704 are 64-bit POWER7 processors operating at 2.4 GHz. They are optimized to achieve maximum performance for both the system and its virtual machines. Couple that performance with PowerVM and you are now enabled for massive workload consolidation to drive maximum system use, predictable performance, and cost efficiency.

POWER7 Intelligent Threads Technology enables workload optimization by selecting the most suitable threading mode (Single thread (per core) or Simultaneous Multi-thread 2 or 4 modes, also called 2-SMT and 4-SMT). The Intelligent Threads Technology can provide improved application performance. In addition, POWER7 processors can maximize cache access to cores, improving performance, using Intelligent Cache technology.

EnergyScale™ Technology offers Intelligent Energy management features, which can dramatically and dynamically conserve power and further improve energy efficiency. These Intelligent Energy features enable the POWER7 processor to operate at a higher frequency if environmental conditions permit, for increased performance per watt. Alternatively, if user settings permit, these features allow the processor to operate at a reduced frequency for significant energy savings.

The PS703 and PS704 come with a standard processor configuration. There are no optional processor configurations for the PS703 and PS704. The PS703 and PS704 processor configurations are as follows:

- The PS703 blade server is a single-wide blade that contains two eight-core, 64-bit POWER7 2.4 GHz processors with 256 KB per processor core L2 cache and 4 MB per processor core L3 cache. No processor options are available.
- The PS704 blade server is a double-wide blade that supports four eight-core, 64-bit POWER7 2.4 GHz processor with 256 KB per processor core L2 cache and 4 MB per processor core L3 cache. No processor options are available.

1.6.6 Memory features

The PS703 and PS704 blade servers uses industry standard VLP DDR3 memory DIMMs. Memory DIMMs must be installed in matched pairs with the same size and speed. For details about the memory subsystem and layout, see 2.4, “Memory subsystem” on page 47.

The PS703 and PS704 blade serves have 16 and 32 DIMM slots, respectively. Memory is available in 4 GB, 8 GB, or 16 GB DIMMs, all operating at a memory speed of 1066 MHz. The memory sizes can be mixed within a system.

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. For details, see 2.4, “Memory subsystem” on page 47. Table 1-7 shows the DIMM features.

Table 1-7 Memory DIMM options

Feature code	Total memory size	Package includes	Speed
8196	8 GB	Two 4 GB DIMMs	1066 MHz
8199	16 GB	Two 8 GB DIMMs	1066 MHz
EM34	32 GB	Two 16 GB DIMMs	1066 MHz

Notes:

- The DDR2 DIMMs used in JS23 and JS43 blade servers are not supported in the POWER7 blade servers.
- The DDR3 DIMMs used in PS700, PS701, and PS702 blade servers are not supported in the PS703 and PS704 blade servers.

The optional Active Memory Expansion is a POWER7 technology that allows the effective maximum memory capacity to be much larger than the true physical memory. Compression and decompression of memory content using processor cycles can allow memory expansion up to 100%. This can allow an AIX 6.1 (or later) partition to do significantly more work with the same physical amount of memory or a server to run more partitions and do more work with the same physical amount of memory. For more information, see 2.5, “Active Memory Expansion” on page 52.

1.6.7 I/O features

The PS703 has one CIOv PCIe expansion card slot and one CFFh PCIe high-speed expansion card slot. The PS704 blade server has two CIOv expansion card slots and two CFFh expansion card slots.

Table 1-8 shows the CIOv and CFFh expansion cards supported in the PS703 and PS704 servers.

Table 1-8 I/O expansion cards supported in the PS703 and PS704

Card Description	Feature Code
CIOv	
QLogic 8 Gb Fibre Channel Expansion Card (CIOv)	8242
QLogic 4 Gb Fibre Channel Expansion Card (CIOv)	8241
Emulex 8 Gb Fibre Channel Expansion Card (CIOv)	8240
3 Gb SAS Passthrough Expansion Card (CIOv)	8246
Broadcom 2-Port Gb Ethernet Expansion Card (CIOv)	8243
CFFh	
QLogic 1Gb Ethernet and 8 Gb Fibre Channel Expansion Card (CFFh)	8271
QLogic 1 Gb Ethernet and 4 Gb Fibre Channel Expansion Card (CFFh)	8252
QLogic 2-port 10 Gb Converged Network Adapter (CFFh)	8275
2-Port QDR 40 GB/s InfiniBand Expansion Card (CFFh)	8272
Broadcom 2/4-Port Ethernet Expansion Card (CFFh)	8291

QLogic 8 Gb Fibre Channel Expansion Card (CIOv)

The QLogic 8 Gb Fibre Channel Expansion Card (CIOv) for IBM BladeCenter, feature #8242, enables high-speed access for IBM blade servers to connect to a Fibre Channel storage area network (SAN). When compared to the previous-generation 4 Gb adapters, the new adapter doubles the throughput speeds for Fibre Channel traffic. As a result, you can manage increased amounts of data and possibly benefit from a reduced hardware cost.

The card has the following features:

- CIOv form factor
- QLogic 2532 8 Gb ASIC
- PCI Express 2.0 host interface
- Support for two full-duplex Fibre Channel ports at 8 Gbps maximum per channel
- Support for Fibre Channel Protocol Small Computer System Interface (FCP-SCSI) and Fibre Channel Internet Protocol (FC-IP)
- Support for Fibre Channel service (class 3)
- Support for switched fabric, point-to-point, and Fibre Channel Arbitrated Loop (FC-AL) connections
- Support for NPIV

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0692.html?Open

QLogic 4 Gb Fibre Channel Expansion Card (CIOv)

The QLogic 4 Gb Fibre Channel Expansion Card (CIOv) for BladeCenter, feature #8241, enables you to connect the BladeCenter servers with CIOv expansion slots to a Fibre Channel SAN. Pick any Fibre Channel storage solution from the IBM System Storage® DS3000, DS4000®, DS5000, and DS8000® series, and begin accessing data over a high-speed interconnect. This card is installed into the PCI Express CIOv slot of a supported blade server. It provides connections to Fibre Channel-compatible modules located in bays 3 and 4 of a supported BladeCenter chassis. A maximum of one QLogic 4 Gb Fibre Channel Expansion Card (CIOv) is supported per single-wide (30 mm) blade server.

The card has the following features:

- CIOv form factor
- PCI Express 2.0 host interface
- Support for two full-duplex Fibre Channel ports at 4 Gbps maximum per channel
- Support for Fibre Channel Protocol SCSI (FCP-SCSI) and Fibre Channel Internet Protocol (FC-IP)
- Support for Fibre Channel service (class 3)
- Support for switched fabric, point-to-point, and Fibre Channel Arbitrated Loop (FC-AL) connections

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0695.html?Open

Emulex 8 Gb Fibre Channel Expansion Card (CIOv)

The Emulex 8 Gb Fibre Channel Expansion Card (CIOv) for IBM BladeCenter, feature #8240, enables high-performance connection to a SAN. The innovative design of the IBM BladeCenter midplane enables this Fibre Channel adapter to operate without the need for an optical transceiver module. This saves significant hardware costs. Each adapter provides dual paths to the SAN switches to ensure full redundancy. The exclusive firmware-based architecture allows firmware and features to be upgraded without taking the server offline or rebooting, and without the need to upgrade the driver.

The card has the following features:

- Support of the 8 Gbps Fibre Channel standard
- Use of the Emulex “Saturn” 8 Gb Fibre Channel I/O Controller (IOC) chip
- Enablement of high-speed and dual-port connection to a Fibre Channel SAN
- Can be combined with a CFFh card on the same blade server

- Comprehensive virtualization capabilities with support for N_Port ID Virtualization (NPIV) and Virtual Fabric
- Simplified installation and configuration using common HBA drivers
- Efficient administration by using HBAnyware for HBAs anywhere in the SAN
- Common driver model that eases management and enables upgrades independent of HBA firmware
- Support of BladeCenter Open Fabric Manager
- Support for NPIV when installed in the PS703 and PS704 blade servers

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0703.html?Open

3 Gb SAS Passthrough Expansion Card (CIOv)

This card, feature #8246, is an expansion card that offers the ideal way to connect the supported BladeCenter servers to a wide variety of SAS storage devices. The SAS connectivity card can connect to the Disk Storage Modules in the BladeCenter S. The card routes the pair of SAS channels from the blade's onboard SAS controller to the SAS switches installed in the BladeCenter chassis.

Tip: This card is also known as the SAS Connectivity Card (CIOv) for IBM BladeCenter.

This card is installed into the CIOv slot of the supported blade server. It provides connections to SAS modules located in bays 3 and 4 of a supported BladeCenter chassis.

The card has the following features:

- CIOv form factor
- Provides external connections for the two SAS ports of the blade server's onboard SAS controller
- Support for two full-duplex SAS ports at 3 Gbps maximum per channel
- Support for SAS, SSP, and SMP protocols
- Connectivity to SAS storage devices

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0701.html?Open

Broadcom 2-Port Gb Ethernet Expansion Card (CIOv)

The Broadcom 2-port Gb Ethernet Expansion Card (CIOv) is an Ethernet expansion card with two 1 Gb Ethernet ports designed for BladeCenter servers with CIOv expansion slots.

The card has the following features:

- PCI Express host interface
- Broadcom BCM5709S communication module
- BladeCenter Open Fabric Manager (BOFM) support
- Connection to 1000BASE-X environments using BladeCenter Ethernet switches
- Full-duplex (FDX) capability, enabling simultaneous transmission and reception of data on the Ethernet local area network (LAN)
- TCPIP checksum offload
- TCP segmentation offload

For more detail see the IBM Redbooks publication *IBM BladeCenter Products and Technology*, SG24-7523, available at the following web page:

http://www.redbooks.ibm.com/abstracts/sg247523.html?Open

QLogic 1 Gb Ethernet and 8 Gb Fibre Channel Expansion Card (CFFh)

The QLogic 1Gb Ethernet and 8Gb Fibre Channel Expansion Card, feature #8271, is a CFFh high speed blade server expansion card with two 8Gb Fibre Channel ports and two 1 Gb Ethernet ports. It provides QLogic 2532 PCI-Express ASIC for 8 Gb 2-port Fibre Channel and Broadcom 5709S ASIC for 1 Gb 2-port Ethernet. This card is used in conjunction with the Multi-Switch Interconnect Module and is installed in the left position of the MSIM and a Fibre Channel capable I/O module is installed in the right position of the MSIM. Both switches do not need to be present at the same time because the Fibre Channel and Ethernet networks are separate and distinct. It can be combined with a CIOv I/O card on the same high-speed blade server.

The card has the following features:

- Broadcom 5709S ASIC with two 1Gb Ethernet ports
- PCI Express host interface
- BladeCenter Open Fabric Manager (BOFM) support
- TCPIP checksum offload
- TCP segmentation offload
- Full-duplex (FDX) capability
- QLogic 2532 ASIC with two 8Gb Fibre Channel ports
- Support for FCP-SCSI and FCP-IP
- Support for point-to-point fabric connection (F-port fabric login)
- Support for Fibre Channel service (classes 2 and 3)
- Support for NPIV when installed in PS703 and PS704 blade servers
- Support for remote startup (boot) operations
- Support for BladeCenter Open Fabric Manager
- Support for Fibre Device Management Interface (FDMI) standard (VESA standard)
- Fibre Channel 8 Gbps, 4 Gbps, or 2 Gbps auto-negotiation

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0690.html?Open

QLogic 1 Gb Ethernet and 4 Gb Fibre Channel Expansion Card (CFFh)

The QLogic Ethernet and 4 Gb Fibre Channel Expansion Card, feature #8252, is a CFFh high speed blade server expansion card with two 4 Gb Fibre Channel ports and two 1 Gb Ethernet ports. It provides QLogic 2432M PCI-Express x4 ASIC for 4 Gb 2-port Fibre Channel and Broadcom 5715S PCI-Express x4 ASIC for 1 Gb 2-port Ethernet. This card is used in conjunction with the Multi-Switch Interconnect Module and is installed in the left position of the MSIM and a Fibre Channel capable I/O module is installed in the right position of the MSIM. Both switches do not need to be present at the same time because the Fibre Channel and Ethernet networks are separate and distinct. It can be combined with a CIOv I/O card on the same high-speed blade server.

The card has the following features:

- Support for FCP-SCSI and FCP-IP
- Support for point-to-point fabric connection (F-port fabric login)
- Support for remote startup (boot) operations
- Support for BladeCenter Open Fabric Manager

For more detail see the IBM Redbooks publication *IBM BladeCenter Products and Technology*, SG24-7523, available at the following web page:

http://www.redbooks.ibm.com/abstracts/sg247523.html?Open

QLogic 2-port 10 Gb Converged Network Adapter (CFFh)

The QLogic 2-port 10 Gb Converged Network Adapter (CFFh) for IBM BladeCenter, feature #8275, offers robust Fibre Channel storage connectivity and 10 Gb networking over a single Converged Enhanced Ethernet (CEE) link. Because this adapter combines the functions of a network interface card and a host bus adapter on a single converged adapter, clients can realize potential benefits in cost, power, and cooling, and data center footprint by deploying less hardware.

The card has the following features:

- CFFh PCI Express 2.0 x8 adapter
- Communication module: QLogic ISP8112
- Support for up to two CEE HSSMs in a BladeCenter H or HT chassis
- Support for 10 Gb Converged Enhanced Ethernet (CEE)
- Support for Fibre Channel over Converged Enhanced Ethernet (FCoCEE)
- Full hardware offload for FCoCEE protocol processing
- Support for IPv4 and IPv6
- Support for SAN boot over CEE, PXE boot, and iSCSI boot
- Support for Wake on LAN

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0716.html?Open

2-Port QDR 40 Gbps InfiniBand Expansion Card (CFFh)

The 2-Port 40 Gbps InfiniBand Expansion Card (CFFh) for IBM BladeCenter is a dual port InfiniBand Host Channel Adapter (HCA) based on proven Mellanox ConnectX IB technology. This HCA, when combined with the QDR switch, delivers end-to-end 40 Gb bandwidth per port. This solution is ideal for low latency, high bandwidth, performance-driven and storage clustering application in a High Performance Compute environment.

The card has the following features:

- 1μs MPI ping latency
- Dual 4X InfiniBand ports at speeds of 10 Gbps, 20 Gbps, or 40 Gbps per port
- CPU offload of transport operations
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization
- Multi-protocol support
- TCP/UDP/IP stateless offload
- 2-port card allows use of two 40 Gb High-Speed Switch Modules (HSSM) in a chassis

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0700.html?Open

Broadcom 2/4-Port Ethernet Expansion Card (CFFh)

The 2/4-Port Ethernet Expansion Card (CFFh) for IBM BladeCenter allows the addition of up to four (in IBM BladeCenter H chassis) or two (in BladeCenter S) extra 1 Gb ports, thereby allowing the use of 6 or 4 ports per blade, respectively.

The card has the following features:

- Based on the Broadcom 5709S module
- PCI Express x4 host interface for high-speed connection
- Connectivity to either standard or high-speed I/O modules bays (depends on chassis)
- Multiple connections from the blade server to the external network

- Ability to function as a 2-port Ethernet NIC in BladeCenter S chassis or a 4-Port Ethernet NIC in a BladeCenter H chassis
- Supports BladeCenter Open Fabric Manager (BOFM)
- Network install and boot support with adapter firmware update

For more information, see the IBM Redbooks at-a-glance guide at the following web page:

http://www.redbooks.ibm.com/abstracts/tips0698.html?Open

1.6.8 Disk features

The PS703 blade servers have one disk bay. The bay supports either of the following:

- One 2.5-inch SAS HDD
- One or two 1.8-inch SATA solid state drives (SSDs)

If you elect to use SSDs, then the Interposer for 1.8-inch Solid State Drives, feature code 4539 must also be installed.

The PS704 blade servers have two disk bays (one on the base card and one in the expansion unit of the blade):

- On the base unit, it can have one 2.5-inch SAS HDD.
- On the base unit, it can have up to two 1.8-inch SATA SSDs.
- On the expansion unit, it can have one 2.5-inch SAS HDD.
- On the expansion unit, it can have up to two 1.8-inch SATA SSDs.

Table 1-6 lists the supported disk features on the PS703 and PS704 blade servers.

Table 1-9 Supported disk drives and options

Feature code	Description
2.5-inch SAS drives	
8274	300 GB 10K SFF SAS HDD
8276	600 GB 10K SFF SAS HDD
1.8-inch solid state drive (SSD) and interposer	
8207	177 GB SATA SSD (requires feature 4539)
4539	Interposer for 1.8-inch Solid State Drives

1.6.9 Standard onboard features

In this section, we describe the standard on-board features.

Service processor

The service processor (or flexible service processor, FSP) is the main integral part of the blade server. It monitors and manages system hardware, resources, and devices. It does the system initialization, configuration, and thermal/power management. It takes corrective action if required.

The PS703 has only one service processor. The PS704 blade server has two FSPs (one on each blade). However, the second service processor is only in IO mode and is not redundant to the one on the base blade.

For more details about service processors, see 2.8, “Service processor” on page 65.

Ethernet ports

The PS703 has a 2-port onboard integrated Ethernet adapter for a total of two Ethernet ports. The PS704 has two 2-port onboard integrated Ethernet adapters with one in the base blade and the second one in the SMP blade for a total of four Ethernet ports.

> **Note:** The PS703 and PS704 do not have the Host Ethernet Adapters (HEAs) and Integrated Virtual Ethernet (IVE) ports that previous Power blade servers have included. A virtual Ethernet can be provided from the Virtual I/O Server virtual network environment.

For more details about Ethernet ports, see 2.7.5, “Embedded Ethernet Controller” on page 63.

SAS Controller

The PS703 blade server has one integrated SAS controller. The PS704 has one integrated SAS controller located on the SMP blade.

The integrated SAS controller is used to drive the local SAS storage. This SAS controller can also support SATA SSD with the addition of the SSD Interposer for 1.8-inch solid-state drives as shown in Figure 1-11.

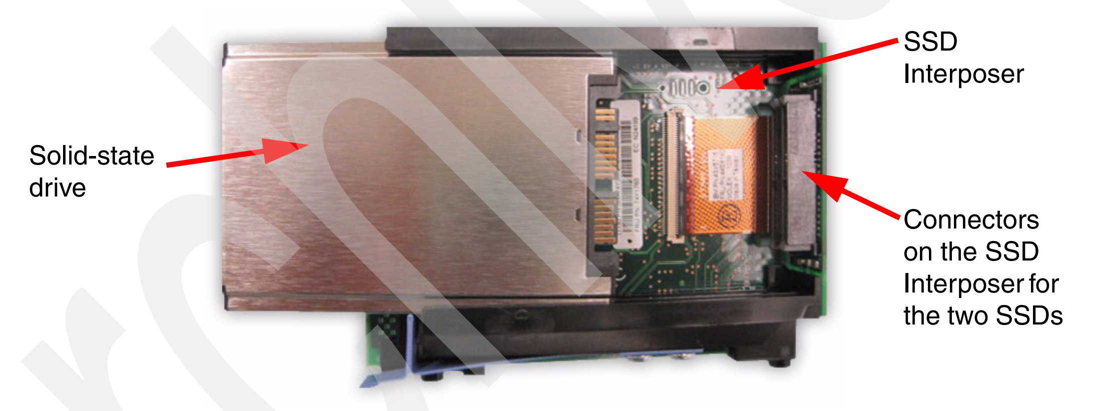


Figure 1-11 Interposer for 1.8-inch Solid State Drives

The integrated SAS controller supports hardware RAID 0, RAID 5, RAID 6, or RAID 10 depending on the number of drives installed.

The 3 Gb SAS Passthrough Expansion Card can be used to connect to the BladeCenter SAS Connectivity Module, which can be connected to the external storage. This SAS pass-through expansion card can also be used to connect to BladeCenter S internal drive SAS drives. See “3 Gb SAS Passthrough Expansion Card (CIOv)” on page 23 for more information. See also “SAS adapter” on page 61 and 2.9, “Internal storage” on page 68.

USB controller

The USB controller connects the USB bus to the midplane, which is then routed to the media tray in the BladeCenter chassis to connect to USB devices (such as an optical drive or diskette drive).

For more information, see 2.7.6, “Embedded USB controller” on page 64.

Serial over LAN (SOL)

The integrated SOL function routes the console data stream over standard dual 1 Gb Ethernet ports to the Advance Management Module. The PS703 and PS704 do not have on-board video chips and do not support KVM connections. Console access is only by SOL connection. Each blade can have a single SOL session; however, there can be multiple telnet or ssh sessions to the BladeCenter AMM, each acting as a SOL connection to a different blade.

For more information, see 2.8.1, “Server console access by SOL” on page 65.

1.7 Supported BladeCenter I/O modules

With IBM BladeCenter, the switches and other I/O modules are installed in the chassis rather than as discrete devices installed in the rack.

The BladeCenter chassis supports a wide variety and range of I/O switch modules. These switch modules are matched to the type, slot location, and form factor of the expansion cards installed in a blade server. For more information, see 1.6.7, “I/O features” on page 21 and 2.7, “Internal I/O subsystem” on page 55.

The I/O switch modules described in the following sections are matched with the on-board Broadcom 2-port BCM5709S network controller ports along with the supported expansion card ports in the PS703 and PS704 blades.

For the latest and most current information about blade, expansion card, switch module, and chassis compatibility and interoperability see the *IBM BladeCenter Interoperability Guide* at the following web page:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073016

1.7.1 Ethernet switch and intelligent pass-through modules

Various types of Ethernet switch and pass-through modules from several manufacturers are available for BladeCenter, and they support different network layers and services. These I/O modules provide external and chassis blade-to-blade connectivity.

The Broadcom 2-port BCM5709S network controller, along with the supported expansion cards in the PS703 and PS704 blades, provides Ethernet connectivity. For more information, see 2.7.5, “Embedded Ethernet Controller” on page 63. There are two physical ports on the PS703 and four physical ports on the PS704. The data traffic from these on-blade 1 Gb Ethernet adapter ports is directed to I/O switch bays 1 and 2 respectively on all BladeCenter chassis except BladeCenter S. On the BladeCenter S the connections for all blade Ethernet ports are wired to I/O switch bay 1.

To provide external network connectivity and a SOL system console through the BladeCenter Advanced Management Module, at least one Ethernet I/O module is required in switch bay 1. For more information, see 2.8.1, “Server console access by SOL” on page 65.

In addition to the onboard Ethernet ports, the QLogic Ethernet and 4 Gb Fibre Channel Expansion Card (CFFh) adapter can provide two additional 1 Gb Ethernet ports per card.

A list of available Ethernet I/O modules that support the on-blade Ethernet ports and expansion card are shown in Table 1-10 on page 29. Not all switches are supported in every configuration of BladeCenter. Complete compatibility matrixes are available on the following web pages:

- ServerProven®:

  http://www.ibm.com/servers/eserver/serverproven/compat/us/eserver.html
- *BladeCenter Interoperability Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073016

Table 1-10 Ethernet switch modules

Part number	Feature code[a]	Option description	Number x type of external ports	Network layers
43W4395	3174	Cisco Catalyst Switch Module 3012	4 x Gigabit Ethernet	Layer 2/3
41Y8523	3173	Cisco Catalyst Switch Module 3110G	4 x Gigabit Ethernet, 2 x StackWise Plus	Layer 2/3
41Y8522	3171	Cisco Catalyst Switch Module 3110X	1 x 10 Gb Ethernet, 2 x StackWise Plus	Layer 2/3
32R1888	Not avail	Cisco Systems Intelligent Gb Fiber Ethernet Switch	4 x Gigabit Ethernet	Layer 2
32R1892	Not avail	Cisco Systems Intelligent Gb Ethernet Switch	4 x Gigabit Ethernet	Layer 2
39Y9324	3220	IBM Server Connectivity Module	6 x Gigabit Ethernet	Layer 2
32R1860	3212	BNT® L2/3 Copper Gigabit Ethernet Switch Module	6 x Gigabit Ethernet	Layer 2/3
32R1861	3213	BNT L2/3 Fibre Gigabit Ethernet Switch Module	6 x Gigabit Ethernet	Layer 2/3
32R1859	3211	BNT Layer 2-7 Gigabit Ethernet Switch Module	4 x Gigabit Ethernet	Layer 2/7
44W4404	1590	BNT 1/10 Gb Uplink Ethernet Switch Module	3 x 10 Gb Ethernet, 6 x Gigabit Ethernet	Layer 2/3

[a] These feature codes are for the Power Systems ordering system (eConfig)

1.7.2 SAS I/O modules

SAS I/O modules provide affordable storage connectivity for BladeCenter chassis using SAS technology to create simple fabric for external shared or non-shared storage attachments.

The SAS RAID Controller Module can perform RAID controller functions inside the BladeCenter S chassis for HDDs installed into Disk Storage Module (DSM). The SAS RAID Controller Module and DSMs in a BladeCenter S provides RAID 0, 5, 6, and 10 support.

The SAS Controller Module (non-RAID) supports the external storage EXP3000 but binds that enclosure to a specific blade.

The DSM, part number 43W3581 feature 4545, must be installed in the BladeCenter S chassis to support external SAS storage devices outside the chassis using the SAS Connectivity Card. No HDDs need to be installed in the DSM to support the external storage.

In the PS703 and PS704 blades, the 3 Gb SAS Passthrough Expansion Card (CIOv) is required for external SAS connectivity. The SAS expansion card requires SAS I/O modules in switch bays 3 and 4 of all supported BladeCenters.

Table 1-11 on page 30 lists the SAS I/O modules and support matrix.

Table 1-11 SAS I/O modules supported by the SAS pass through card

Part number	Feature code[a]	Description	3 Gb SAS pass-thru card	BC-E	BC-H	BC-HT	BC-S	MSIM	MSIM-HT
39Y9195	3267	SAS Connectivity Module	Yes	Yes	Yes	Yes	Yes	No	No
43W3584	3734	SAS RAID Controller Module	Yes	No	No	No	Yes	No	No

[a] These feature codes are for the Power Systems ordering system (eConfig)

1.7.3 Fibre Channel switch and pass-through modules

Fibre Channel I/O modules are available from several manufacturers. These I/O modules can provide full SAN fabric support up to 8 Gb.

The following 4 Gb and 8 Gb Fibre Channel cards are CIOv form factor and require a Fibre Channel switch or Intelligent Pass-Through module in switch bays 3 and 4 of all supported BladeCenters. The CIOv expansion cards are as follows:

- Emulex 8 Gb Fibre Channel Expansion Card (CIOv)
- QLogic 4 Gb Fibre Channel Expansion Card (CIOv)
- QLogic 8 Gb Fibre Channel Expansion Card (CIOv)

Additional 4 Gb and 8 Gb Fibre Channel ports are also available in the CFFh form factor expansion cards. These cards require the use of the MSIM in a BladeCenter H or the MSIM-HT in a BladeCenter HT, plus Fibre Channel I/O modules. The CFFh Fibre Channel cards are as follows:

- QLogic Ethernet and 4 Gb Fibre Channel Expansion Card (CFFh)
- QLogic 8 Gb Fibre Channel Expansion Card (CFFh)

A list of available Fibre Channel I/O modules that support the CIOv and CFFh expansion cards is shown in Table 1-12 on page 31. Not all modules are supported in every configuration of BladeCenter. Complete compatibility matrixes are available on the following web pages:

- ServerProven:

  `http://www.ibm.com/servers/eserver/serverproven/compat/us/eserver.html`

- *BladeCenter Interoperability Guide*

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073016`

Table 1-12 Fibre Channel I/O modules

Part number	Feature code[a]	Description	Number of external ports	Port interface bandwidth
32R1812	3206	Brocade 20-port SAN Switch Module	6	4 Gbps
32R1813	3207	Brocade 10-port SAN Switch Module[b]	6	4 Gbps
42C1828	5764	Brocade Enterprise 20-port 8 Gb SAN Switch Module	6	8 Gbps
44X1920	5869	Brocade 20-port 8 Gb SAN Switch Module	6	8 Gbps
44X1921	5045	Brocade 10-port 8 Gb SAN Switch Module	6	8 Gbps
39Y9280	3242	Cisco Systems 20-port 4 Gb FC Switch Module	6	4 Gbps
39Y9284	3241	Cisco Systems 10-port 4 Gb FC Switch Module[b]	6	4 Gbps
43W6725	3244	QLogic 20-port 4 Gb SAN Switch Module	6	4 Gbps
43W6724	3243	QLogic 10-port 4 Gb SAN Switch Module[b]	6	4 Gbps
43W6723	3245	QLogic 4 Gb Intelligent Pass-thru Module[c]	6	4 Gbps
44X1905	3284	QLogic 20-Port 8 Gb SAN Switch Module	6	8 Gbps
44X1907	5449	QLogic 8 Gb Intelligent Pass-thru Module[c]	6	8 Gbps
44W4483	5452	Intelligent Copper Pass-thru Module	14	1 Gbps

a. These feature codes are for the Power Systems ordering system (eConfig)

b. Only 10 ports are activated on these switches. An optional upgrade to 20 ports (14 internal + 6 external) is available.

c. Can be upgraded to full fabric switch

1.7.4 Converged networking I/O modules

There are two basic solutions to implement Fibre Channel over Ethernet (FCoE) over a converged network with a BladeCenter.

- The first solution uses a top-of-rack FCoE-capable switch in conjunction with converged-network-capable 10 Gb Ethernet I/O modules in the BladeCenter. The FCoE-capable top-of-rack switch provides connectivity to the SAN.
- The second BladeCenter H solution uses a combination of converged-network-capable 10 Gb Ethernet switch modules and fabric extension modules to provide SAN connectivity, all contained within the BladeCenter H I/O bays.

Implementing either solution with the PS703 and PS704 blades requires the QLogic 2-port 10 Gb Converged Network Adapter (CFFh). The QLogic Converged Network Adapter (CNA) provides 10 Gb Ethernet and 8 Gb Fibre Channel connectivity over a single CEE link. This card is a CFFh form factor with connections to BladeCenter H and HT I/O module bays 7 and 9.

Table 1-13 on page 32 shows the currently available I/O modules that are available to provide an FCoE solution.

Table 1-13 Converged network modules supported by the QLogic CNA

Part Number	Feature Code[a]	Description	Number of external ports
46C7191	3248	BNT Virtual Fabric 10 Gb Switch Module for IBM BladeCenter[b,c]	10 x 10 Gb SFP+
46M6181	5412	10 Gb Ethernet Pass-Thru Module for BladeCenter[b]	14 x 10 Gb SFP+
46M6172	3268	QLogic Virtual Fabric Extension Module for IBM BladeCenter[d,e]	6 x 8 Gb FC SFP
46M6071	2241	Cisco Nexus 4001I Switch Module for IBM BladeCenter[b]	6 x 10 Gb SFP+
69Y1909	Not available	Brocade Converged 10 GbE Switch Module for IBM BladeCenter	8 X 10 GB Ethernet 8 x 8 Gb FC

[a] These feature codes are for the Power Systems ordering system (eConfig).
[b] Used for top-of-rack solution.
[c] Use with Fabric Extension Module for self contain BladeCenter solution.
[d] Also requires BNT Virtual Fabric 10 Gb Switch Module.
[e] BladeCenter H only.

For the latest interoperability information see the BladeCenter Interoperability Guide, available from:

http://ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5073016

1.7.5 InfiniBand switch module

The Voltaire 40 Gb InfiniBand Switch Module for BladeCenter provides InfiniBand QDR connectivity between the blade server and external InfiniBand fabrics in non-blocking designs, all on a single device. Voltaire's high speed module also accommodates performance-optimized fabric designs using a single BladeCenter chassis or stacking multiple BladeCenter chassis without requiring an external InfiniBand switch.

The InfiniBand switch module offers 14 internal ports, one to each server, and 16 ports out of the chassis per switch.

The module's HyperScale architecture also provides a unique interswitch link or mesh capability to form highly scalable, cost-effective, and low latency fabrics. Because this switch has 16 uplink ports, they can create a meshed architecture and still have unblocked access to data using the 14 uplink ports. This solution can scale from 14 to 126 nodes and offers latency of less than 200 nanoseconds, allowing applications to operate at maximum efficiency.

The PS703 and PS704 blades connect to the Voltaire switch through the 2-port 40 Gb InfiniBand Expansion Card. The card is only supported in a BladeCenter H and the two ports are connected to high speed I/O switch bays 7/8 and 9/10.

Details about the Voltaire 40 Gb InfiniBand Switch Module for the BladeCenter H are shown in Table 1-14.

Table 1-14 InfiniBand switch module for IBM BladeCenter

Part number	Feature code	Description	Number of external ports	Type of external ports
46M6005	3204	Voltaire 40 Gb InfiniBand Switch Module	16	4X QDR (40 Gbps)

1.7.6 Multi-switch Interconnect Module

The MSIM is a switch module container that fits in the high speed switch bays (bays 7 and 8 or bays 9 and 10) of the BladeCenter H chassis. Up to two MSIMs can be installed in the BladeCenter H. The MSIM supports most standard switch modules. I/O module to MSIM compatibility matrixes can be reviewed at the following web pages:

- ServerProven:

  http://www.ibm.com/servers/eserver/serverproven/compat/us/eserver.html

- *BladeCenter Interoperability Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073016

With PS703 and PS704 blades, the following expansion cards require an MSIM in a BladeCenter H chassis:

- QLogic Ethernet and 4 Gb Fibre Channel Expansion Card (CFFh)
- QLogic 8 Gb Fibre Channel Expansion Card (CFFh)

The MSIM is shown in Figure 1-12.

> **Note:** The MSIM comes standard without any I/O modules installed. They must be ordered separately. In addition, the use of MSIM modules requires that all four power modules be installed in the BladeCenter H chassis.

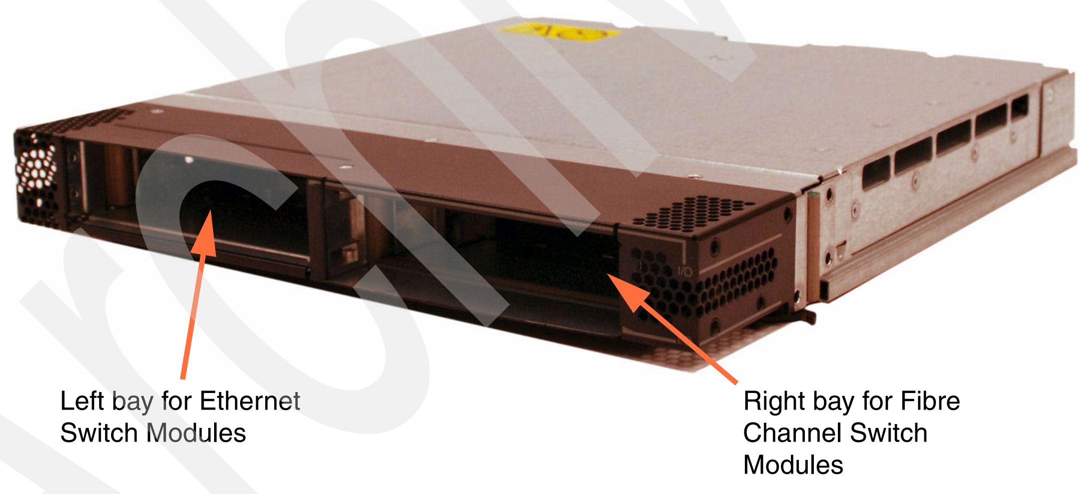

Left bay for Ethernet Switch Modules
Right bay for Fibre Channel Switch Modules

Figure 1-12 Multi-switch Interconnect Module

Table 1-15 shows MSIM ordering information.

Table 1-15 MSIM ordering information

Description	Part number	Feature code[a]
MSIM for IBM BladeCenter	39Y9314	3239

[a] These feature codes are for the Power Systems ordering system (eConfig).

1.7.7 Multi-switch Interconnect Module for BladeCenter HT

The Multi-switch Interconnect Module for BladeCenter HT (MSIM-HT) is a switch module container that fits in the high-speed switch bays (bays 7 and 8 or bays 9 and 10) of the BladeCenter HT chassis. Up to two MSIMs can be installed in the BladeCenter HT. The MSIM-HT accepts two supported standard switch modules as shown in Figure 1-13.

The MSIM-HT has a reduced number of supported standard I/O modules compared to the MSIM.

I/O module to MSIM-HT compatibility matrixes can be viewed at the following web pages:

- ServerProven:

  http://www.ibm.com/servers/eserver/serverproven/compat/us/eserver.html

- *BladeCenter Interoperability Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073016

With PS703 and PS704 blades the QLogic Ethernet and 4 Gb Fibre Channel Expansion Card (CFFh) requires an MSIM-HT in a BladeCenter HT chassis.

Note: The MSIM-HT comes standard without any I/O modules installed. They must be ordered separately. In addition, the use of MSIM-HT modules requires that all four power modules be installed in the BladeCenter HT chassis.

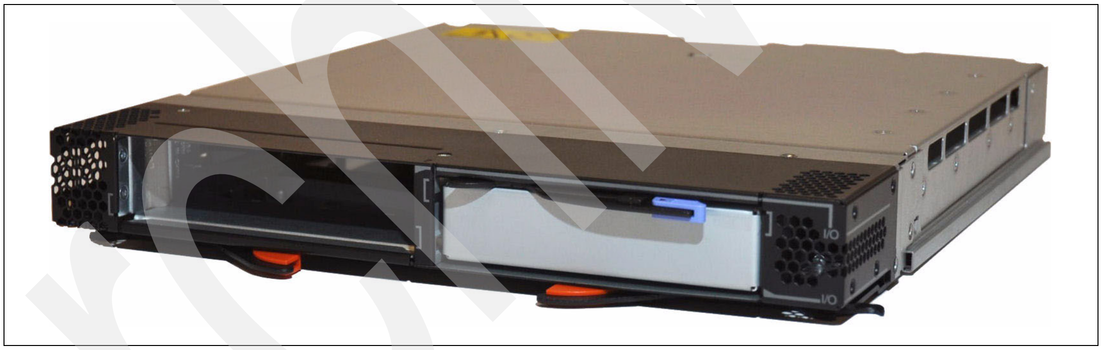


Figure 1-13 Multi-switch Interconnect Module for BladeCenter HT

Table 1-16 shows MSIM-HT ordering information.

Table 1-16 MSIM-HT ordering information

Description	Part number	Feature code[a]
Multi-switch Interconnect Module for BladeCenter HT	44R5913	5491

[a] These feature codes are for the Power Systems ordering system (eConfig).

1.8 Building to order

You can perform a build to order configuration using the IBM Configurator for e-business (e-config). The configurator allows you to select a pre-configured Express model or to build a system to order. Use this tool to specify each configuration feature that you want on the system, building on top of the base-required features.

1.9 Model upgrades

The PS703 and PS704 are new serial-number blade servers. There are no upgrades from POWER5™, POWER6, POWER7, and the POWER7 PS700, PS701, and PS702 blade servers to the POWER7 PS703 and PS704 blade servers, which retain the serial number.

Unlike the upgrade which exists from the PS701 to the PS702, there are no upgrades from the PS703 to the PS704.

2

Architecture and technical overview

This chapter discusses the overall system architecture of the POWER7 processor-based blade servers and provides details about each major subsystem and technology.

The topics covered are:

Note: The bandwidths that are provided throughout the chapter are theoretical maximums used for reference.

2.1 Architecture

The overall system architecture is shown in Figure 2-1, with the major components described in the following sections. Figure 2-1 shows the PS703 layout.

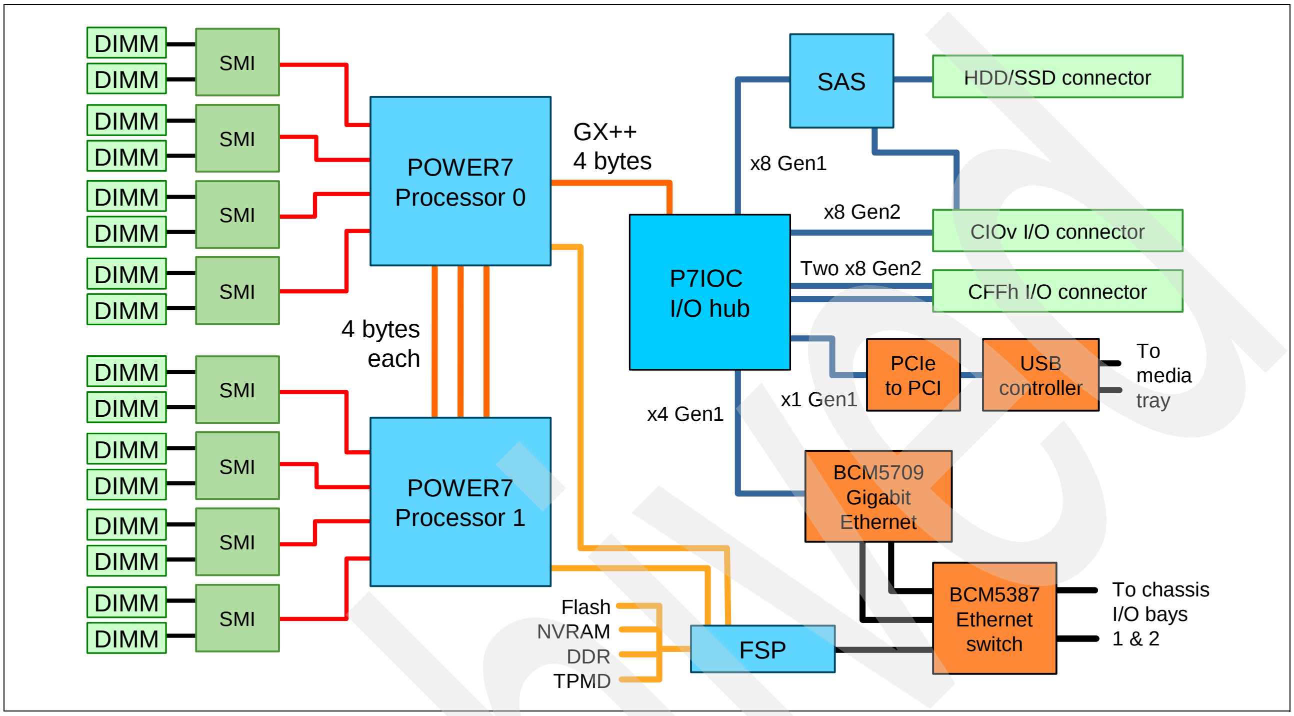


Figure 2-1 PS703 block diagram

The PS704 double-wide blade base planar holds the same components as the PS703 single-wide blade with the exclusion of the SAS controller, which is located on the SMP planar. The PS704 double-wide blade SMP planar also has the same components with the exclusion of the USB controller. This means that the components are doubled for the PS704 double-wide blade as compared to the PS703 single-wide blade except for the SAS controller and USB controller. See 2.7, “Internal I/O subsystem” on page 55 for more details.

Figure 2-2 on page 39 shows the PS704 layout.

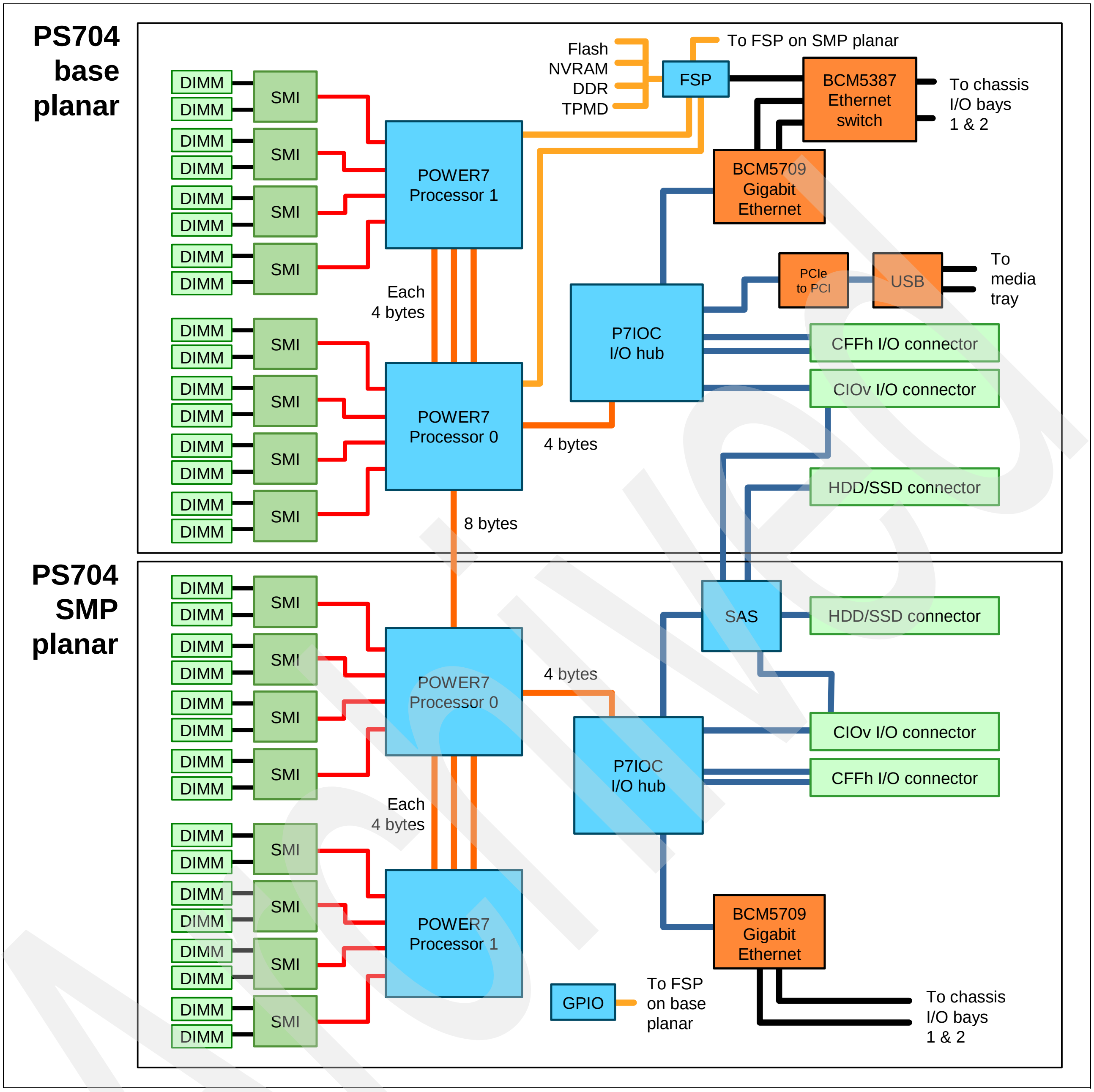


Figure 2-2 PS704 block diagram

2.2 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 a wide range of related technologies to deliver leading throughput, efficiency, scalability, and reliability, availability, and serviceability (RAS).

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

IBM has used innovative methods 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) the following elements:

- 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

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

- 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 2-3 shows the POWER7 processor die layout with the major areas identified: eight POWER7 processor cores, L2 cache, L3 cache and chip power bus Interconnect, simultaneous multiprocessing (SMP) links, GX++ interface, and two memory controllers.

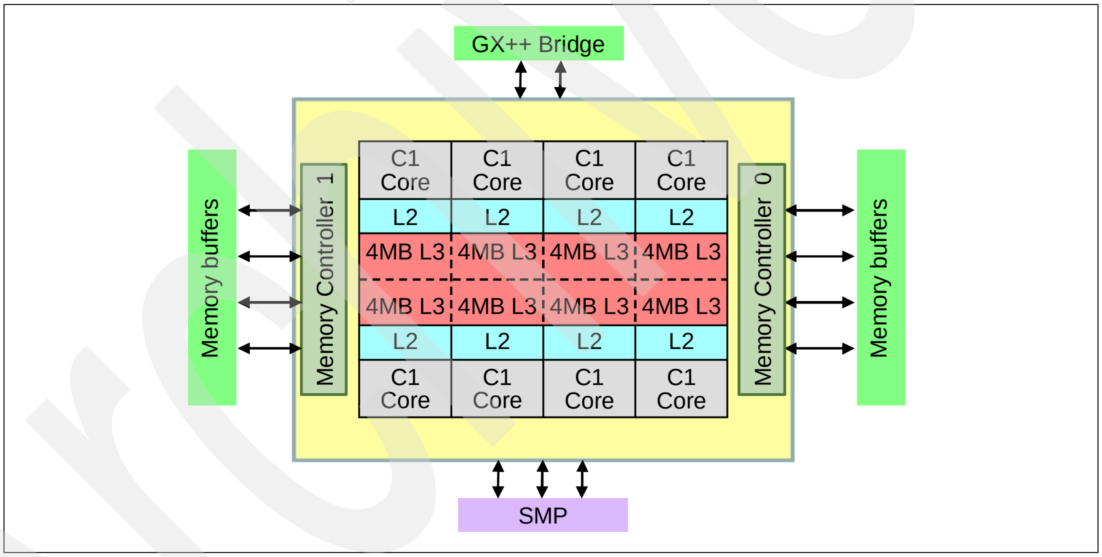


Figure 2-3 POWER7 processor architecture

2.2.1 POWER7 processor overview

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

The POWER7 processor chip is 567 mm^2 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 scale effectively, the POWER7 processor uses a combination of local and global SMP links with high coherency bandwidth and makes use of the IBM dual-scope broadcast coherence protocol.

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

Table 2-1 Summary of POWER7 processor technology

Technology	POWER7 processor
Die size	567 mm^2
Fabrication technology	► 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 2.2.6, "On-chip L3 intelligent cache" on page 44)
Processor cores	8
Max execution threads core/chip	4/32
L2 cache per core/per chip	256 KB / 2 MB
On-chip L3 cache per core/per chip	4 MB / 32 MB
DDR3 memory controllers	2
SMP design-point	Up to 32 sockets with IBM POWER7 processors
Compatibility	With prior generation of POWER processor

2.2.2 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 as follows:

- Instruction cache: 32 KB
- Data cache: 32 KB
- L2 cache: 256 KB, implemented in fast SRAM
- L3 cache: 4MB eDRAM

2.2.3 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. Figure 2-4 shows the evolution of simultaneous multithreading.

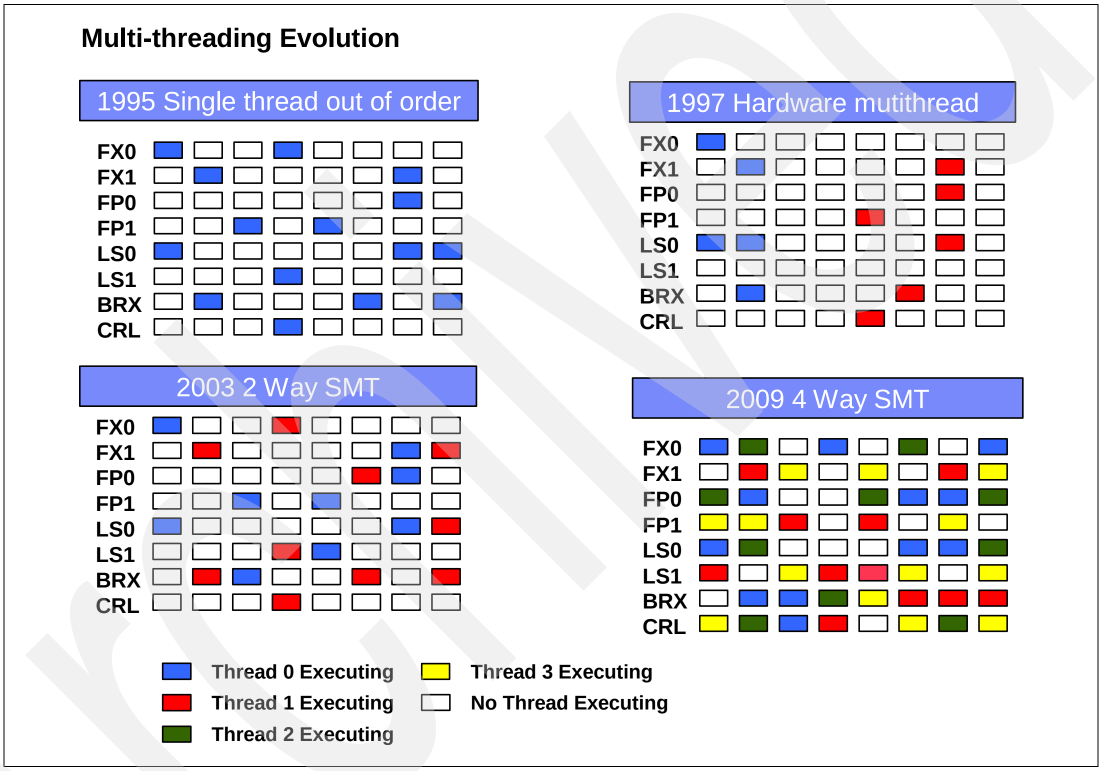


Figure 2-4 Evolution of simultaneous multithreading

The various SMT modes offered by the POWER7 processor allow flexibility, enabling users to select the threading technology that meets a combination of objectives (such as performance, throughput, energy use, and workload enablement).

Intelligent threads

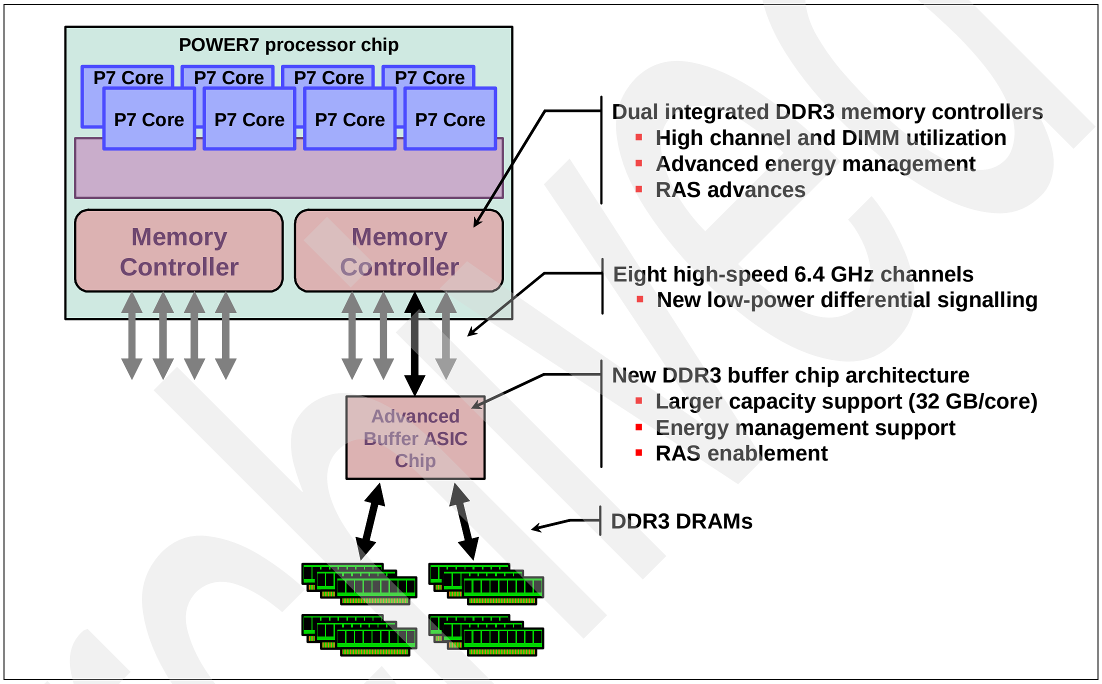
The POWER7 processor features *intelligent threads*, which can vary based on the workload demand. The system either automatically selects (or the system administrator can manually select) whether a workload benefits from dedicating as much capability as possible to a single thread of work, or if the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

2.2.4 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 certain POWER7 processor-based systems (including the PS700, PS701, PS702, PS703 and PS704) only one memory controller is active.

Figure 2-5 gives a simple overview of the POWER7 processor memory access structure.



POWER7 processor chip
P7 Core
P7 Core
P7 Core
P7 Core
P7 Core
P7 Core
P7 Core
P7 Core
Memory Controller
Memory Controller
Advanced Buffer ASIC Chip
Dual integrated DDR3 memory controllers
High channel and DIMM utilization
Advanced energy management
RAS advances
Eight high-speed 6.4 GHz channels
New low-power differential signalling
New DDR3 buffer chip architecture
Larger capacity support (32 GB/core)
Energy management support
RAS enablement
DDR3 DRAMs

Figure 2-5 Overview of POWER7 memory access structure

2.2.5 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 fast processors that handle high transaction rates at high speeds. Web workloads typically benefit more from processors with many threads that allow the breakdown of Web requests into many parts and handle them in parallel. POWER7 processors have the unique ability to provide leadership performance in either case.

POWER7 processor cores

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. For the PS703 and PS704 blades, only the full 8-core version is used.

The L3 cache associated with the implementation is dependant on the number of active cores. For the 8-core version, this means that 8 x 4 = 32 MB of L3 cache is available.

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 2-6 shows the physical packaging options that are supported with POWER7 processors.

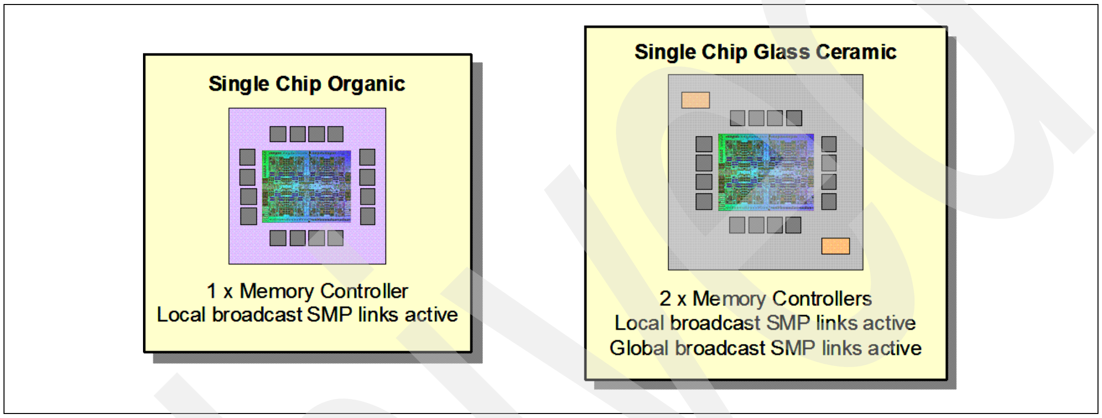

Single Chip Organic
1 x Memory Controller
Local broadcast SMP links active
Single Chip Glass Ceramic
2 x Memory Controllers
Local broadcast SMP links active
Global broadcast SMP links active

Figure 2-6 Outline of the POWER7 processor physical packaging

2.2.6 On-chip L3 intelligent cache

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

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

Figure 2-7 shows the FLR-L3 cache regions for the cores on the POWER7 processor die.

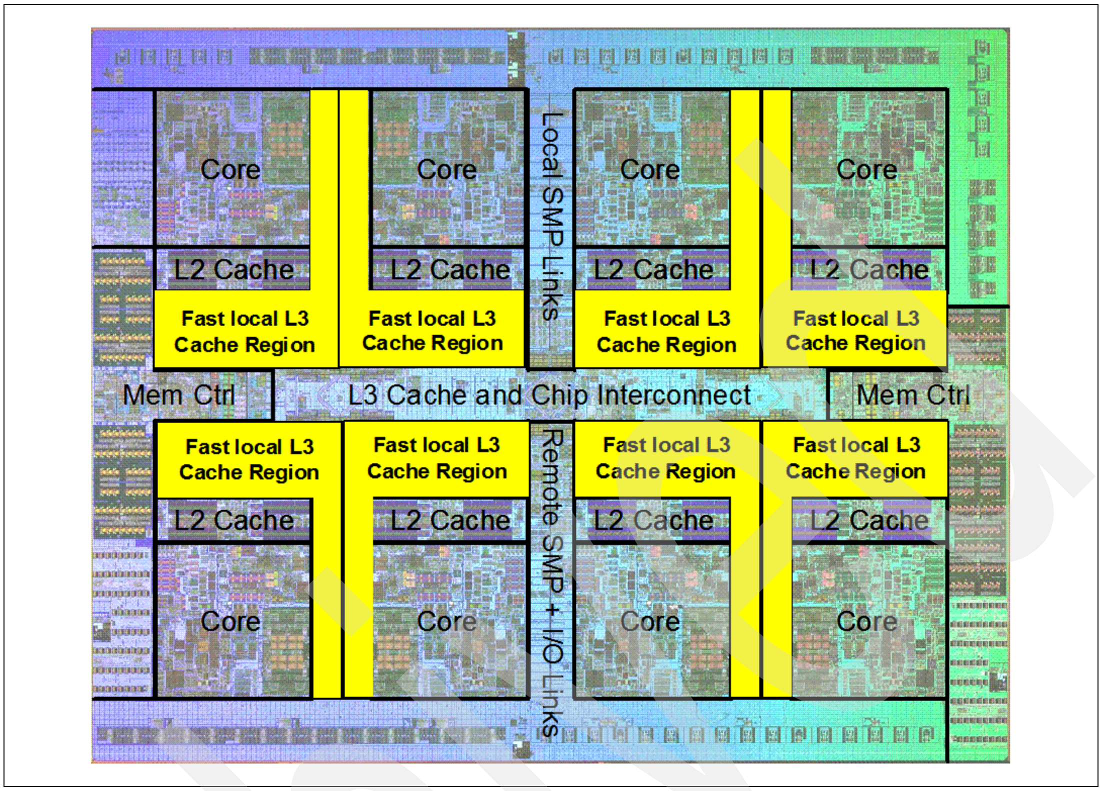

Core
Core
Core
Core
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Fast local L3 Cache Region
Fast local L3 Cache Region
Fast local L3 Cache Region
Fast local L3 Cache Region
Local SMP Links
Mem Ctrl
L3 Cache and Chip Interconnect
Mem Ctrl
Fast local L3 Cache Region
Fast local L3 Cache Region
Fast local L3 Cache Region
Fast local L3 Cache Region
Remote SMP + I/O Links
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Core
Core
Core
Core

Figure 2-7 FLR-L3 cache regions 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.

2.2.7 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 optimize energy usage and performance dynamically, so that the best possible balance is maintained. Intelligent energy features (such as EnergyScale) work with the BladeCenter Advanced Management Module (AMM) and IBM Systems Director Active Energy Manager™ to optimize processor speed dynamically, based on thermal conditions and system use.

For more information about the POWER7 energy management features see the following document:

Adaptive Energy Management Features of the POWER7 Processor

http://www.research.ibm.com/people/l/lefurgy/Publications/hotchips22_power7.pdf

TurboCore mode

Note: TurboCore mode is not available on the PS703 and PS704 blades.

TurboCore mode is a feature of the POWER7 processor but is not implemented in the PS703 and PS704 servers. It uses four cores per POWER7 processor chip with access to the entire 32 MB of L3 cache (8 MB per core) and at a faster processor core frequency, which delivers higher performance per core, and might save on software costs for those applications that are licensed per core.

2.2.8 Comparison of the POWER7 and POWER6 processors

Table 2-2 compares characteristics of various generations of POWER7 and POWER6 processors.

Note: This shows the characteristics of the POWER7 processors in general, but not necessarily as implemented in the POWER7 processor-based blade servers. Implementation specifics are noted.

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

Feature	POWER7 (PS703, PS704)	POWER7 (PS700, PS701, PS702)	POWER6+	POWER6
Technology	45 nm	45 nm	65 nm	65 nm
Die size	567 mm^2	567 mm^2	341 mm^2	341 mm^2
Maximum cores	8	8	2	2
Maximum SMT threads per core	4 threads	4 threads	2 threads	2 threads
L2 Cache	256 KB per core	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	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

Feature	POWER7 (PS703, PS704)	POWER7 (PS700, PS701, PS702)	POWER6+	POWER6
CPU frequency	2.4 GHz	3.0 GHz	5.0 GHz	4.2 GHz
Memory support	DDR3	DDR3	DDR2	DDR2
I/O Bus	Two GX++	Two GX++ (but operate in GX+ mode)	One GX+	One GX+
Enhanced Cache Mode (TurboCore)	No	No	No	No
Sleep & Nap Mode	Both	Both	Nap only	Nap only

2.3 POWER7 processor-based blades

The PS703 and PS704 are follow-ons to the previous generation blades, the PS700, PS701 and PS702. The PS700 blade contains a single processor socket with a 4-core processor and eight DDR3 memory DIMM slots. The PS701 blade contains a single processor socket with an 8-core processor and 16 DDR3 memory DIMM slots. The PS702 blade, a double-wide server, contains two processor sockets, each with an 8-core processor and a total of 32 DDR3 memory DIMM slots.

The PS703 blade contains two processor sockets, each with an 8-core processor and a total of 16 DDR3 memory DIMM slots. The PS704 blade contains four processor sockets, each with an 8-core processor and a total of 32 DDR3 memory DIMM slots. The cores in the PS700, PS701, and PS702 blades run at 3.0 GHz. The cores in the PS703 and PS704 blades run at 2.4 GHz.

POWER7 processor-based blades support POWER7 processors with various processor core counts. Table 2-3 summarizes the POWER7 processors for the PS700, PS701, PS702, PS703, and PS704.

Table 2-3 Summary of POWER7 processor options for the PS700, PS701, PS702, PS703, and PS704 blades

Model	Cores per POWER7 processor	Number of POWER7 processors	Total cores	Frequency (GHz)	L3 cache size per POWER7 processor (MB)
PS700	4	1	4	3.0	16
PS701	8	1	8	3.0	32
PS702	8	2	16	3.0	32
PS703	8	2	16	2.4	32
PS704	8	4	32	2.4	32

2.4 Memory subsystem

Each POWER7 processor has two intergrated memory controllers in the chip. However, the POWER7 blades only use one memory controller per processor and the second memory controller of the processor is unused. The PS703's two 8-core processors use a single memory controller per processor, which connects to four memory buffers per CPU, providing access to a total of 8 memory buffers and therefore 16 DDR3 DIMMS. The PS704's four 8-core processor chips use a single memory controller per processor chip, which connects to four memory buffers per CPU, providing access to a total of 16 memory buffers and therefore 32 DDR3 DIMMS.

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

2.4.1 Memory placement rules

The supported memory minimum and maximum for each server is listed in Table 2-4.

Table 2-4 Memory limits

Blade	Minimum memory	Maximum memory
PS703	16 GB	256 GB (16x 16 GB DIMMs)
PS704	32 GB	512 GB (32x 16 GB DIMMs)

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

Figure 2-9 shows the PS703 and PS704 physical memory DIMM topology.

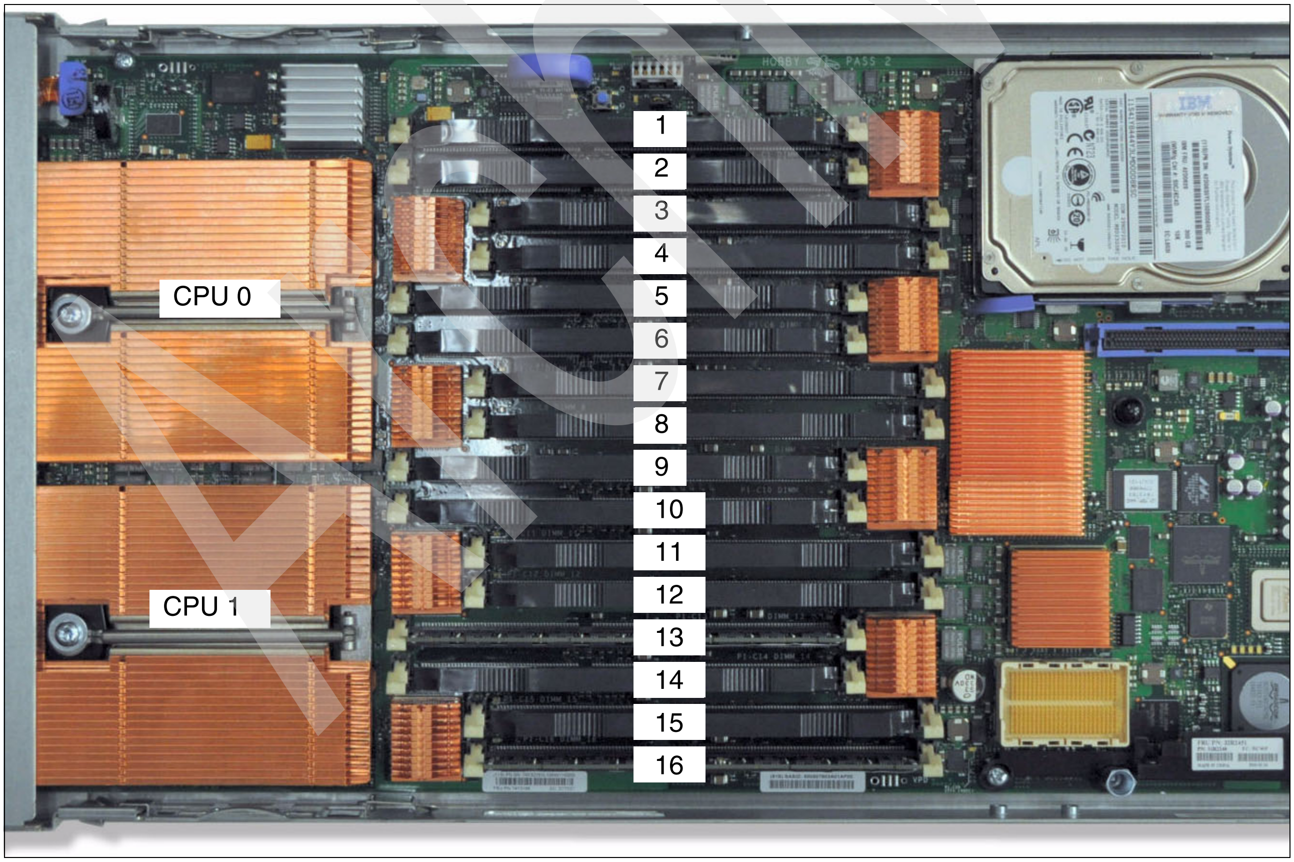


Figure 2-8 Memory DIMM topology for the PS703 and PS704

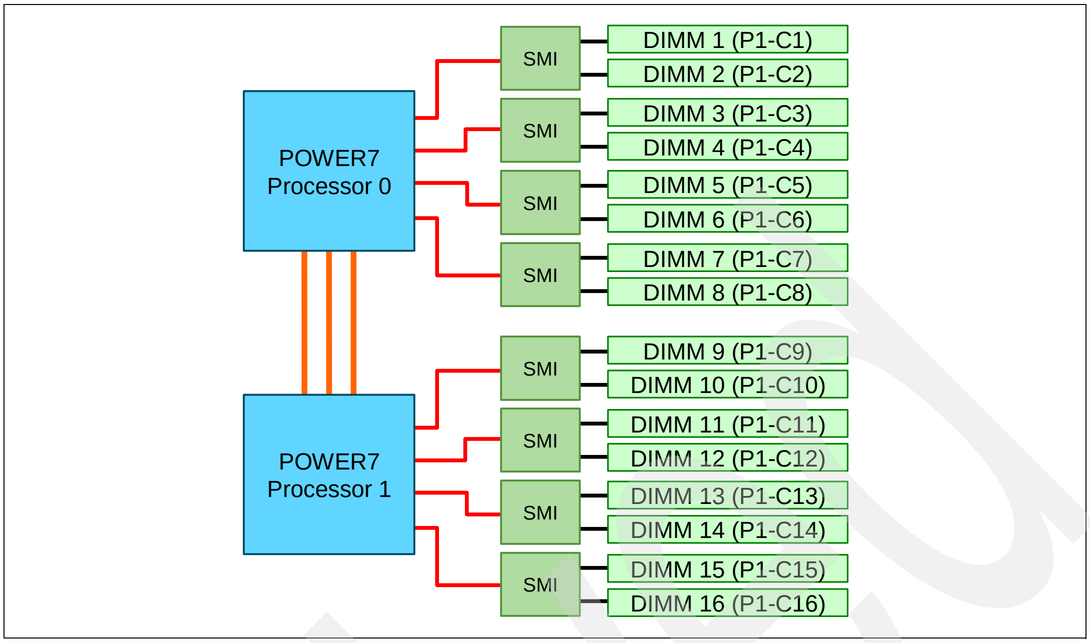


Figure 2-9 Memory DIMM topology

There are 16 buffered DIMM slots on the PS703 and PS704 base blade shown in Figure 2-9, with an additional 16 slots on the PS704 expansion unit. The PS703 and the PS704 base blade have slots labelled P1-C1 through P1-C16 as shown in Figure 2-9. For the PS704 expansion unit the numbering is the same except for the reference to the second planar board. The numbering is from P2-C1 through P2-C16.

The memory-placement rules are as follows:

- Install DIMM fillers in unused DIMM slots to ensure proper cooling.
- Install DIMMs in pairs (1 and 4, 5 and 8, 9 and 12, 13 and 16, 2 and 3, 6 and 7, 10 and 11, and 14 and 15).
- Both DIMMs in a pair must be the same size, speed, type, and technology. You can mix compatible DIMMs from different manufacturers.
- Each DIMM within a processor-support group (1-4, 5-8, 9-12, 13-16) must be the same size and speed.
- Install only supported DIMMs, as described on the ServerProven web site. See:

  `http://www.ibm.com/servers/eserver/serverproven/compat/us/`

DIMMs should be installed in specific DIMM sockets depending on the number of DIMMs to install. This is described in the following tables. See Figure 2-9 for DIMM socket physical layout compared to the DIMM location codes.

For the PS703, Table 2-5 shows the required placement of memory DIMMs depending on the number of DIMMs installed.

Table 2-5 PS703 DIMM placement rules

DIMM socket number	DIMM socket location code	PS703							
		Number of DIMMs to install							
		2	**4**	**6**	**8**	**10**	**12**	**14**	**16**
1	**P1-C1**	x	x	x	x	x	x	x	x
2	**P1-C2**					x	x	x	x
3	**P1-C3**					x	x	x	x
4	**P1-C4**	x	x	x	x	x	x	x	x
5	**P1-C5**			x	x	x	x	x	x
6	**P1-C6**							x	x
7	**P1-C7**							x	x
8	**P1-C8**			x	x	x	x	x	x
9	**P1-C9**		x	x	x	x	x	x	x
10	**P1-C10**						x	x	x
11	**P1-C11**						x	x	x
12	**P1-C12**		x	x	x	x	x	x	x
13	**P1-C13**				x	x	x	x	x
14	**P1-C14**								x
15	**P1-C15**								x
16	**P1-C16**				x	x	x	x	x

For the PS704, Table 2-6 shows the required placement of memory DIMMs depending on the number of DIMMs installed. The recommended practice is to match the DIMMs between the two system planars.

Table 2-6 PS704 DIMM placement

DIMM socket number	DIMM socket location code	PS704							
		Number of DIMMs to install							
		4	8	12	16	20	24	28	32
1	P1-C1	x	x	x	x	x	x	x	x
2	P1-C2					x	x	x	x
3	P1-C3					x	x	x	x
4	P1-C4	x	x	x	x	x	x	x	x
5	P1-C5			x	x	x	x	x	x
6	P1-C6							x	x
7	P1-C7							x	x
8	P1-C8			x	x	x	x	x	x
9	P1-C9		x	x	x	x	x	x	x
10	P1-C10						x	x	x
11	P1-C11						x	x	x
12	P1-C12		x	x	x	x	x	x	x
13	P1-C13				x	x	x	x	x
14	P1-C14								x
15	P1-C15								x
16	P1-C16				x	x	x	x	x
17	P2-C1	x	x	x	x	x	x	x	x
18	P2-C2					x	x	x	x
19	P2-C3					x	x	x	x
20	P2-C4	x	x	x	x	x	x	x	x
21	P2-C5			x	x	x	x	x	x
22	P2-C6							x	x
23	P2-C7							x	x
24	P2-C8			x	x	x	x	x	x
25	P2-C9		x	x	x	x	x	x	x
26	P2-C10						x	x	x
27	P2-C11						x	x	x
28	P2-C12		x	x	x	x	x	x	x
29	P2-C13				x	x	x	x	x
30	P2-C14								x
31	P2-C15								x
32	P2-C16				x	x	x	x	x

2.5 Active Memory Expansion

Optional Active Memory Expansion is a POWER7 technology that allows the effective maximum memory capacity to be much larger than the true physical memory. Innovative compression/decompression of memory content using processor cycles can allow memory expansion up to 100%.

This can allow an AIX 6.1 or later partition to do significantly more work with the same physical amount of memory, or a server to run more partitions and do more work with the same physical amount of memory.

Active Memory Expansion uses CPU resources to compress/decompress the memory contents. 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 workloads showed excellent results for many workloads in terms of memory expansion per additional CPU utilized. 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. After being turned on, there are monitoring capabilities in standard AIX performance tools such as **lparstat**, **vmstat**, **topas**, and **svmon**.

Figure 2-10 represents the percentage of CPU used to compress memory for two partitions with various profiles. The green curve corresponds to a partition that has spare processing power capacity, while the blue curve corresponds to a partition constrained in processing power.

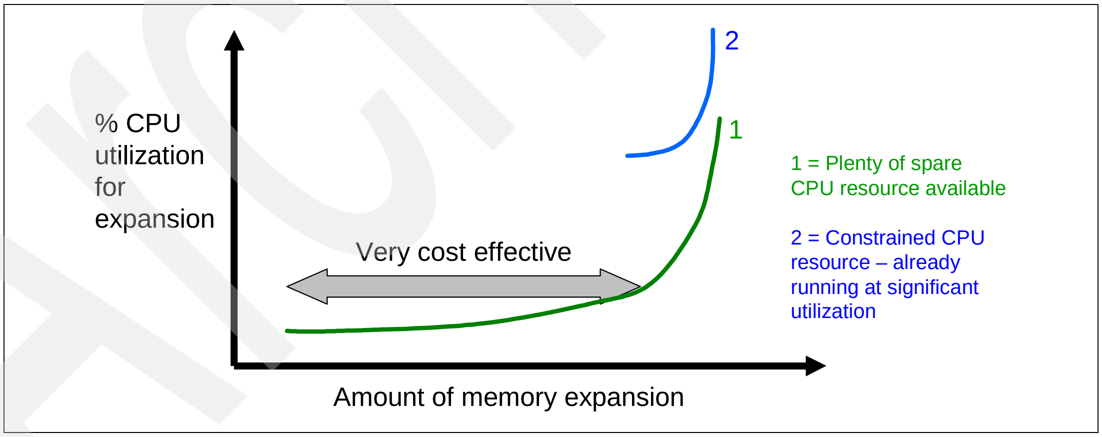


Figure 2-10 CPU usage versus memory expansion effectiveness

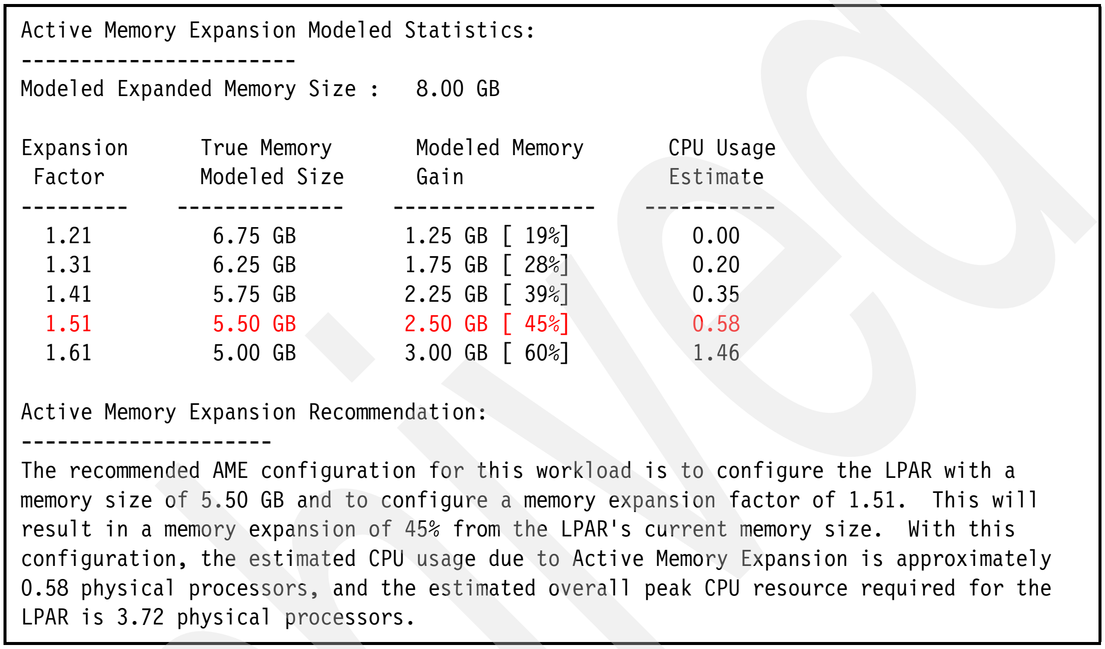
Both cases shows that there is a knee-of-curve relationship for CPU resources required for memory expansion:

- Busy processor cores do not have resources to spare for expansion.
- The more memory expansion is done, the more CPU resource is 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 perform this study, a planning tool is included with AIX 6.1 Technology Level 4 or later 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 2-11 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 and indicates one particular combination. In this example, the tool proposes to allocate 58% of a processor core, to benefit from 45% extra memory capacity.



```
Active Memory Expansion Modeled Statistics:
-----------------------
Modeled Expanded Memory Size :   8.00 GB

Expansion      True Memory       Modeled Memory       CPU Usage
 Factor        Modeled Size      Gain                 Estimate
---------    --------------    -----------------    -----------
  1.21          6.75 GB         1.25 GB [ 19%]         0.00
  1.31          6.25 GB         1.75 GB [ 28%]         0.20
  1.41          5.75 GB         2.25 GB [ 39%]         0.35
  1.51          5.50 GB         2.50 GB [ 45%]         0.58
  1.61          5.00 GB         3.00 GB [ 60%]         1.46

Active Memory Expansion Recommendation:
---------------------
The recommended AME configuration for this workload is to configure the LPAR with a
memory size of 5.50 GB and to configure a memory expansion factor of 1.51.  This will
result in a memory expansion of 45% from the LPAR's current memory size.  With this
configuration, the estimated CPU usage due to Active Memory Expansion is approximately
0.58 physical processors, and the estimated overall peak CPU resource required for the
LPAR is 3.72 physical processors.
```

Figure 2-11 Output form Active Memory Expansion planning tool

For more information on this topic see the white paper, *Active Memory Expansion: Overview and Usage Guide*, available from:

http://www.ibm.com/systems/power/hardware/whitepapers/am_exp.html

2.6 Technical comparison

Table 2-7 shows a comparison of the technical characteristics of the PS700, PS701, PS702, PS703, and PS704.

Table 2-7 Comparison of technical characteristics between PS700, PS701, PS702, PS703, and PS704

Systems characteristic	PS700	PS701	PS702	PS703	PS704
Single Wide (SW) or Double Wide (DW)	SW	SW	DW	SW	DW
Processor	4-cores at 3.0 GHz	8-cores at 3.0 GHz	16-cores at 3.0 GHz	16-cores at 2.4 GHz	32-cores at 2.4 GHz
L3 cache	4 MB On-chip eDRAM per core	4 MB On-chip eDRAM per core	4 MB On-chip eDRAM per core	4 MB On-chip eDRAM per core	4 MB On-chip eDRAM per core
Memory buffers	2	4	8	8	16
DIMM slots	8	16	32	16	32
DIMM speed	1066 MHz	1066 MHz	1066 MHz	1066 MHz	1066 MHz
Max Mem/core	16 GB	16 GB	16 GB	8 GB	8 GB
Anchor SVPD card	1	1	1	1	1
GX generation	GX+	GX+	GX+	GX++	GX++
GX speed/BW	1.25GHz@4B = 5Gb/s	1.25GHz@4B = 5Gb/s	1.25GHz@4B = 5Gb/s	2.5GHz@4B = 10Gb/s	2.5GHz@4B = 10Gb/s
PCIe Gen1 support	Yes	Yes	Yes	No	No
PCIe Gen2 support	No	No	No	Yes	Yes
PCI-X support	No	No	No	No	No
Integrated 1 Gb Ethernet ports	2x HEA	2x HEA	4x HEA	2x BCM5709S	4x BCM5709S
I/O Hub	1 P5IOC2	1 P5IOC2	2 P5IOC2	1 P7IOC	2 P7IOC
Embedded SAS controller	1	1	1	1	1
2.5” SAS HDD option	2	1	2	1	2
SSD option	No support	No support	No support	2x 1.8” SATA	4x 1.8” SATA
RAID function	0, 10	0	0, 10	0, 10	0, 10, 5, 6
BCM5387 5-port Ethernet switch	1	1	2	1	2
CIOv slot	1	1	2	1	2
CFFh slot	1	1	2	1	2

2.7 Internal I/O subsystem

Each POWER7 processor as implemented in the PS703 and PS704 blades utilizes a single GX++ bus from CPU0 to connect to the I/O subsystem as shown in Figure 2-12. The I/O subsystem is a GX++ multifunctional host bridge ASIC chip (“P7IOC” in Figure 2-12). The GX++ IO hub chip connects 6 PCIe ports, which provide access to the following devices:

- GX++ primary interface to the processor
- BCM5709S internal Ethernet controller
- USB controller
- CIOv card slot
- Embedded 3Gb SAS controller
- CFFh card slot

> **Note:** Table 2-2 on page 46 indicates there are two GX buses in the POWER7 processor; however, only one of them is active in the PS700, PS701, and PS703, and each planar in the PS702 and PS704.

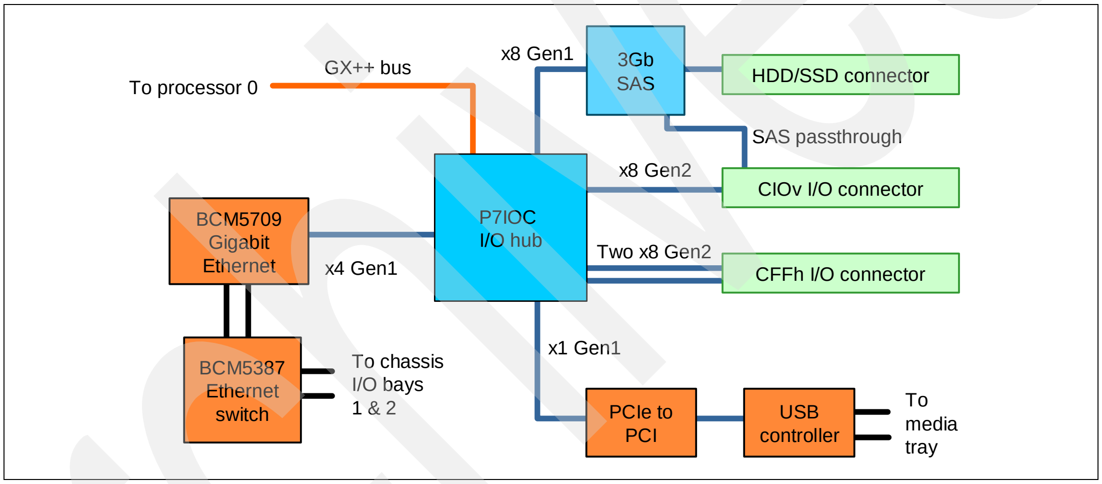


Figure 2-12 PS703 I/O Hub subsystem architecture

Figure 2-13 on page 56 shows the architecture of the I/O hub subsystem in the PS704 server, showing the base planar and SMP planar. The PS704 with four POWER7 processors has two GX++ multifunctional host bridge chips, one on each planar. The I/O system of the two hubs is duplicated on both planars except for the embedded SAS controller and the USB controller.

The PS704 has only one embedded 3Gb SAS controller located on the SMP planar that connects to both disk bays and CIOv card slots of the planars. When a CIOv SAS Pass-through card is used on the base planar the embedded SAS controller also connects to the ports of that card as well as the CIOv slot in the SMP planar. There is only one USB controller located on the base planar as depicted in the diagram.

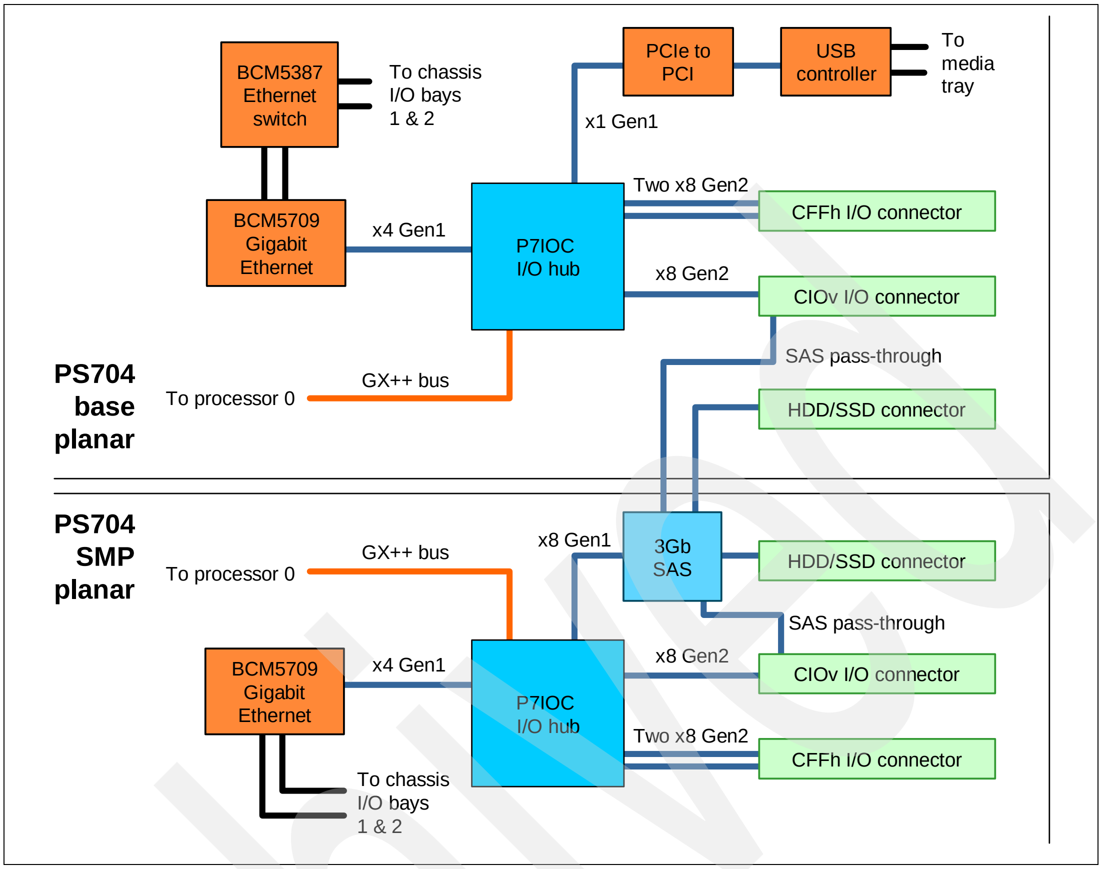


Figure 2-13 PS704 I/O Hub subsystem architecture

2.7.1 PCI Express bus

PCIe uses a serial interface and allows for point-to-point interconnections between devices using a 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 5 Gbps. These two pairs of wires is called a lane. A PCIe link might be comprised of multiple lanes. In such configurations, the connection is labeled as x1, x2, x8, x12, x16, or x32, where the number is the number of lanes.

The PCIe expansion card options for the PS700, PS701, PS702, PS703, and PS704 blades support Extended Error Handling (EEH). The card ports are routed through the BladeCenter mid-plane to predetermined I/O switch bays. The switches installed in these switch bays must match the type of expansion card installed, Ethernet, Fibre Channel, and so forth.

2.7.2 PCIe slots

The two PCIe slots are connected to the three x8 Gen2 PCIe links on the GX++ multifunctional host bridge chip. One of the x8 links supports the CIOv connector and the other two links support the CFFh connector on the blade. All PCIe slots are Enhanced Error Handling (EEH). PCI EEH-enabled adapters respond to a special data packet generated from the affected PCIe 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.

Expansion card form factors

There are two PCIe card form factors supported on the PS703 and PS704 blades:

- CIOv
- CFFh

CIOv form factor

A CIOv expansion card uses the PCI Express 2.0 x8 160 pin connector. A CIOv adapter requires compatible switch modules to be installed in bay 3 and bay 4 of the BladeCenter chassis. The CIOv card can be used in any BladeCenter that supports the PS703 and PS704 blades.

CFFh form factor

The CFFh expansion card attaches to the 450 pin PCIe Express connector of the blade server. In addition, the CFFh adapter can only be used in servers that are installed in the BladeCenter H, BladeCenter HT, or BladeCenter S chassis.

A CFFh adapter requires that either:

- A Multi-Switch Interconnect Module (MSIM) or MSIM-HT (BladeCenter HT chassis) is installed in bays 7 and 8, bays 9 and 10, or both.
- A high speed switch module be installed in bay 7 and bay 9.
- In the BladeCenter S, a compatible switch module is installed in bay 2.

The requirement of either the MSIM, MSIM-HT, or high-speed switch modules depends on the type of CFFh expansion card installed. The MSIM or MSIM-HT must contain compatible switch modules. See 1.7.6, “Multi-switch Interconnect Module” on page 33, or 1.7.7, “Multi-switch Interconnect Module for BladeCenter HT” on page 34, for more information about the MSIM or MSIM-HT.

The CIOv expansion card can be used in conjunction with a CFFh card in BladeCenter H, HT, and in certain cases a BladeCenter S chassis, depending on the expansion card type.

Table 2-8 lists the slot types, locations, and supported expansion card form factor types of the PS703 and PS704 blades.

Table 2-8 Slot configuration of the PS703 and PS704 blades

Card location	Form factor	PS703 location	PS704 location
Base blade	CIOv	P1-C19	P1-C19
Base blade	CFFh	P1-C20	P1-C20
Expansion blade	CIOv	Not present	P2-C19
Expansion blade	CFFh	Not present	P2-C20

Figure 2-14 shows the locations of the PCIe CIOv and CFFh connectors, and the physical location codes for the PS703.

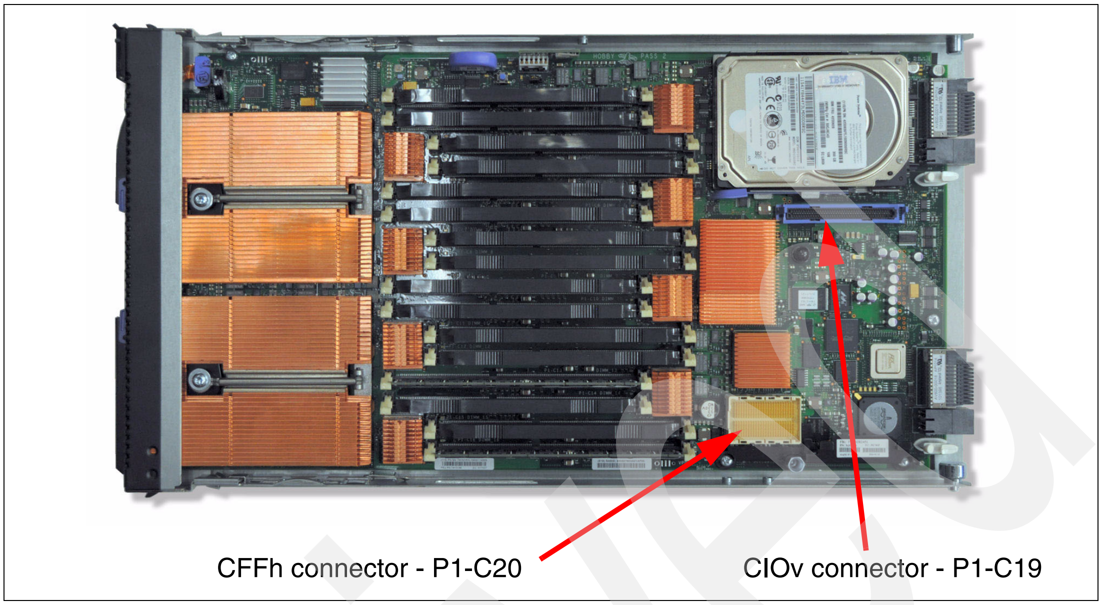


Figure 2-14 PS703 location codes for PCIe expansion cards

Figure 2-15 shows the locations of the PCIe CIOv and CFFh connectors for the PS704 base planar and the physical location codes. The expansion unit for the PS704 uses the prefix P2 for the slots on the second planar.

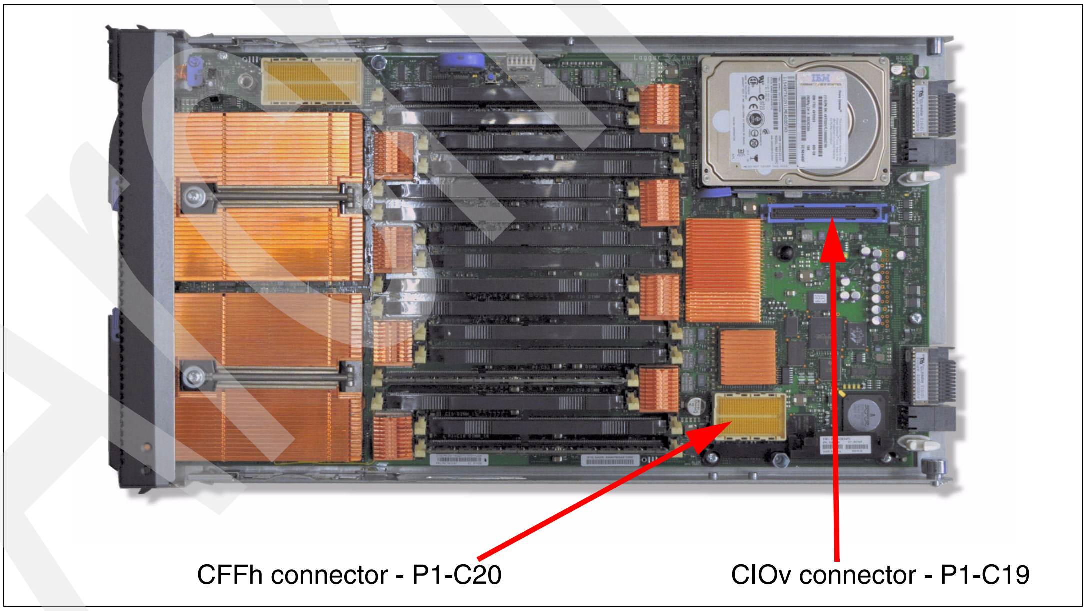


Figure 2-15 PS704 base location codes for PCIe expansion cards

Figure 2-16 shows the locations of the PCIe CIOv and CFFh connectors for the PS702 expansion blade (feature code 8358) and the physical location codes.

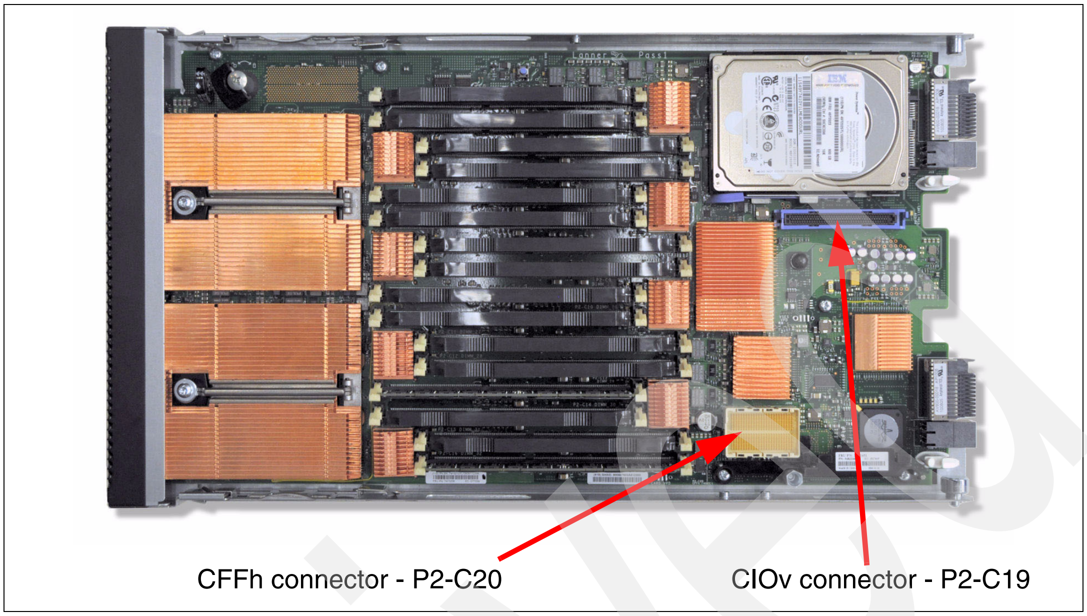


Figure 2-16 PS704 expansion blade location codes for PCIe expansion cards

BladeCenter I/O topology

There are no externally accessible ports on the PS703 and PS704 blades; all I/O is routed through a BladeCenter midplane to the I/O modules bays.

The I/O ports on all expansion cards are typically set up to provide a redundant pair of ports. Each port has a separate path through the mid-plane of the BladeCenter chassis to a specific I/O module bay. Figure 2-17 on page 60 through Figure 2-19 on page 61 show the supported BladeCenter chassis and the I/O topology for each.

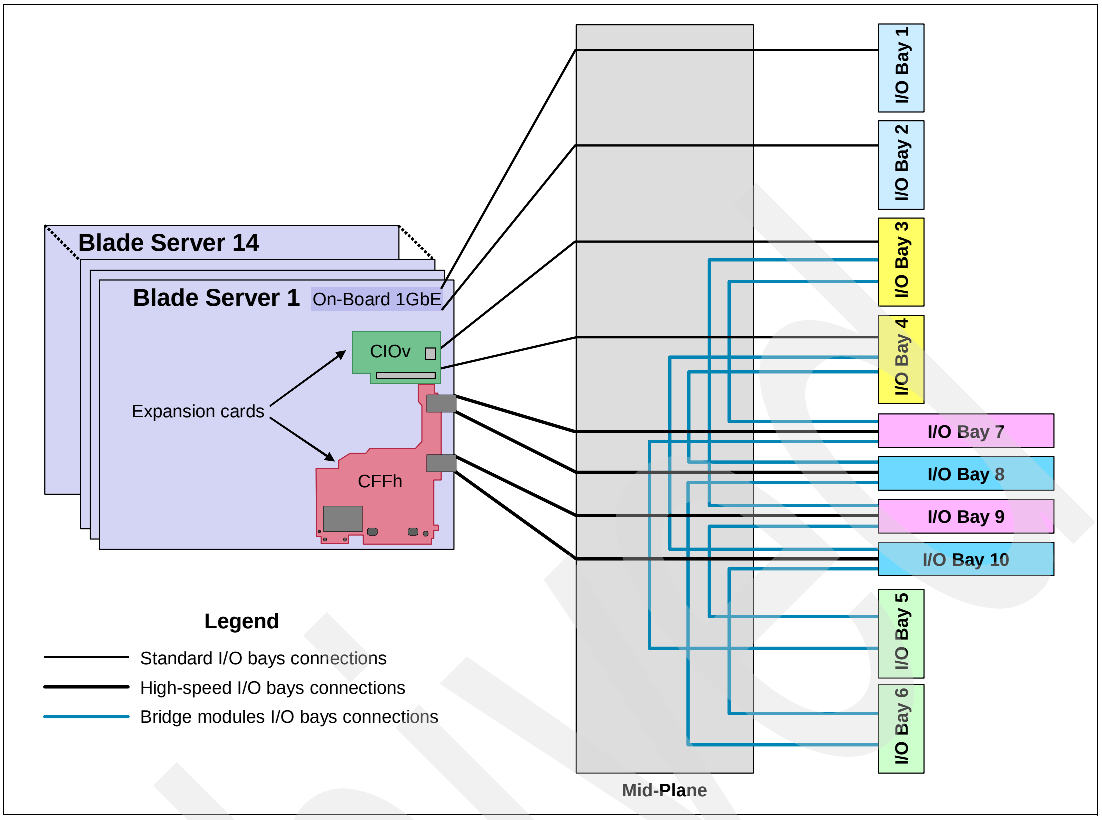


Figure 2-17 BladeCenter H I/O topology

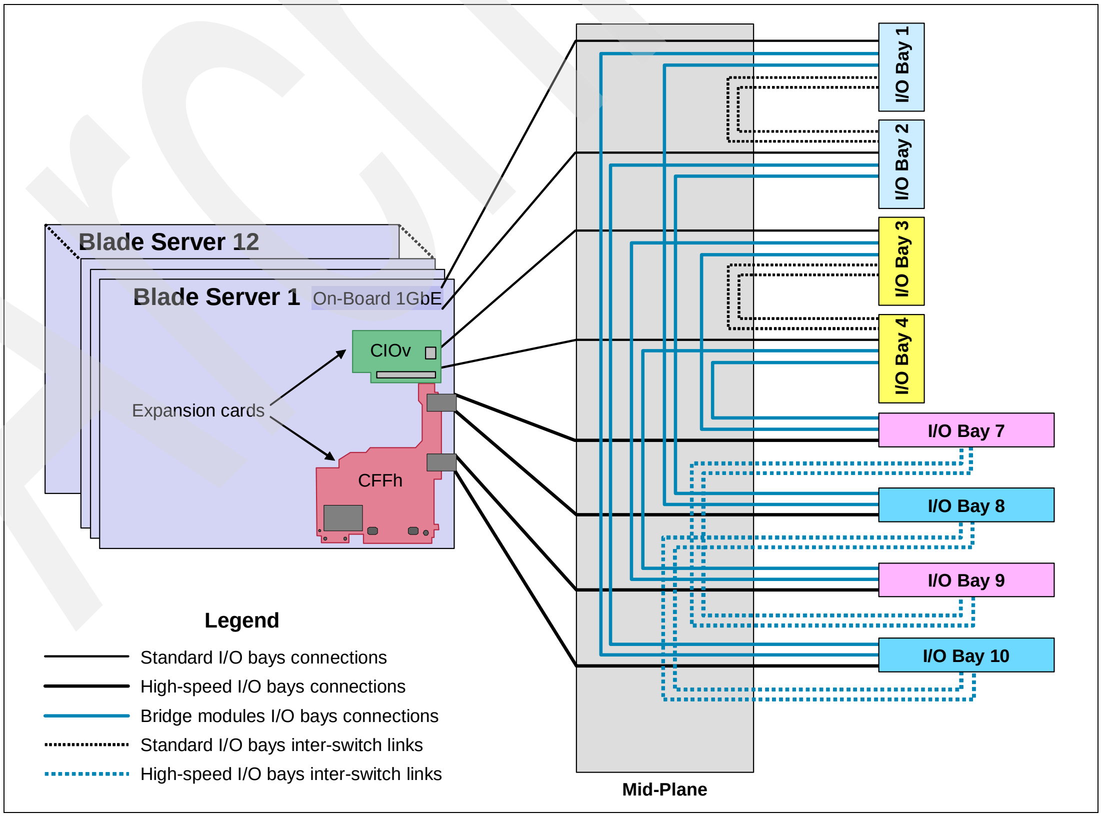

I/O Bay 1
I/O Bay 2
I/O Bay 3
I/O Bay 4
I/O Bay 7
I/O Bay 8
I/O Bay 9
I/O Bay 10
Blade Server 12
Blade Server 1
On-Board 1GbE
CIOv
Expansion cards
CFFh
Legend
Standard I/O bays connections
High-speed I/O bays connections
Bridge modules I/O bays connections
Standard I/O bays inter-switch links
High-speed I/O bays inter-switch links
Mid-Plane

Figure 2-18 BladeCenter HT I/O topology

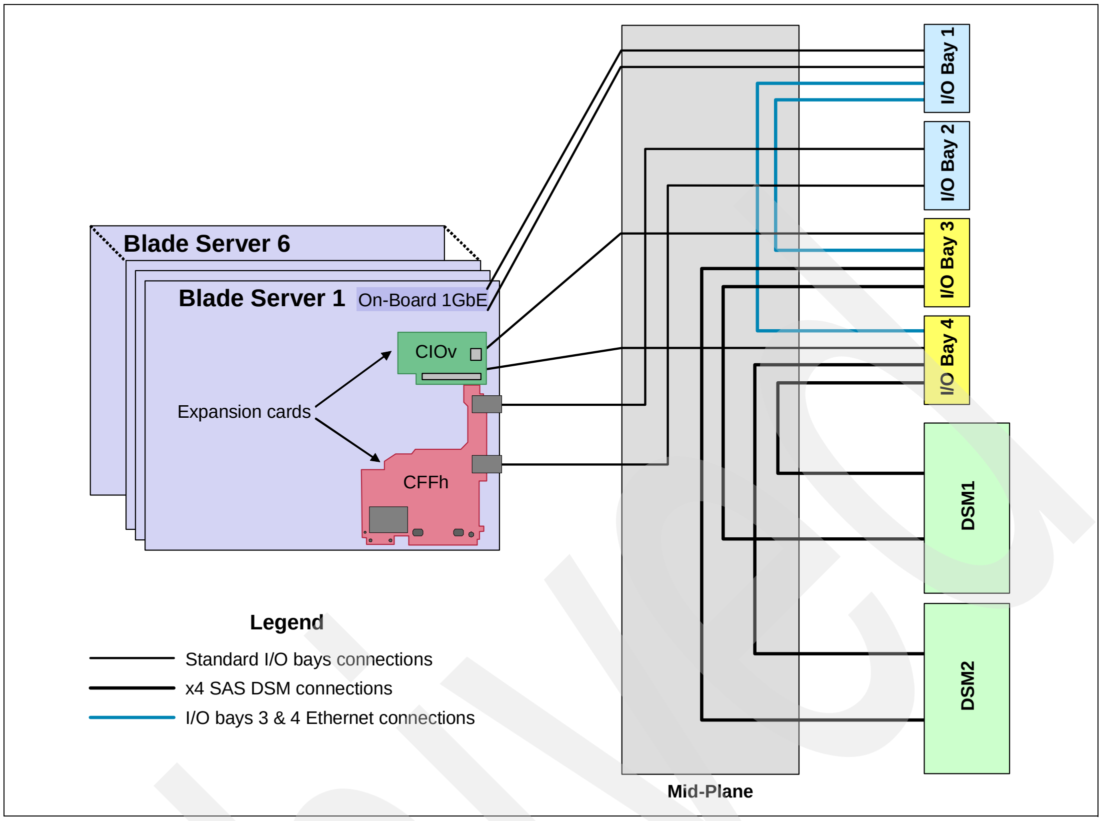


Figure 2-19 BladeCenter S I/O topology

2.7.3 I/O expansion cards

I/O expansion cards can provide additional resources that can be used by a native operating system, the Virtual I/O Server (VIOS), or assigned directly to a LPAR by the VIOS.

See 1.6.7, “I/O features” on page 21 for details about each supported card.

LAN adapters

In addition to the onboard 2-port Broadcom BCM5709S Ethernet controller, Ethernet ports can be added with LAN expansion card adapters. The PS703 and PS704 support expansion cards with Ethernet controllers as listed in Table 1-8 on page 21.

CIOv adapters require that Ethernet switches be installed in bays 3 and 4 of the BladeCenter chassis.

CFFh expansion cards are supported in BladeCenter H and HT. The Broadcom 2/4-Port Ethernet Expansion Card (CFFh) is also supported in the BladeCenter S. In the BC-H and BC-HT, the CFFh adapters require that Ethernet switches be installed in bays 7 and 9, and the Fibre Channel ports to switch bays 8 and 10. In the BladeCenter S only the Ethernet ports are usable and the connection is to Bay 2.

SAS adapter

To connect to external SAS devices, including the BladeCenter S storage modules, the 3 Gb SAS Passthrough Expansion Card and BladeCenter SAS Connectivity Modules are required.

The 3 Gb SAS Passthrough Expansion Card is a 2-port PCIe CIOv form factor card. The output from the ports on this card is routed through the BladeCenter mid-plane to I/O switch bays 3 and 4.

Fibre Channel adapters

The PS703 and PS704 blades support direct or SAN connection to devices using Fibre Channel adapters and the appropriate pass-through or Fibre Channel switch modules in the BladeCenter chassis. Fibre Channel expansion cards are available in both form factors and in 4 Gb and 8 Gb data rates.

The two ports on CIOv form factor expansion cards are connected to BladeCenter I/O switch module bays 3 and 4. The two Fibre Channel ports on a CFFh expansion card connect to BladeCenter H or HT I/O switch bays 8 and 10. The Fibre Channel ports on a CFFh form factor adapter are not supported for use in a BladeCenter S chassis.

Fibre Channel over Ethernet (FCoE)

A new emerging protocol, 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 Ethernet). This evolution of FCoE makes network consolidation a reality; the combination of Fibre Channel and Ethernet enables a consolidated network that maintains the resiliency, efficiency, and seamlessness of the existing FC-based data center.

Figure 2-20 shows a configuration using BladeCenter FCoE components.

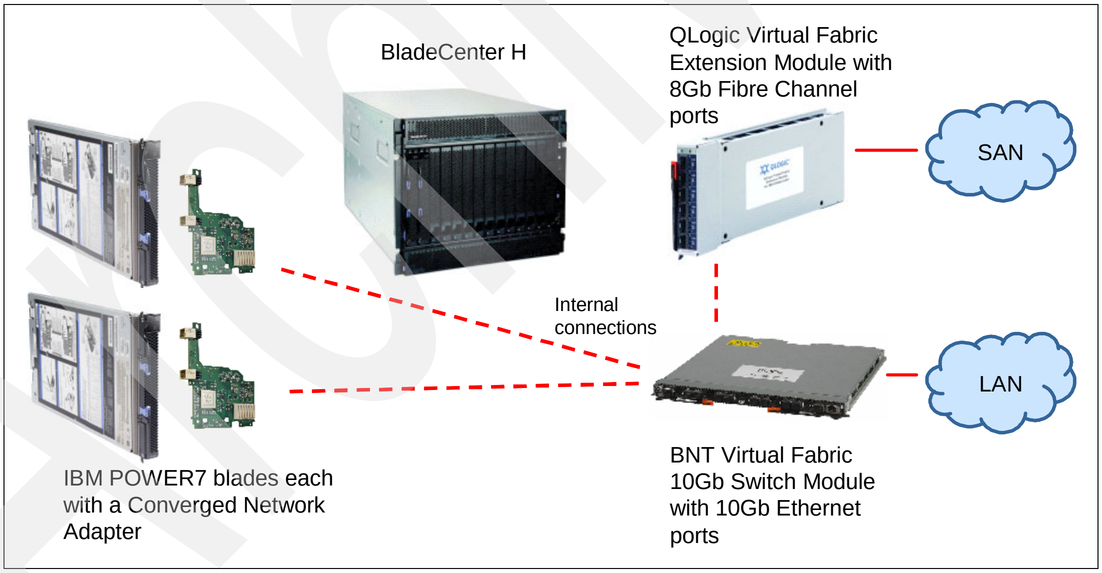

BladeCenter H
QLogic Virtual Fabric Extension Module with 8Gb Fibre Channel ports
SAN
Internal connections
LAN
IBM POWER7 blades each with a Converged Network Adapter
BNT Virtual Fabric 10Gb Switch Module with 10Gb Ethernet ports

Figure 2-20 FCoE connections in IBM BladeCenter

For more information about FCoE, read *An Introduction to Fibre Channel over Ethernet, and Fibre Channel over Convergence Enhanced Ethernet*, REDP-4493, available from the following web page:

http://www.redbooks.ibm.com/abstracts/redp4493.html

The QLogic 2-port 10 Gb Converged Network Adapter is a CFFh form factor card. The ports on this card are connected to BladeCenter H and HT I/O switch module bays 7 and 9. In these bays a passthrough or FCoE-capable I/O module can provide connectivity to a top-of-rack switch. A combination of the appropriate I/O switch module in these bays and the proper Fibre Channel-capable modules in bays 3 and 5 can eliminate the top-of-rack switch requirement. See 1.7, “Supported BladeCenter I/O modules” on page 28.

InfiniBand Host Channel adapter

The InfiniBand Architecture (IBA) is an industry-standard architecture for server I/O and interserver 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 is an open set of interconnected standards and specifications. The main InfiniBand specification has been published by the InfiniBand Trade Association and is available at the following web page:

http://www.infinibandta.org/

InfiniBand is based on a switched fabric architecture of serial point-to-point links. These InfiniBand 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, available from the following web page:

http://www.redbooks.ibm.com/abstracts/sg247767.html

The 4X InfiniBand QDR Expansion Card is a 2-port CFFh form factor card and is only supported in a BladeCenter H chassis. The two ports are connected to the BladeCenter H I/O switch bays 7 and 8, and 9 and 10. A supported InfiniBand switch is installed in the switch bays to route the traffic either between blade servers internal to the chassis or externally to an InfiniBand fabric.

2.7.4 Embedded SAS Controller

The embedded 3 Gb SAS controller is connected to one of the Gen1 PCIe x8 buses on the GX++ multifunctional host bridge chip. The PS704 uses a single embedded SAS controller located on the SMP expansion blade.

More information about the SAS I/O subsystem can be found in 2.9, “Internal storage” on page 68.

2.7.5 Embedded Ethernet Controller

The Broadcom 2-port BCM5709S network controller has its own Gen1 x4 connection to the GX++ multifunctional host bridge chip. The WOL, TOE, iSCSI, and RDMA functions of the BCM5709S are not implemented on the PS703 and PS704 blades.

The connections are routed through the 5-port Broadcom BCM5387 Ethernet switch ports 3 and 4. Then port 0 and port 1 of the BCM5387 connect to the Bladecenter chassis. Port 2 of the BCM5387 is connected to the FSP to provide its connection to the chassis for SOL connectivity.

See 2.8.1, “Server console access by SOL” on page 65 for more details concerning the SOL connection.

> **Note:** The PS703 and PS704 blades do not provide two Host Ethernet Adapters (HEA) as the previous PS700, PS701, and PS703 blades did. This also means the Integrated Virtual Ethernet (IVE) feature is not available on the PS703 and PS704 blades. Broadcom 5709S Ethernet ports can be virtualized by PowerVM VIOS software.

MAC addresses for BCM5709S Ethernet ports

Each of the two BCM5709S Ethernet ports is assigned one physical MAC address. This is different from the HEA, where each logical port of the HEA had its own MAC address. When VIOS (Virtual IO System) is used on the PS703 or PS704 blades, it assigns the MAC virtual address that corresponds to the BMC5709S physical addresses as appropriate. Thus the total number of required MAC addresses for each PS703 and PS704 base planar is four, two for FSP and two for Broadcom BCM5709S. On the PS704 SMP planar, only two MAC addresses are needed for 5709S because there is no FSP.

Each planar has a label that lists the MAC addresses. The first two listed are those of FSP enet0 and enet1 respectively. The next two listed are for BCM5709S port0 and port1 respectively.

2.7.6 Embedded USB controller

The USB controller complex is connected to the Gen1 PCIe x1 bus of the GX++ multifunctional host bridge chip as shown in Figure 2-1 on page 38.

This embedded USB controller provides support for four USB 2.0 root ports, which are routed to the BladeCenter chassis midplane. However, only two are used:

- Two USB ports are connected to the BladeCenter Media Tray, providing access to the optical drive (if installed) and external USB ports to the blade server.
- The other two ports are not connected. (On other servers, these USB connections are for keyboard and mouse function, which the PS703 and PS704 blades do not implement.)

> **Note:** The PS703 and PS704 blades do not support the KVM function from the AMM. If the mouse and keyboard are plugged into the blade center, they are not operational with the PS703 and PS704 blades. You must use SOL via the AMM or an SDMC virtual console to connect to the blade. See 5.5.6, “Virtual consoles from the SDMC” on page 177 for more details.

The BladeCenter Media Tray, depending on the BladeCenter chassis used, can contain up to two USB ports, one optical drive and system status LEDs.

For information about the different media tray options available by BladeCenter model see *IBM BladeCenter Products and Technology*, SG24-7523 available from:

`http://www.redbooks.ibm.com/abstracts/sg247523.html`

The media tray is a shared USB resource that can be assigned to any single blade slot at one time, providing access to the chassis USB optical drive and USB ports.

2.8 Service processor

The Flexible Service Processor (FSP) is used to monitor and manage the system hardware resources and devices. In a POWER7-based blade implementation the external network connection for the service processor is routed through an on-blade BCM5387 Ethernet switch, through the BladeCenter midplane, chassis switches and to the AMM. The Serial over LAN (SOL) connection for a system console uses this same connection. When the blade is in standby power mode the service processor responds to AMM instructions and can detect Wake-on-LAN (WOL) packets. The PS703 and PS704 blades have only one network port available for configuring the service processor IP address in the AMM.

The PS703 has a single service processor. The PS704 has a second service processor in the expansion unit. However, it is only used for controlling and managing the hardware on this second planar.

2.8.1 Server console access by SOL

The PS703 and PS704 blades do not have an on-board video chip and do not support KVM connections. Server console access is obtain by a SOL connection only. The AMM direct KVM and remote control feature are not available to the PS703 and PS704 blades.

SOL provides a means to manage servers remotely by using a command-line interface (CLI) over a Telnet or secure shell (SSH) connection. SOL is required to manage servers that do not have KVM support or that are attached to an SDMC. SOL provides console redirection for both System Management Services (SMS) and the blade server operating system. The SOL feature redirects server serial-connection data over a LAN without requiring special cabling by routing the data via the AMM network interface. The SOL connection enables blade servers to be managed from any remote location with network access to the AMM.

SOL offers the following advantages:

- Remote administration without keyboard, video, or mouse (headless servers)
- Reduced cabling and no requirement for a serial concentrator
- Standard Telnet interface, eliminating the requirement for special client software

The IBM BladeCenter AMM CLI provides access to the text-console command prompt on each blade server through a SOL connection, enabling the blade servers to be managed from a remote location.

In the BladeCenter environment, the SOL console data stream from a blade is routed from the blade's service processor to the AMM through the Ethernet switch on the blade's system board. The signal is then routed through the network infrastructure of the BladeCenter unit to the Ethernet switch modules installed in bay 1 or 2.

> **Note:** Link Aggregation is not supported with the SOL Ethernet port.

Figure 2-21 on page 66 shows the SOL traffic flow and the Gigabit Ethernet production traffic flow.

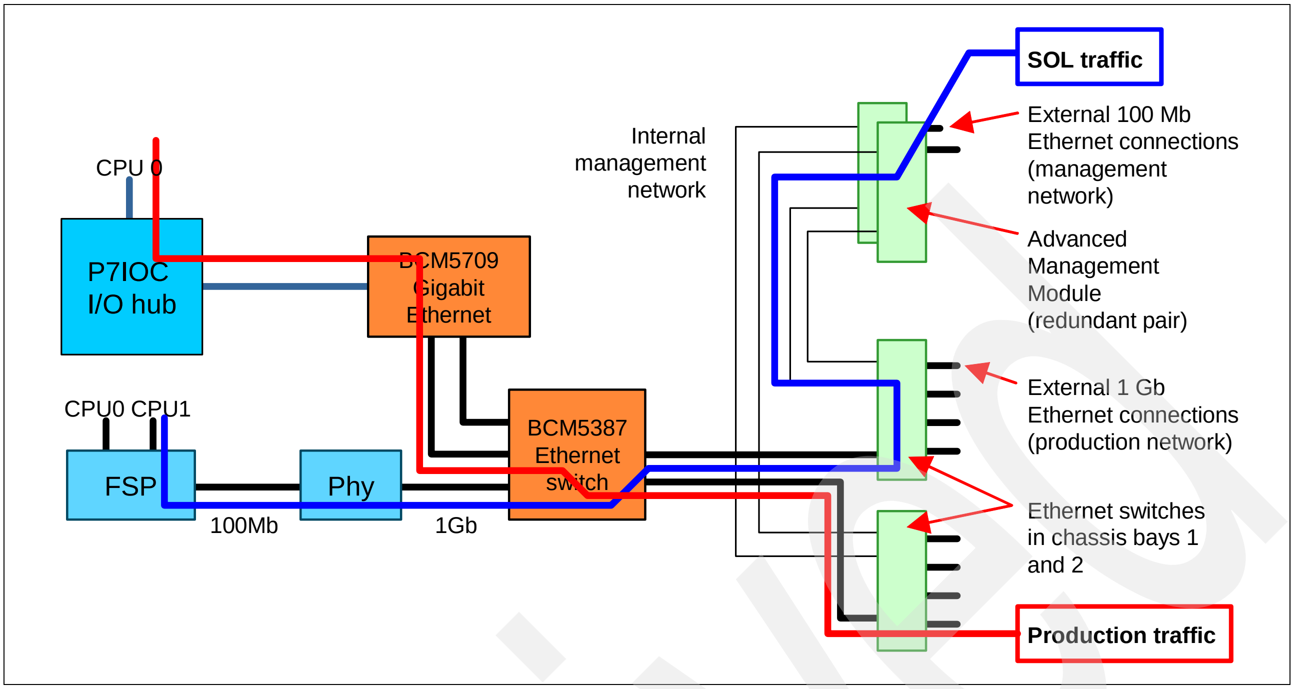

SOL traffic
External 100 Mb Ethernet connections (management network)
Advanced Management Module (redundant pair)
External 1 Gb Ethernet connections (production network)
Ethernet switches in chassis bays 1 and 2
Production traffic
Internal management network
CPU 0
P7IOC I/O hub
BCM5709 Gigabit Ethernet
CPU0 CPU1
FSP
Phy
BCM5387 Ethernet switch
100Mb
1Gb

Figure 2-21 SOL service processor to AMM connection

BladeCenter components are configured for SOL operation through the BladeCenter AMM. The AMM also acts as a proxy in the network infrastructure to couple a client running a Telnet or SSH session with the management module to an SOL session running on a blade server, enabling the Telnet or SSH client to interact with the serial port of the blade server over the network.

Because all SOL traffic is controlled by and routed through the AMM, administrators can segregate the management traffic for the BladeCenter unit from the data traffic of the blade servers. To start an SOL connection with a blade server, follow these steps:

1. Start a Telnet or SSH CLI session with the AMM.
2. Start a remote-console SOL session with any blade server in the BladeCenter unit that is set up and enabled for SOL operation.

You can establish up to 20 separate web-interface, Telnet, or SSH sessions with a BladeCenter AMM. For a BladeCenter unit, this step enables you to have 14 simultaneous SOL sessions active (one for each of up to 14 blade servers) with six additional CLI sessions available for BladeCenter unit management.

With a BladeCenter S unit you have six simultaneous SOL sessions active (one for each of up to six blade servers) with 14 additional CLI sessions available for BladeCenter unit management. If security is a concern, you can use Secure Shell (SSH) sessions, or connections made through the serial management port that is available on the AMM, to establish secure Telnet CLI sessions with the BladeCenter management module before starting an SOL console-redirect session with a blade server.

SOL has the following requirements:

- An Ethernet switch module or Intelligent Pass-Thru Module must be installed in bay 1 of a BladeCenter (SOL does not operate with the I/O module in bay 2).
- SOL must be enabled for those blades that you want to connect to with SOL.

- The Ethernet switch module must be set up correctly.

> **Note:** The AMM has an option called *management channel auto-discovery* (MCAD) which allows certain blades to dynamically change the path for the management network within the chassis in order to use any IO module slot. The JS20, JS21, JS12, JS22, JS23, JS43, PS700, PS701, PS702, PS703, and PS704 do not support the use of MCAD. You must have a switch in IO module bay 1.
>
> For more information about MCAD see the InfoCenter site at:
>
> http://publib.boulder.ibm.com/infocenter/bladectr/documentation/index.jsp?topic=/com.ibm.bladecenter.advmgtmod.doc/kp1bb_bc_mmug_usemcad.html

For details about setting up SOL, see the *BladeCenter Serial Over LAN Setup Guide*, which can be found at the following web page:

http://ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-54666

This guide contains an example of how to establish a Telnet or SSH connection to the management module and then an SOL console.

2.8.2 Anchor card

The anchor card contains the Smart VPD chip that stores system-specific information. The same anchor card is used for both the single-wide base blade and for the double-wide SMP blade. The pluggable anchor card provides a means for the system-specific information to be transferable from a faulty system planar to the replacement planar.

Before the service processor knows what system it resides on, it reads the SmartChip VPD to obtain system information. There is only one anchor card for the PS703 and the PS704, as shown in Figure 2-22. The PS704 base planar holds the anchor card; the SMP planar does not have an anchor card.

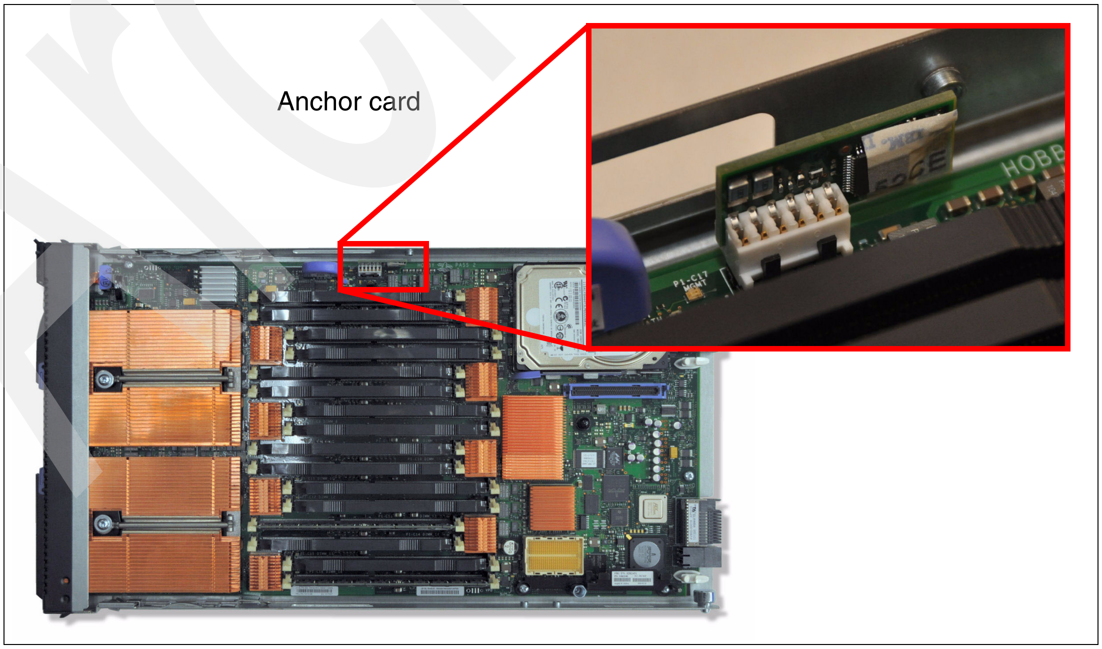


Figure 2-22 Anchor card location

Note: The anchor card used in the PS700-PS704, unlike the card used in the JS series of blades, does not contain LPAR or virtual server configuration data (such as CPU and mmmemory allocations). When this card is swapped during a blade replacement operation to the new blade the configuration information must be restored from a profile.bak file generated on the old blade by the VIOS prior to the replacement. The command to generate the profile.bak file is **`bkprofdata`**. The VIOS command to restore profile data is **`rstprofdata`**. Generating a profile.bak should be part of good administrative practices for POWER-based blades.

2.9 Internal storage

PS703 and PS704 blades use a single integrated 3 Gb SAS controller. The controller attaches to the IO hub PCIe Gen1 connector operating at 2.5 Gbps. The PS704 has a single embedded SAS controller located on the SMP planar. The PS704 base planar has no SAS controller.

The PS703 blade embedded SAS controller provides a standard 2.5” connector for the internal drive, and 2 ports through the CIOv connector to the optional 3 Gb SAS Passthrough Expansion Card to the BladeCenter SAS switch modules. The SAS controller ports used for the internal disk drives can support a single 2.5-inch SAS hard disk drive (HDD) or two 1.8-inch SATA solid state drives (SSD) via the SAS to SATA interposer card at each DASD bay location, as shown in Figure 2-23.

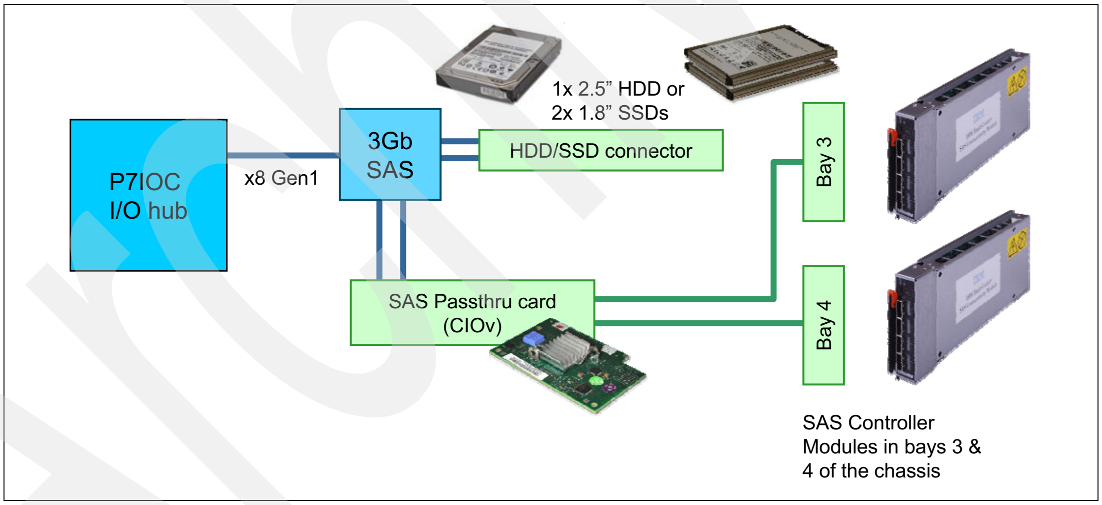


Figure 2-23 PS703 SAS configuration

Figure 2-24 shows the physical locations and codes for the HDDs in the PS703.

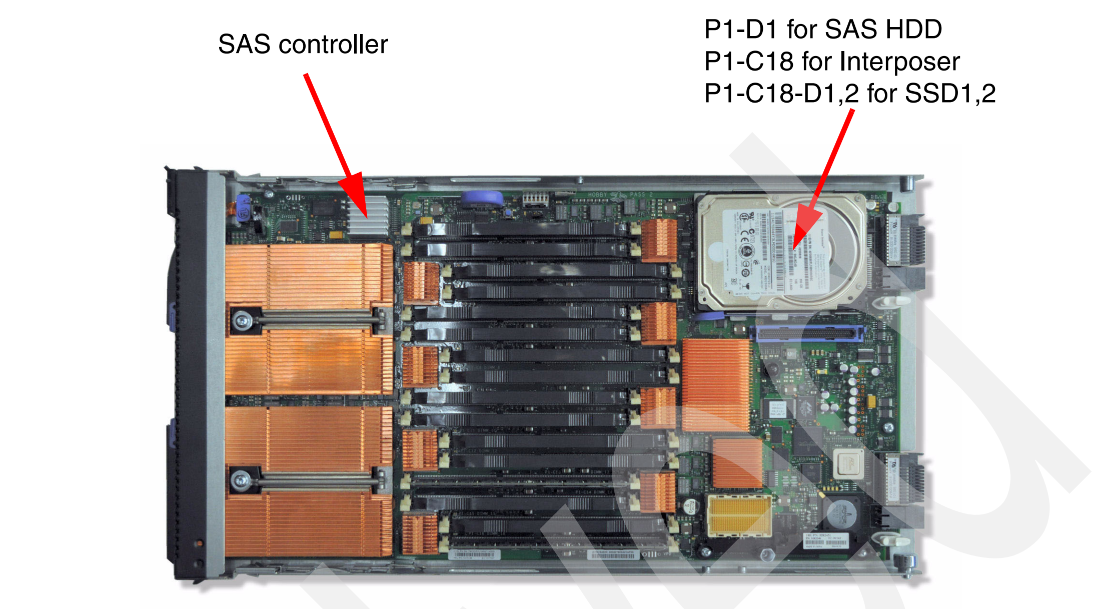


Figure 2-24 HDD location and physical location code PS703

In the PS704 blade, the SAS controller is located on the SMP planar. A total of eight SAS ports are used in the PS704 blade, four of which are used on the SMP planar and the other four are routed from the SMP planar to the base planar. So for each planar of the PS704 there are two ports for each DASD slot and two ports for each CIOv connector, as shown in Figure 2-25.

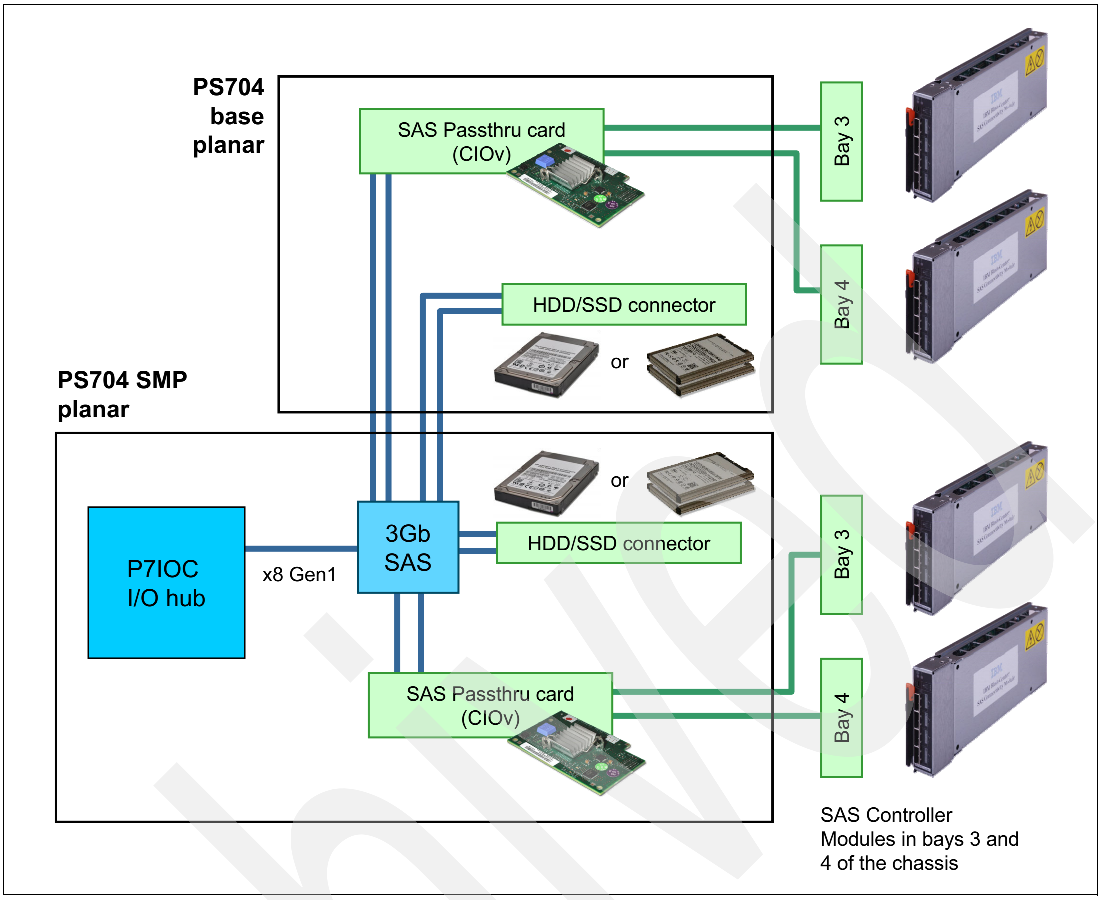


Figure 2-25 PS704 SAS configuration

Figure 2-26 shows the physical location and code for a HDD in a PS704 base planar.

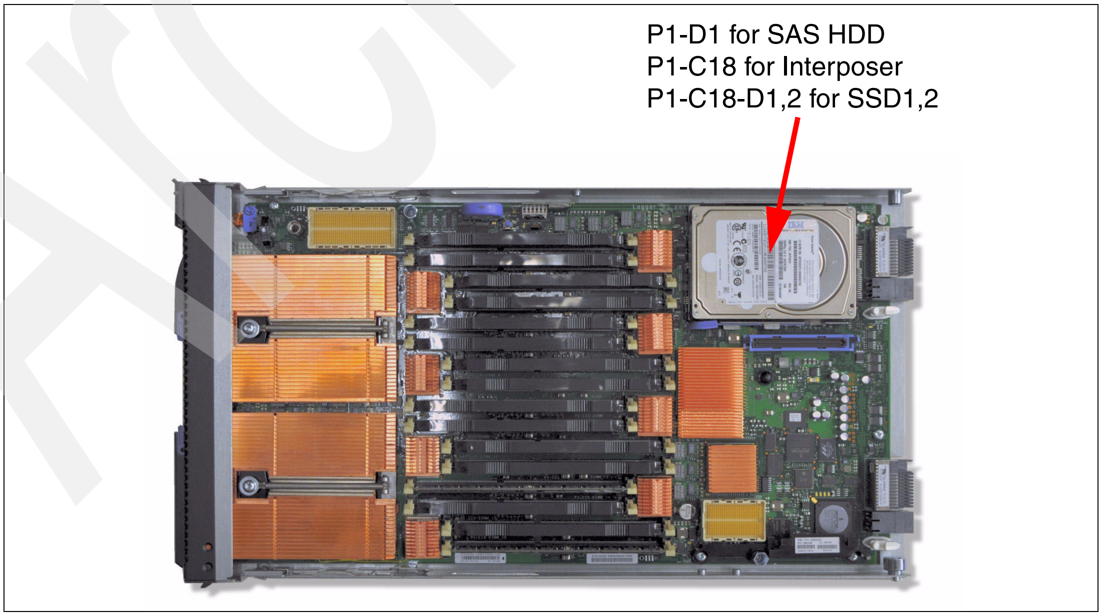


Figure 2-26 HDD location and physical location code PS704 base planar

Figure 2-27 shows the physical location and code for a HDD in a PS704 SMP planar.

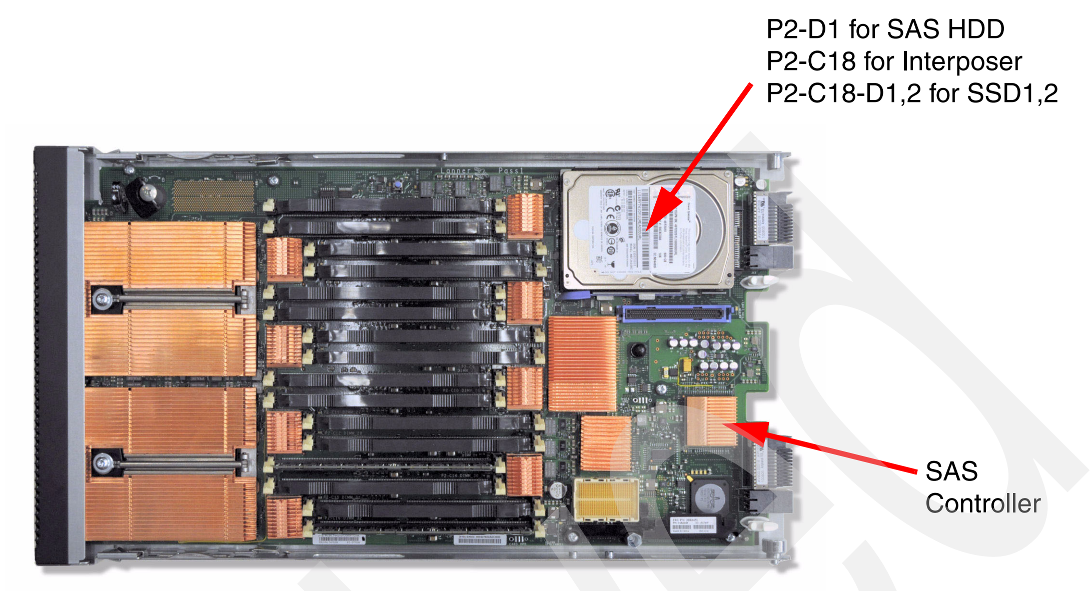
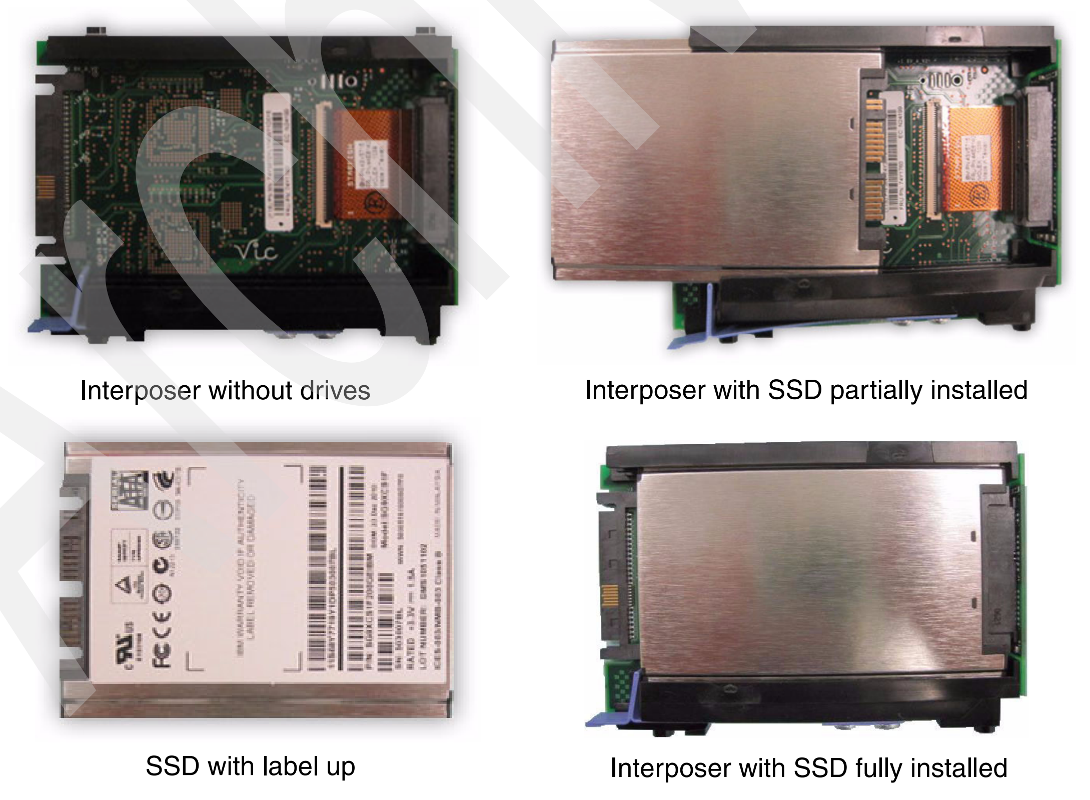


Figure 2-27 HDD location and physical location code PS704 SMP planar

Figure 2-28 shows the SATA SSD interposer card used to connect the 1.8-inch SATA SSD drives to the drive bay.



Interposer without drives

Interposer with SSD partially installed

SSD with label up

Interposer with SSD fully installed

Figure 2-28 SATA SSD Interposer card

Note: The SSDs used in the PS703 and PS704 servers are formatted in 528-byte sectors and are only useable in RAID arrays or as hot spares. Each device contains metadata written by the onboard 3Gb SAS controller to verify the RAID configuration. Error logs will exist if problems are encountered. Refer to the appropriate service documentation should errors occur.

2.9.1 Hardware RAID function

For the PS703, the supported RAID functions are as follows:

- 1 HDD - RAID 0
- 1 SSD - RAID 0
- 2 SSDs - RAID 0, 10

For the PS704, the supported RAID functions are as follows:

- 2 HDDs - RAID 0, 10
- 1 HDD and 1 SSD - RAID 0 on each disk (Combining HDD and SDD is not allowed in other RAID configurations.)
- 2 SSDs - RAID 0, 10
- 3 SSDs - RAID 0, 5, 10 (RAID 10 on 2 disks; the third disk then can be a hot spare.)
- 4 SSDs - RAID 0, 5, 6, 10 (Hot spare with RAID 5 is possible, that is, 3 disks in RAID 5 and one disk as a hot spare.)

Drives in the PS703 or PS704 blade server can be used to implement and manage various types of RAID arrays in operating systems that are on the ServerProven list. For the blade server, you must configure the RAID array through the Disk Array Manager. The AIX Disk Array Manager is packaged with the Diagnostics utilities on the Diagnostics CD for instances when the operating system has not been installed yet. Use **`smit sasdam`** to use the AIX Disk Array Manager to configure the disk drives for use with the SAS controller when there is an operating system installed. The Disk Array Manager only provides the ability to configure RAID 0, 5, 6 or 10. To achieve RAID1 mirror functionality use the RAID 10 option with two internal disk drives.

You can configure a hot spare disk if there are enough disks available. A hot spare is a disk that can be used by the controller to automatically replace a failed disk in a degraded RAID 5, 6, or 10 disk array. A hot spare disk is useful only if its capacity is greater than or equal to the capacity of the smallest disk in an array that becomes degraded.

For more information about hot spare disks, see "Using hot spare disks" in the Systems Hardware Information Center at:

`http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/arebj/sasusinghotsparedisks.htm`

Note: Before you can create a RAID array, you must reformat the drives so that the sector size of the drives changes from 512 bytes to 528 bytes. If you later decide to remove the drives, delete the RAID array before you remove the drives. If you decide to delete the RAID array and reuse the drives, you might need to reformat the drives so that the sector size of the drives changes from 528 bytes to 512 bytes.

For more information, see “Using the Disk Array Manager” in the Systems Hardware Information Center at:

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp?topic=/arebj/sasusingthesasdiskarraymanager.htm

2.9.2 External SAS connections

The onboard SAS controller in the PS703 and PS704 does not provide a direct access external SAS port. However, by using a 3 Gb SAS Passthrough Expansion Card and BladeCenter SAS Connectivity Modules, two ports on the SAS controller (four in the PS704 with a second SAS card on the expansion unit) are expanded, providing access to BladeCenter S Disk Storage Modules (DSM) or an external SAS disk sub-system.

The Disk Storage Module, part number 43W3581 feature 4545 must be installed in the BladeCenter S chassis to support external SAS storage devices outside the chassis using the SAS Connectivity Card. No disk drives need to be installed in the DSM to support the external storage.

Note: To also use external drives in a RAID array using the embedded SAS controller, the drives must be formatted to 528 sectors.

2.10 External disk subsystems

This section describes the external disk subsystems that are supported IBM System Storage family of products.

For up-to-date compatibility information for Power blades and IBM Storage, go to the Storage System Interoperability Center at the following link:

http://ibm.com/systems/support/storage/config/ssic

For N Series Storage compatibility with Power blades, go to:

http://ibm.com/systems/storage/network/interophome.html

2.10.1 IBM BladeCenter S integrated storage

A key feature of the IBM BladeCenter S chassis is support for integrated storage. The BladeCenter S supports up to two storage modules. These modules provide integrated SAS storage functionality to the BladeCenter S chassis.

There are two ways to implement the integrated storage solution for BladeCenter S with the PS703 and PS704:

- Using the SAS Connectivity Module
- Using the SAS RAID Controller Module

These methods are detailed in the following sections.

Basic local storage using SAS Connectivity Module

The main feature of basic local storage is the ability to assign physical disks in disk storage modules (DSMs) to the blade server, and create volumes by using the RAID function of the on-board SAS controller on the blade itself in conjunction with the SAS Passthrough Card installed into the blade server.

Table 2-9 lists the basic local storage solution components for BladeCenter S.

Table 2-9 Basic local storage solution components for BladeCenter S

Component description	Part number	Min/max quantity
Disk Storage Module (DSM)	43W3581	1 / 2
SAS Connectivity Module	39Y9195	1 / 2
3 Gb SAS Passthrough Card (CIOv)[a]	43W4068	1 per PS703 2 per PS704
PS703 or PS704	7891	1 / 6

[a] Also known as the SAS Connectivity Card (CIOv)

Table 2-10 lists hard disk drives supported in DSMs by SAS Connectivity Modules.

Table 2-10 Hard disk drives supported in DSMs by SAS Connectivity Modules

Description	Part number	Max quantity
3.5” Hot Swap SATA		
1000 GB Dual Port Hot Swap SATA HDD	43W7630	12 (6 per one DSM)
3.5” Hot Swap NL SAS		
IBM 1 TB 7200 NL SAS 3.5'' HS HDD	42D0547	12 (6 per one DSM)
IBM 1 TB 7.2K 6 Gbps NL SAS 3.5'' HDD	42D0777	12 (6 per one DSM)
IBM 2 TB 7.2K 6 Gbps NL SAS 3.5'' HDD	42D0767	12 (6 per one DSM)
3.5” Hot Swap SAS		
IBM 300 GB 15K 6 Gbps SAS 3.5" Hot-Swap HDD	44W2234	12 (6 per one DSM)
IBM 450 GB 15K 6 Gbps SAS 3.5" Hot-Swap HDD	44W2239	12 (6 per one DSM)
IBM 600 GB 15K 6 Gbps SAS 3.5" Hot-Swap HDD	44W2244	12 (6 per one DSM)

Figure 2-29 on page 75 shows a sample connection topology for basic local storage with one SAS Connectivity Module installed.

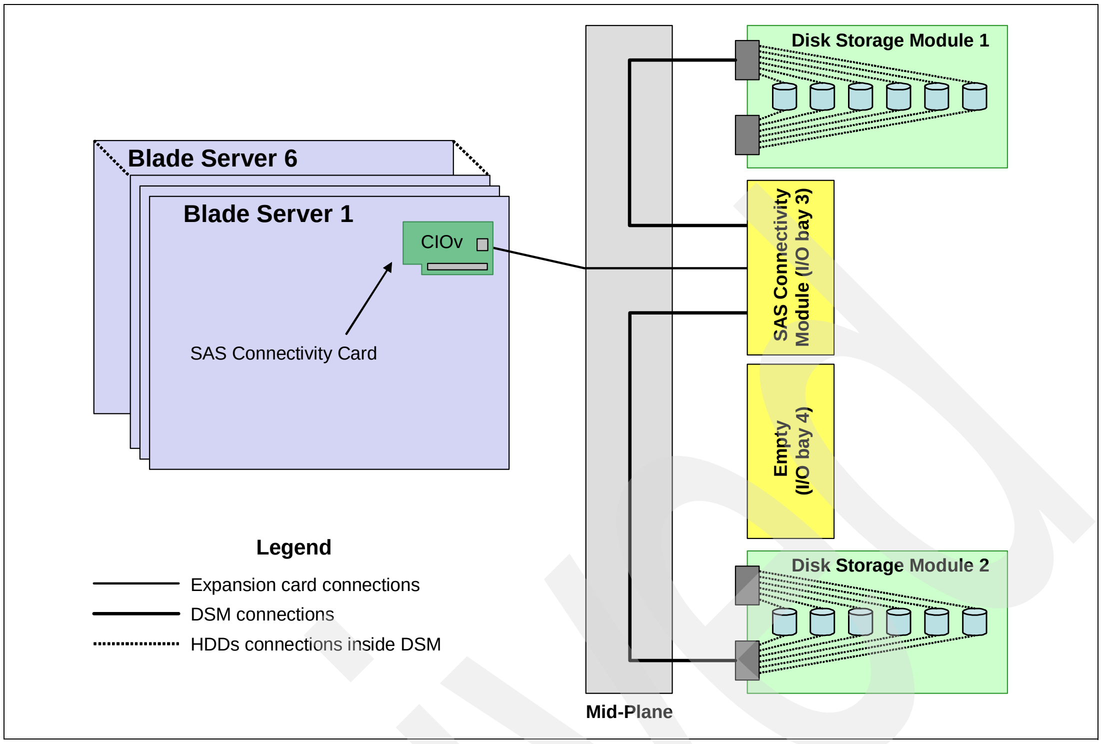


Figure 2-29 SAS I/O connections with one SAS Connectivity Module installed

Figure 2-30 shows a sample connection topology for basic local storage with two SAS Connectivity Modules installed.

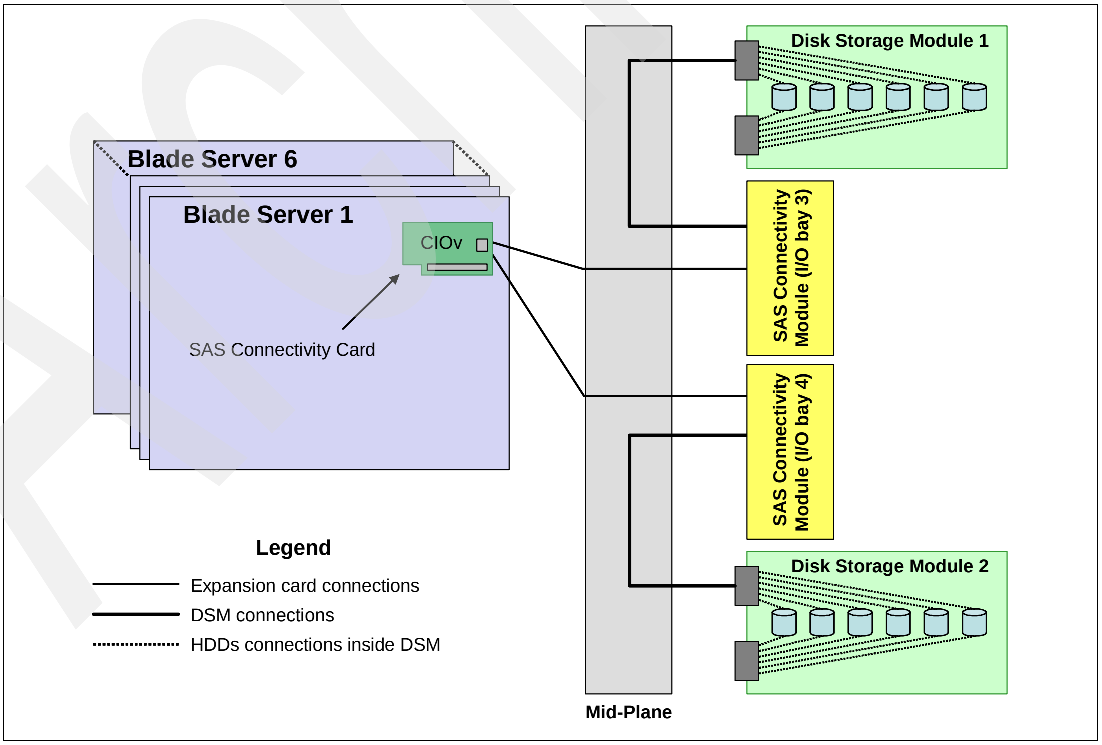


Figure 2-30 SAS I/O connections with two SAS Connectivity Modules installed

Keep the following considerations in mind when planning BladeCenter S basic local storage implementations:

- Every blade requiring integrated storage connectivity must have one SAS Connectivity Card installed.
- At least one DSM must be installed into the BladeCenter S chassis; a maximum of two DSMs are supported in one chassis. The IBM BladeCenter S does not ship with any DSMs as standard.
- At least one SAS Connectivity Module must be installed into the BladeCenter S chassis. A maximum of two SAS modules are supported for redundancy and high availability purposes.

  If two SAS connectivity modules are installed, the module in I/O module bay 3 controls access to storage module 1, and the module in I/O module bay 4 controls access to storage module 2.
- Each physical hard disk drive in any DSM can be assigned to the one blade server only, and the blade itself sees this physical disk (or disks) as its own hard disk drives connected via SAS expansion card. That is, there are no LUNs, storage partitions, and so forth as in shared storage systems. Instead, each blade has its own set of physical disks residing in DSM (or DSMs).
- The SAS Connectivity Module controls disk assignments by using zoning-like techniques, which results in the SAS module maintaining single, isolated, dedicated paths between physical HDDs and the blade server. Configuration of HDD assignments is done by the administrator, and the administrator uses predefined templates or creates custom configurations.
- The RAID functionality is supplied by the onboard SAS controller when SAS Connectivity Cards are used in the blade servers.
- The maximum number of drives supported by the IM volume is two, plus one optional global hot spare. The IME volume supports up to ten HDDs plus two optional hot spares. The IS volume supports up to ten HDDs. The IS volume does not support hot spare drives.
- When creating a RAID-1 array, we recommend that you span both disk storage modules. This maximizes the availability of data if one of the paths to the disks is lost, because there is only one connection to each disk storage module, as shown in Figure 2-30 on page 75.
- Mixing HDDs of different capacities in a single volume is supported. However, the total volume size is aligned with the size of the smallest HDD, and excess space on larger-sized HDDs is not used.

  Supported combinations of volumes include:

  – Two IM or IME volume per blade server
  – One IM or IME volume and one IS volume per blade server
  – Two IS volumes per blade server
- Each blade with an SAS expansion card has access to its assigned HDDs in both DSMs, even if only one SAS Connectivity module is present. Potentially, all 12 drives in both DSMs can be assigned to the single blade server. However, only 10 HDDs can be used in a single volume. You can create either two volumes to utilize the capacity of all drives, or designate the remaining two drives as hot spares.
- Both SAS and SATA hot swap HDDs are supported, and an intermix of SAS and SATA drives is supported, as well. However, each volume must have hard disks of the same type; that is, SAS or SATA.
- External disk storage attachments are not supported.

Advanced shared storage using the SAS RAID Controller Module

The main feature of advanced shared storage for BladeCenter S is the ability to:

- Create storage pools from hard disks in disk storage modules
- Create logical volumes in these pools
- Assign these volumes rather than physical disks to the blade servers
- Map a single logical volume to several blade servers simultaneously

Table 2-11 lists the advanced local storage components for BladeCenter S.

Table 2-11 Advanced local storage solution components for BladeCenter S

Component description	Part number	Min/max quantity
Disk Storage Module (DSM)	43W3581	1 / 2
SAS RAID Controller Module	43W3584	2 / 2
SAS Connectivity Card (CIOv)	43W4068	1 per PS703 2 per PS704
Ethernet Switch in I/O bay 1	Varies	1 / 1
PS703 or PS704	7891	1 / 6

Table 2-12 lists hard disk drives supported by SAS RAID Controller Modules.

Table 2-12 Hard disk drives supported in DSMs by SAS RAID Controller Modules

Description	Part number	Max quantity
3.5” Hot Swap NL SAS		
IBM 1 TB 7200 NL SAS 3.5'' HS HDD	42D0547	12 (6 per one DSM)
IBM 1 TB 7.2K 6 Gbps NL SAS 3.5'' HDD	42D0777	12 (6 per one DSM)
IBM 2 TB 7.2K 6 Gbps NL SAS 3.5'' HDD	42D0767	12 (6 per one DSM)
3.5” Hot Swap SAS		
IBM 300 GB 15K 6 Gbps SAS 3.5" Hot-Swap HDD	44W2234	12 (6 per one DSM)
IBM 450 GB 15K 6 Gbps SAS 3.5" Hot-Swap HDD	44W2239	12 (6 per one DSM)
IBM 600 GB 15K 6 Gbps SAS 3.5" Hot-Swap HDD	44W2244	12 (6 per one DSM)

Figure 2-31 shows a sample topology for BladeCenter S with two SAS RAID Controller Modules.

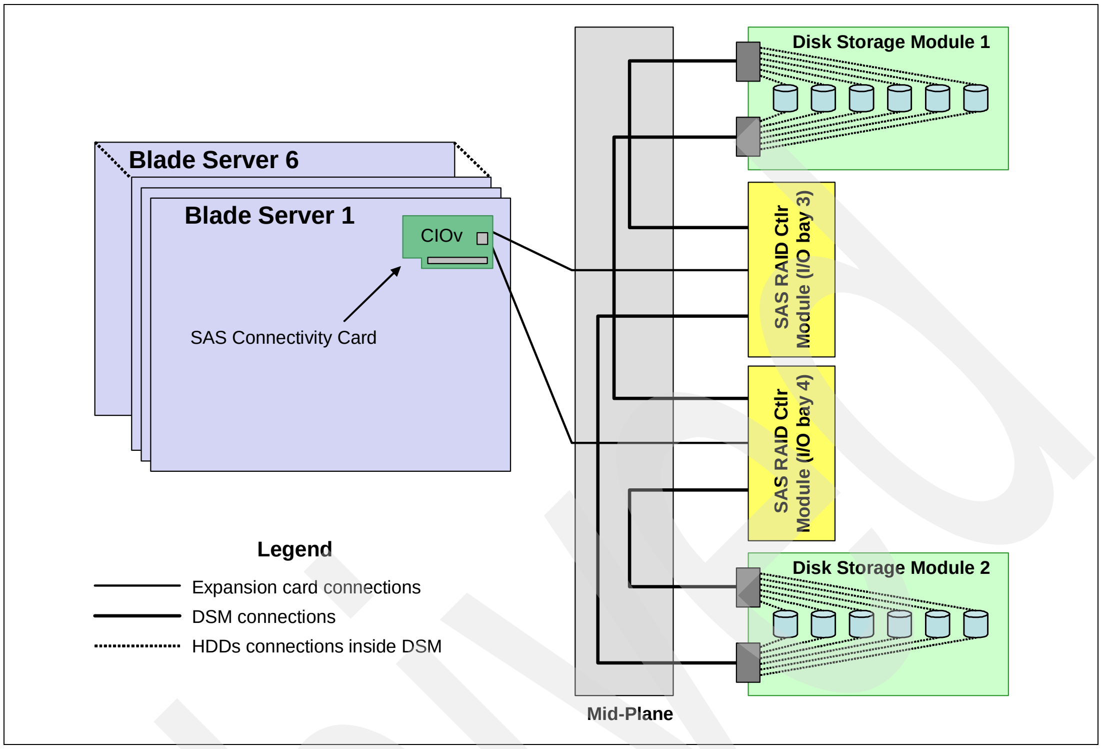


Figure 2-31   BladeCenter S SAS RAID Controller connections topology

New features: Starting from firmware release 1.2.0, SAS RAID Controller Module now supports online volume expansion, online storage pool expansion, concurrent code update, and online DSM replacement for RAID-1 and 10 configurations. Prior to this firmware level, the system must be stopped to perform capacity expansion, firmware upgrade, or DSM servicing.

Keep these considerations in mind when planning BladeCenter S advanced shared storage implementations:

- Every blade requiring integrated storage connectivity must have one SAS Expansion Card installed.
- At least one DSM must be installed into the BladeCenter S chassis; a maximum of two DSMs are supported in one chassis. The IBM BladeCenter S does not ship with any DSMs as standard.
- Two SAS RAID Controller Modules installed into the BladeCenter S chassis.
- The SAS RAID Controller Module creates storage pools (or arrays), and the RAID level is defined for these storage pools. Logical volumes are created from storage pools. Volumes can be assigned to a specific blade, or can be shared by several blade servers.
- Zoning is supported by the SAS RAID controller module. However, zoning should not be used for regular operations (in other words, for purposes other than troubleshooting).
- RAID functionality is supplied by the SAS RAID Controller Modules installed into the BladeCenter S chassis.
  - RAID levels supported: 0, 5, 6, 10.
  - Maximum volume size is limited by size of storage pool.

  - Maximum number of volumes is 16 per blade server (maximum of 128 volumes per chassis).
  - One volume can be mapped to all 6 blades in the chassis.
- Mixing HDDs of different capacities in a single volume is supported. However, the total volume size is aligned with the size of the smallest HDD, and excess space on larger-sized HDDs is not used.
- Both SAS and Near-line SAS (NL SAS) hot swap HDDs are supported, and intermixing SAS/NL SAS drives is supported as well. However, each storage pool must have hard disks of the same type; that is, SAS or NL SAS. SATA drives are not supported by SAS RAID Controller Module.
- Global hot-spare drives are supported. The drive designated as a hot-spare should be as large as, or larger than, other drives in the system.
- Blade boot from logical volume is supported.
- Path failover is supported with IBM Subsystem Device Driver Device Specific Module (SDD DSM) for Windows® and Device Mapper Multipath (DMM) for Red Hat/Novell SUSE Linux.
- External tape attachments are supported.

For more information about the SAS RAID Controller solution, see:

- *SAS RAID Controller Installation and User Guide*:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5078040
- *SAS RAID Controller Module Interoperability Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5078491
- *SAS RAID Controller Module Detailed Host Attachment Guide* (Remote Boot included):

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5078491

2.10.2 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, management, and system efficiency at a reduced total cost of ownership. Several enhancements have been incorporated into the N series product line to complement and reinvigorate this portfolio of solutions. The enhancements include:

- The new SnapManager® for Hyper-V provides extensive management for backup, restoration, and replication for Microsoft® Hyper-V environments.
- The new N series Software Packs provides the benefits of a broad set of N series solutions at a 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 page:

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 (such as 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 page:

http://www.ibm.com/systems/storage/disk/ds3000/

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 the following features:

- 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 page:

http://www.ibm.com/systems/storage/disk/ds5020/

IBM System Storage DS5000

New DS5000 enhancements help reduce costs by reducing power per performance by introducing SSD drives. Also, with the new EXP5060 expansion unit supporting 60 1-TB SATA drives in a 4 U 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 providing 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 costs.

For more information, see the following web page:

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

For more information, see the following web page:

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

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 need information to be reliably available around the clock and with minimal delay, the DS8000 series can be an ideal solution. 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 needed for the most demanding business applications.

For more information, see the following web page:

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

IBM Storwize V7000 Midrange Disk System

IBM Storwize V7000 is a virtualized storage system to complement virtualized server environments that provides unmatched performance, availability, advanced functions, and highly scalable capacity never seen before in midrange disk systems. IBM Storwize V7000 is a powerful midrange disk system that is easy to use and enables rapid deployment without additional resources. Storwize V7000 is virtual storage that offers greater efficiency and flexibility through built-in SSD optimization and "thin provisioning" technologies. Storwize V7000 advanced functions also enable non-disruptive migration of data from existing storage, simplifying implementation, and minimizing disruption to users. IBM Storwize V7000 also enables you to virtualize and reuse existing disk systems, supporting a greater potential return on investment.

For more information, see the following web page:

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

2.11 IVM

IVM is a simplified hardware management solution that is part of the PowerVM implementation on the PS703 and PS704 blades. POWER processor-based blades do not include an option for attachment to a Hardware Management Console (HMC). POWER processor-based blades do, however, include an option for attachment to an IBM Systems Director Management Console (SDMC).

IVM inherits most of the HMC features and capabilities, and enables the exploitation of PowerVM technology. It manages a single server, avoiding the need for an independent appliance. It provides a solution that enables the administrator to reduce system setup time and to make hardware management easier, at a lower cost.

IVM is an addition to the Virtual I/O Server, 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. See Table 2-13.

For a complete description of the possibilities offered by IVM, see *Integrated Virtualization Manager on IBM System p5*, REDP-4061, available the following web page:

http://www.redbooks.ibm.com/abstracts/redp4061.html

Table 2-13 Comparison of IVM, HMC, and SDMC

Characteristic	IVM	HMC	SDMC
General characteristics			
Delivery vehicle	Integrated into the server	A desktop or rack-mounted appliance	Hardware/Software Appliance
Footprint	Runs in 60 MB memory and requires minimal CPU as it runs stateless.	2-Core x86, 2 GB RAM, 80 GB HD	4-core CPU (Intel Nehalem or better), 8 GB RAM, 500 GB HD, 2 to 4 network interface cards
Installation	Installed with the Virtual I/O Server (optical or network). Preinstall option available on certain systems.	Appliance is preinstalled. Reinstall through optical media or network is supported.	Hardware appliance is preinstalled. Software appliance requires x86 to be installed on.
Multiple system support	One IVM per server	One HMC can manage multiple (non-blade) servers (48 systems/ 1024 LPARS)	One SDMC can manage multiple servers (48 systems/1024 virtual servers)
High-end servers	Not supported	Supported	Supported (hardware appliance only)
Low end & midrange servers	Supported	Supported	Supported
Blade servers	Supported	Not Supported	Supported
Server Families Supported	P5, P5+: Yes P6, P6+: Yes P7: Yes	P5, P5+: Yes P6, P6+: Yes P7: Yes	P5, P5+: No P6, P6+: Yes P7: Yes
User interface	Web browser (no local graphical display) and telnet session	Web browser (local or remote)	Web browser (local or remote)
Scripting and automation	VIOS command-line interface (CLI) and HMC compatible CLI.	HMC CLI	SMCLI
RAS characteristics			
Redundancy and HA of manager	Only one IVM per server	Multiple HMCs can manage the same system for HMC redundancy (active/active).	Multiple SDMCs can manage the same system (active/backup).
Multiple VIOS	No, single VIOS	Yes	Yes
Fix or update process for manager	VIOS fixes and updates	HMC e-fixes and release updates	Update Manager
Adapter microcode updates	Inventory scout through RMC	Inventory scout through RMC	Update Manager

Characteristic	IVM	HMC	SDMC
Firmware updates	Inband through OS; not concurrent	Service Focal Point™ with concurrent firmware updates	Update Manager, concurrent firmware updates
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	Service and Support Manager for consolidated management of operating system- and firmware-detected errors
PowerVM function			
Full PowerVM Capability	Partial	Full	Full
Capacity on Demand	Entry of PowerVM codes only	Full Support	Full Support
I/O Support for IBM i	Virtual Only	Virtual and Direct	Virtual and Direct
Multiple Shared Processor Pool	No, default pool only	Yes	Yes
Workload Management (WLM) Groups Supported	One	254	254
Support for multiple profiles per partition	No	Yes	Yes
SysPlan Deploy & mksysplan	Limited no POWER7 support, no Deploy on blades	Yes	Not in initial release, no blade support

2.12 Operating system support

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

- AIX
- IBM i
- Linux

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

> **Note:** For details about the software available on IBM POWER servers, see Power Systems Software™ at the following web page:
>
> http://www.ibm.com/systems/power/software/

The PS703 and PS704 blades support the operating system versions identified in this section.

Virtual I/O Server

Virtual I/O Server 2.2.0.12-FP24 SP02 or later

IBM regularly updates the Virtual I/O Server code. To find information about the latest updates, see the Virtual I/O Server at the following web page:

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

IBM AIX Version 5.3

AIX Version 5.3 with the 5300-12 Technology Level with Service Pack 4 or later.

AIX Version 5.3 with the 5300-11 Technology Level with 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 AIX 5L operating system. Information about these packages, downloading, and obtaining the CD-ROM is on the Fix Central web page:

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

The Service Update Management Assistant can help you to automate the task of checking and downloading operating system downloads, and is part of the base operating system. For more information about the **suma** command functionality, go to the following web page:

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

AIX Version 6.1

AIX 6.1 with the 6100-04 Technology Level with Service Pack 10 or later

AIX 6.1 with the 6100-05 Technology Level with Service Pack 6 or later

AIX 6.1 with the 6100-06 Technology Level with Service Pack 5 or later

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

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

AIX Version 7.1

AIX 7.1 with Service Pack 3 or later

For information regarding AIX V7.1 maintenance and support, go to the Fix Central web page:

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

IBM i

Virtual I/O Server is required to install IBM i in a LPAR on PS703 and PS704 blades and all I/O must be virtualized.

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

For a detailed guide on installing and operating IBM i with Power Blades, see the following web page:

http://ibm.com/systems/resources/systems_power_hardware_blades_i_on_blade_readme.pdf

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.

At the time of this writing, the supported versions of Linux on POWER7 processor technology based servers are as follows:

- SUSE Linux Enterprise Server 11 SP1 for POWER or later, with current maintenance updates available from Novell to enable all planned functionality
- Red Hat RHEL 5.6 for POWER, or later
- Red Hat RHEL 6.0 for POWER, or later

Linux operating system licenses are ordered separately from the hardware. You can obtain Linux operating system licenses from IBM, to be included with your POWER7 processor technology-based servers, or from other Linux distributors.

> **Note:** For systems ordered with the Linux operating system, IBM ships the most current version available from the distributor. If you require a different version than that shipped by IBM, you must obtain it via download from the Linux distributor's website. Information concerning access to a distributor's website is located on the product registration card delivered to you as part of your Linux operating system order.

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

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

For information about SUSE Linux Enterprise Server, go to the following web page:

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

For information about Red Hat Enterprise Linux Advanced Server, go to the following web page:

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

Supported virtualization features are listed in 3.4.10, “Supported PowerVM features by operating system” on page 117.

> **Note:** Users should also update their systems with the latest Linux for Power service and productivity tools from IBM's website at
>
> http://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/home.html

2.13 IBM EnergyScale

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.

The BladeCenter AMM and IBM Systems Director Active Energy Manager exploit EnergyScale technology, enabling advanced energy management features to conserve power and 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 reduce frequency to save energy. This feature is called Turbo-Mode.

Tip: Turbo-Mode, discussed here, and TurboCore mode, discussed in “TurboCore mode” on page 46 are two different technologies.

2.13.1 IBM EnergyScale technology

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

IBM EnergyScale consists of the following elements:

- A built-in EnergyScale device (formally known as Thermal Power Management Device or TPMD).
- Power executive software. IBM Systems Director Active Energy Manager, an IBM Systems Directors plug-in and BladeCenter AMM.

IBM EnergyScale functions include the following elements:

- 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 might use such information to predict data center energy consumption at various times of the day, week, or month.

- Thermal reporting

  IBM Systems 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 the processor frequency and voltage by a fixed amount, reducing the energy consumption of the system and 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% from nominal. Power Saver Mode is not supported during boot or reboot 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 use of the POWER7 processors. The user must configure this setting from the BladeCenter AMM or IBM Director Active Energy Manager. Processor frequency and use are inversely proportional for most workloads, implying that as the frequency of a processor increases, its use 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 use. When a system is idle, the system firmware lowers the frequency and voltage to Power Saver Mode values. When fully used, 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-used, 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 can 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 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 BladeCenter AMM or IBM Systems Director Active Energy Manager user interface.

- 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.

- 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.

In addition to the normal EnergyScale functions, the EnergyScale device in the PS703 and PS704 blades incorporates the following BladeCenter functions:

- Transition from over-subscribed power consumption to nominal power consumption when commanded by the BladeCenter AMM. This transition is signaled by the AMM as a result of a redundant power supply failure in the BladeCenter.

- Report blade power consumption to the AMM through the service processor.
- Report blade system voltage levels to the AMM through the service processor.
- Accommodate BladeCenter/AMM defined thermal triggers such as warning temperature, throttle temperature, and critical temperature.

2.13.2 EnergyScale device

The EnergyScale device 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 cores versus two cores, four threads versus two threads per core, and asynchronous processor core chipset
- Advanced Idle Power Management functions
- Advanced Dynamic Power Management (DPM) functions in all units in hardware (processor cores, processor core chiplet, chip-level nest unit level, and chip level)
- Advanced Actuators/Control
- Advanced Accelerators

The new EnergyScale device 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.

3

Virtualization

IBM Advance POWER Virtualization (PowerVM) is a feature use to consolidate workload to deliver cost savings and improve infrastructure responsiveness. As you look for ways to maximize the return on your IT infrastructure investments, consolidating workloads and increasing server use becomes an attractive proposition.

IBM Power Systems, combined with PowerVM technology, are designed to help you consolidate and simplify your IT environment. The following list details key capabilities:

- Improve server use by consolidating diverse sets of applications.
- Share CPU, memory, and I/O resources to reduce total cost of ownership.
- Improve business responsiveness and operational speed by dynamically re-allocating resources to applications as needed, to better anticipate changing business needs.
- Simplify IT infrastructure management by making workloads independent of hardware resources, enabling you to make business-driven policies to deliver resources based on time, cost, and service-level requirements.
- Move running workloads between servers to maximize availability and avoid planned downtime

This chapter discusses the following virtualization technologies and features on IBM POWER7 processor-based blade servers:

3.1 PowerVM Version 2.2 enhancements

The latest available PowerVM Version 2.2 contains the following enhancements:

- Support for up to 160 LPARs on PS703
- Support for up to 320 LPARs on PS704
- Support for up to 80 LPARs on Power 710 and 720
- Support for up to 160 LPARs on Power 730 and 740
- Support for up to 320 LPARs on Power750
- Support for up to 640 LPARs on 770 and 780
- Support for up to 1000 LPARs on Power 795
- PowerVM support for sub-chip per-core licensing on Power 710, 720, 730, and 740
- Role Based Access Control (RBAC)

  RBAC brings an added level of security and flexibility in the administration of VIOS. With RBAC, you can create a set of authorizations for the user management commands. You can assign these authorizations to a role UserManagement and this role can be given to any other user. So a normal user with the role UserManagement can manage the users on the system but will not have any further access.

  With RBAC, the Virtual I/O Server has the capability of splitting management functions that presently can be done only by the *padmin* user, providing better security by giving only the necessary access to users, and easy management and auditing of system functions.

- Suspend/Resume

  Using Suspend/Resume, clients can provide long-term suspension (greater than 5-10 seconds) of partitions, saving partition state (memory, NVRAM, and VSP state) on persistent storage, freeing server resources that were in use by that partition, restoring partition state to server resources, and resuming operation of that partition and its applications either on the same server or on a different server.

  Requirements for Suspend/Resume: All resources must be virtualized prior to suspending a partition. If the partition is to be resumed on a different server, then the shared external I/O (disk and LAN) should remain identical. Suspend/Resume works with AIX and Linux workloads when managed by HMC.

- Shared storage pools

  VIOS 2.2 allows the creation of storage pools that can be accessed by VIOS partitions deployed across multiple Power Systems servers so that an assigned allocation of storage capacity can be efficiently managed and shared.

- Thin provisioning

  VIOS 2.2 supports highly efficient storage provisioning, whereby virtualized workloads in VMs can have storage resources from a shared storage pool dynamically added or released as required.

- VIOS grouping

  Multiple VIOS 2.2 partitions can utilize a common shared storage pool to more efficiently utilize limited storage resources and simplify the management and integration of storage subsystems.

The IBM PowerVM Workload Partitions Manager™ for AIX, Version 2.2 has the following enhancements:

- When used with AIX 6.1 Technology Level 6, the following support applies:
  - Support for exporting VIOS SCSI disk into a WPAR. Compatibility analysis and mobility of WPARs with VIOS SCSI disk. In addition to Fibre Channel devices, now VIOS SCSI disks can be exported into a workload partition (WPAR).
  - WPAR Manager Command-Line Interface (CLI). The WPAR Manager CLI allows federated management of WPARs across multiple systems by command line.
  - Support for workload partition definitions. The WPAR definitions can be preserved after WPARs are deleted. These definitions can be deployed at a later time to any WPAR-capable system.
- In addition to the feature supported on AIX 6.1 Technology Level 6, the following applies to AIX 7.1:
  - Support for AIX 5.2 Workload Partitions for AIX 7.1. Lifecycle management and mobility enablement for AIX 5.2 Technology Level 10 SP8 Version WPARs.
  - Support for trusted kernel extension loading and configuration from WPARs. Enables exporting a list of kernel extensions that can then be loaded inside a WPAR, yet maintaining isolation.

3.2 POWER Hypervisor

Combined with features designed into the POWER7 processors, the POWER Hypervisor delivers functions that enable capabilities, including dedicated processor partitioning, micro-partitioning, virtual processors, IEEE VLAN compatible virtual switch, virtual SCSI adapters, virtual Fibre Channel adapters, and virtual consoles.

The user interface into the POWER Hypervisor on POWER-based blades has traditionally been through the Integrated Virtualization Manager or IVM. Now a second method of systems management is available, the Systems Director Management Console or SDMC. The SDMC brings additional capabilities to POWER6- and POWER7-based blades that brings them closer to traditional rack-based POWER platforms. Additional details can be found in Chapter 5, “Systems Director Management Console” on page 151.

Terminology note: The SDMC, being IBM Systems Director based, uses the term *virtual server*. This term is used interchangeably with *logical partitions* or *LPARs*

The POWER Hypervisor technology is integrated with all IBM POWER servers, including the POWER7 processor-based blade servers. The hypervisor orchestrates and manages system virtualization, including creating logical partitions and dynamically moving resources across multiple operating environments. The POWER Hypervisor is a basic component of the system firmware that is layered between the hardware and operating system. POWER Hypervisor offers the following functions:

- Provides an abstraction layer between the physical hardware resources and the logical partitions using them
- Enforces partition integrity by providing a security layer between logical partitions
- Controls the dispatch of virtual processors to physical processors and 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 Ethernet switches between logical partitions that help to reduce the need for physical Ethernet adapters for interpartition 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
- Uses micro-partitioning to allow multiple instances of the operating system to run on POWER5, POWER6, and POWER7 processor-based servers and POWER6 and POWER7 processor-based blades

The POWER Hypervisor is always installed and activated, regardless of system configuration. The POWER Hypervisor does not own any physical I/O devices; all physical I/O devices in the system are owned by logical partitions or the Virtual I/O Server.

Memory is required to support the resource assignment to the logical partitions on the server. The amount of memory required by the POWER Hypervisor firmware varies according to several factors. The following factors influence POWER Hypervisor memory requirements:

- Number of logical partitions
- Number of physical and virtual I/O devices used by the logical partitions
- Maximum memory values specified in the logical partition profiles

The minimum amount of physical memory 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 system, as shown in Table 3-1 on page 93.

Table 3-1 Configured memory-to-default LMB size

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

Physical memory assigned to partitions are 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 I/O adapters are defined by system administrators during logical partition definition. Configuration information for the adapters is presented to the partition operating system.

Virtual SCSI

The POWER Hypervisor provides a virtual SCSI mechanism for virtualization of storage devices. Virtual SCSI allows secure communications between a logical partition and the IO Server (VIOS). The storage virtualization is accomplished by pairing two adapters: a virtual SCSI server adapter on the VIOS, and a virtual SCSI client adapter on IBM i, Linux, or AIX partitions. The combination of Virtual SCSI and VIOS provides the opportunity to share physical disk adapters in a flexible and reliable manner.

Virtual Ethernet

The POWER Hypervisor provides an IEEE 802.1Q VLAN-style virtual Ethernet switch that allows partitions on the same server to use fast and secure communication without any need for physical connection.

Virtual Ethernet support starts with AIX Version 5.3, or the appropriate level of Linux supporting virtual Ethernet devices (see 3.4.10, “Supported PowerVM features by operating system” on page 117). 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 up to 65408 bytes in size. Therefore, the maximum MTU for the corresponding interface can be up to 65394 (65408 minus 14 for the header) in the non-VLAN case and to 65390 (65408 minus 14, minus 4) 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).
- An AIX partition supports 256 virtual Ethernet adapters for each logical partition. 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 operating system partition 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 configured in a VIOS partition (see 3.4.4, “VIOS” on page 108 for more details about shared Ethernet). The device configured is know as a Shared Ethernet Adapter or SEA.

> **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 for inter-partition communication.

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 VIOS logical partition. The VIOS logical partition provides the connection between the virtual Fibre Channel adapters on the VIOS logical partition and the physical Fibre Channel adapters on the managed system.

NPIV is a standard technology for Fibre Channel networks. It enables you to connect multiple logical partitions to one physical port of a physical Fibre Channel adapter. Each logical partition is identified by a unique WWPN, which means that you can connect each logical partition to independent physical storage on a SAN.

> **Note:** To enable NPIV on a managed system, VIOS at version 2.1 or later is required. NPIV is only supported on 8 Gb Fibre Channel and Converged Network (FCoE) adapters on POWER-based systems.

You can only configure virtual Fibre Channel adapters on client logical partitions that run the following operating systems:

- AIX 6.1 Technology Level 2, or later
- AIX 5.3 Technology Level 9, or later
- IBM i version 6.1.1, 7.1, or later
- SUSE Linux Enterprise Server 11, or later
- RHEL 5.5, 6, or later

For details on which expansion cards support NPIV see 3.4.9, “N_Port ID Virtualization (NPIV)” on page 116.

Systems that are managed by the Integrated Virtualization Manager (IVM) or a Systems Director Management Console (SDMC) can dynamically add and remove virtual Fibre Channel adapters from logical partitions. Figure 3-1 on page 95 depicts the connections between the client partition virtual Fibre Channel adapters and the external storage.

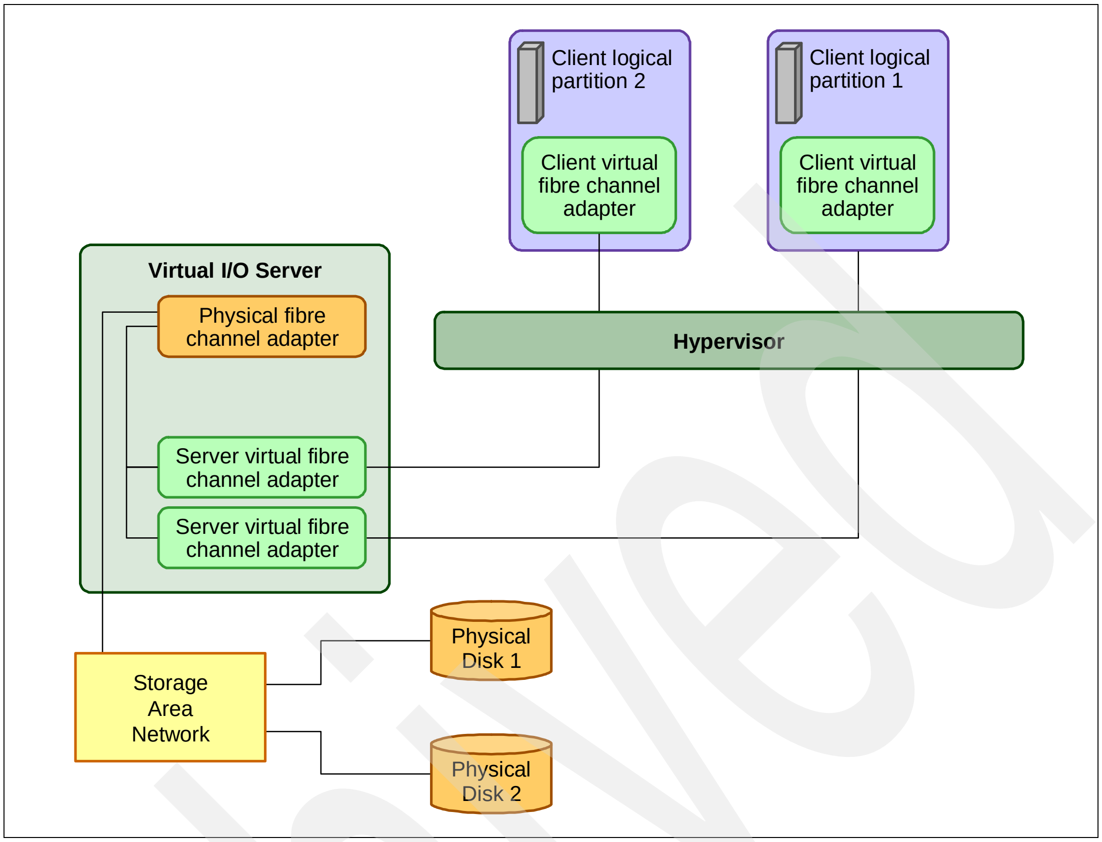

Client logical partition 2
Client virtual fibre channel adapter
Client logical partition 1
Client virtual fibre channel adapter
Virtual I/O Server
Physical fibre channel adapter
Hypervisor
Server virtual fibre channel adapter
Server virtual fibre channel adapter
Physical Disk 1
Storage Area Network
Physical Disk 2

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

Virtual serial adapters (TTY) console

Virtual serial adapters provide a point-to-point connection from one logical partition to another, the Hardware Management Console (HMC), or the Systems Director Management Console (SDMC) to each logical partition on the managed system. Virtual serial adapters are used primarily to establish terminal or console connections to logical partitions.

Each partition needs to have access to a system console. Tasks such as operating system installation, network setup, and certain 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 it. Virtual TTY does not require the purchase of any additional features or software such as the PowerVM Edition features.

For POWER6- or POWER7-based blades, the operating system console can be accessed from the IVM user interface or SDMC depending on which method is being used to manage the system, or from Serial Over LAN (SOL) through the BladeCenter Advanced Management Module (AMM) for the LPAR or virtual server that owns the first Ethernet port. Not all methods can be used concurrently.

3.3 POWER processor modes

Although, strictly speaking, they are 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:

  http://www.power.org/resources/reading/PowerISA_V2.05.pdf

- POWER6+ compatibility mode

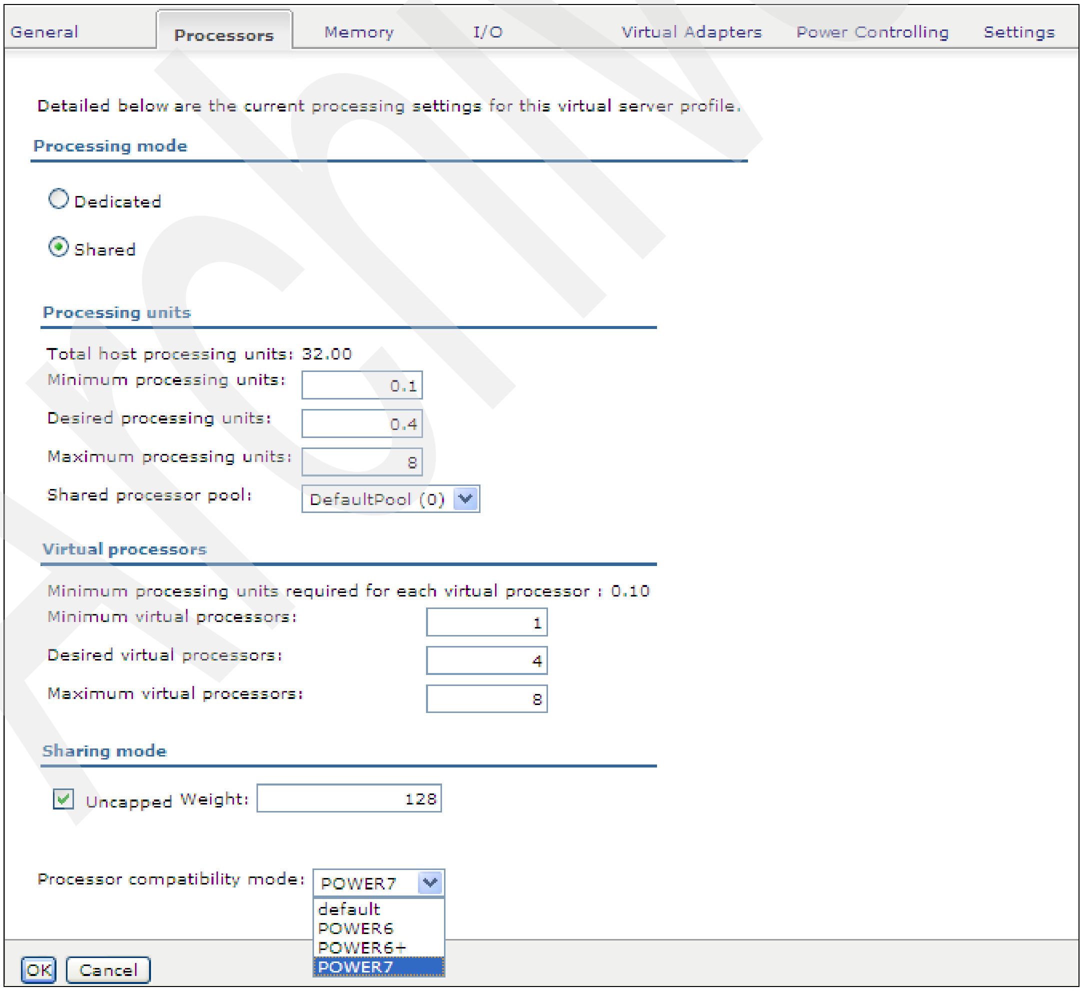
  This mode is similar to POWER6, with 8 additional Storage Protection Keys.

- POWER7 mode

  This is the native mode for POWER7 processors, implementing v2.06 of the Power Instruction Set Architecture. For more information, see:

  http://www.power.org/resources/downloads/PowerISA_V2.06_PUBLIC.pdf

Figure 3-2 shows how the SDMC displays available processor compatibility modes.



Detailed below are the current processing settings for this virtual server profile.

Processing mode

○ Dedicated

● Shared

Processing units

Total host processing units: 32.00

Minimum processing units: 0.1

Desired processing units: 0.4

Maximum processing units: 8

Shared processor pool: DefaultPool (0)

Virtual processors

Minimum processing units required for each virtual processor : 0.10

Minimum virtual processors: 1

Desired virtual processors: 4

Maximum virtual processors: 8

Sharing mode

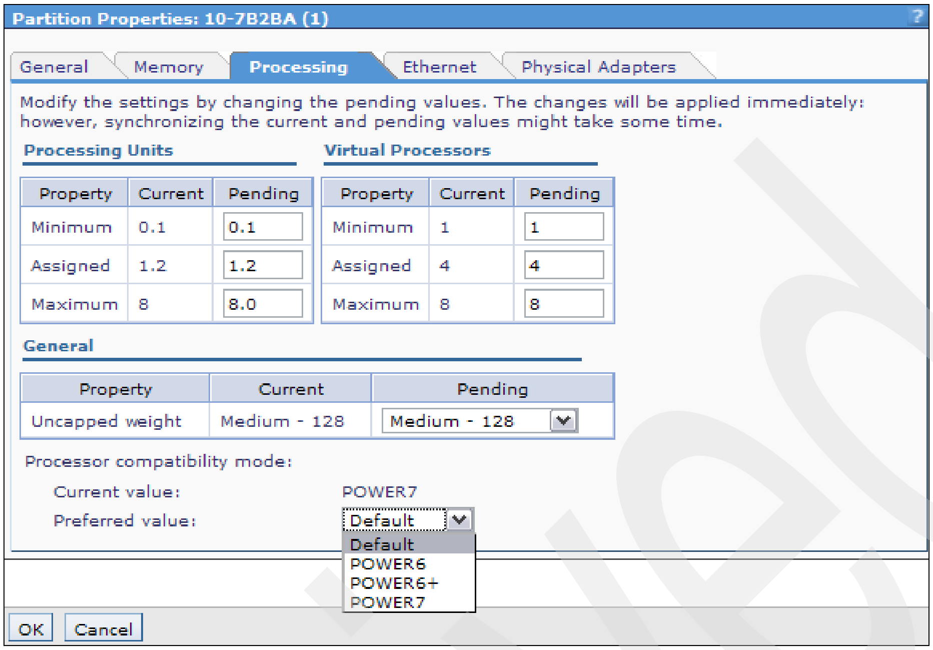
☑ Uncapped Weight: 128

Processor compatibility mode: POWER7

default
POWER6
POWER6+
POWER7

Figure 3-2 Selecting processor compatibility mode from SDMC

Figure 3-3 shows how to choose a processor compatibility mode from IVM.



Modify the settings by changing the pending values. The changes will be applied immediately: however, synchronizing the current and pending values might take some time.

Processing Units

Property	Current	Pending
Minimum	0.1	0.1
Assigned	1.2	1.2
Maximum	8	8.0

Virtual Processors

Property	Current	Pending
Minimum	1	1
Assigned	4	4
Maximum	8	8

General

Property	Current	Pending
Uncapped weight	Medium - 128	Medium - 128

Processor compatibility mode:

Current value: POWER7

Preferred value: Default

- Default
- POWER6
- POWER6+
- POWER7

Figure 3-3 Configuring partition profile compatibility mode from IVM

Table 3-2 lists the differences between these modes.

Table 3-2 Differences between POWER6 and POWER7 mode

POWER6 and POWER6+ mode	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 for graphic and scientific workload.
Affinity OFF by default	3-tier memory, micro-partition Affinity	Improved system performance for system images spanning sockets and nodes.
► Barrier Synchronization ► Fixed 128-byte Array; Kernel Extension Access	► Enhanced Barrier Synchronization ► Variable Sized Array; User Shared Memory Access	High performance computing parallel programming synchronization facility.
64-core and 128-thread scaling	► 32-core and 128-thread scaling ► 64-core and 256-thread scaling ► 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.

3.4 PowerVM

The PowerVM platform is the family of technologies, capabilities, and offerings that deliver industry-leading virtualization on the IBM Power Systems. Leading technologies such as Logical Partitioning, Micro-Partitioning™, POWER Hypervisor, Virtual I/O Server (VIOS), PowerVM Lx86, Live Partition Mobility (LPM), Active Memory Sharing (AMS), Suspend/Resume, and N_Port ID Virtualization (NPIV) are included within the PowerVM family. As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and value-added software.

3.4.1 PowerVM editions

This section provides information about the PowerVM editions on POWER7 processor-based blade servers.

- PowerVM Express Edition

  This edition is intended for evaluations, pilots, and proofs of concepts, generally in single-server projects. This edition supports up to three partitions per system (VIOS, AIX, Linux, and IBM i) that share processors and I/O. The Express Edition allows users to try out the Integrated Virtualization Manager (IVM) and VIOS. The Express Edition also allows attachment to an SDMC.

- PowerVM Standard Edition

  This edition is intended for production deployments, and server consolidation. This edition makes the POWER7 systems an ideal platform for consolidation of AIX, Linux, and IBM i operating system applications, helping clients reduce infrastructure complexity and cost.

- PowerVM Enterprise Edition

  The Enterprise edition is suitable for large server deployments such as multi-server deployments and cloud infrastructure. This edition includes all the features of PowerVM Standard Edition plus Live Partition Mobility and Active Memory Sharing.

Note: PowerVM Express Edition, PowerVM Standard Edition, and PowerVM Enterprise Edition are optional when running AIX or Linux. PowerVM Express Edition, PowerVM Standard Edition or PowerVM Enterprise Edition is required when running the IBM i operating system on the POWER7 PS703 and PS704 Blade Servers.

Table 3-3 lists the PowerVM editions available on each model of POWER7 processor-based blade servers and their feature codes.

Table 3-3 PowerVM Edition and feature codes

Blade Servers	Power VM Express	Power VM Standard	PowerVM Enterprise
PS703	#5225	#5227	#5228
PS704	#5225	#5227	#5228

Note: It is possible to upgrade from the Express Edition to the Standard or Enterprise Edition, and from the Standard to the Enterprise Edition.

Table 3-4 on page 99 lists the offerings of the three PowerVM editions for POWER7 blades.

Table 3-4 PowerVM capabilities by edition for POWER7-based blades

PowerVM Offerings	Express	Standard	Enterprise
Maximum VMs	Up to 3 per server	10 per core	10 per core
Management	VMControl, IVM, SDMC	VMControl, IVM, SDMC	VMControl, IVM, SDMC
Virtual I/O Server	Yes	Yes (Dual)	Yes (Dual)
PowerVM Lx86	Yes	Yes	Yes
Suspend/Resume	No	Yes	Yes
NPIV	Yes	Yes	Yes
Shared Processor Pools	No	Yes	Yes
Shared Storage Pools	No	Yes	Yes
Thin Provisioning	No	Yes	Yes
Active Memory Sharing	No	No	Yes
Live Partition Mobility	No	No	Yes

The PowerVM Editions Web site also contains useful information:

http://www.ibm.com/systems/power/software/virtualization/editions

3.4.2 Logical partitions

Logical partitions (LPARs) and virtualization increase use of system resources and add a new level of configuration possibilities. This section provides details and configuration specifications. A logical partition can be regarded as a virtual server, capable of starting an operating system and running a workload. SDMC-managed systems use the term virtual server.

Dynamic 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 multiple logical partitions, allowing each logical partition 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 as they are executing. IBM AIX Version 5.2, with necessary enhancements to enable dynamic LPAR, was introduced in 2002. The ability to reconfigure dynamic LPARs encourages system administrators to redefine available system resources dynamically to reach the optimum capacity for each defined dynamic LPAR.

Micro-partitioning

Virtualization of physical processors in POWER5, POWER6, and POWER7 systems introduces an abstraction layer that is implemented in POWER Hypervisor. Micro-partitioning is the ability to distribute the processing capacity of one or more physical processors among one or more logical partitions. Thus, processors are shared among logical partitions. Micro-partitioning technology allows you to allocate fractions of processors to a logical partition.

The POWER Hypervisor abstracts the physical processors and presents a set of virtual processors to the operating system within the micro-partitions on the system. The operating system sees only the virtual processors and dispatches runable tasks to them in the normal course of running a workload.

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*.

When defining a shared processor partition, several options must be defined:

- Processing units

  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.

- Cap or Uncap partition

  Select whether or not the partition can access extra processing power to “fill up” its virtual processors beyond its capacity entitlement, selecting either to cap or uncap your partition. If spare processing power is available in the 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. A virtual processor is a depiction or a representation of a physical processor that is presented to the operating system running in a micro-partition.

The POWER Hypervisor calculates a partition’s processing power based on minimum, desired, and maximum values, processing mode and on other active partitions’ requirements. 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 can be 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.

Partitioning maximums on the POWER7-based blades are as follows:

- The PS703 can have 16 dedicated partitions or up to 160 micro-partitions
- The PS704 can have 32 dedicated partitions or up to 320 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.

The following list details additional information about virtual processors:

- A virtual processor can be running (dispatched) either on a physical processor or as standby waiting for a physical processor to became available.

- Virtual processors do not introduce any additional abstraction level. They are only a dispatch entity. On a physical processor, virtual processors run at the same speed as the physical processor.
- Each partition's profile defines CPU entitlement, which 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 the pool.
- The number of virtual processors can be changed dynamically through a dynamic LPAR operation.

Processor 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 defines the processing mode of the logical partition.

Figure 3-4 shows a diagram of the concepts discussed in the remaining sections.

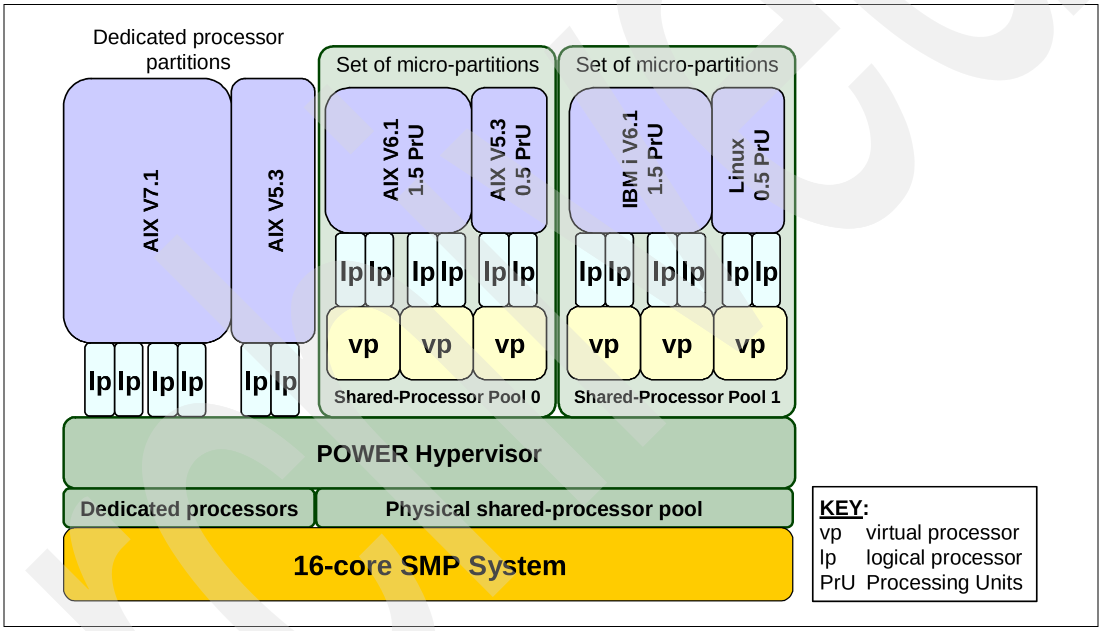


Figure 3-4 Concepts on dedicated and shared processor modes

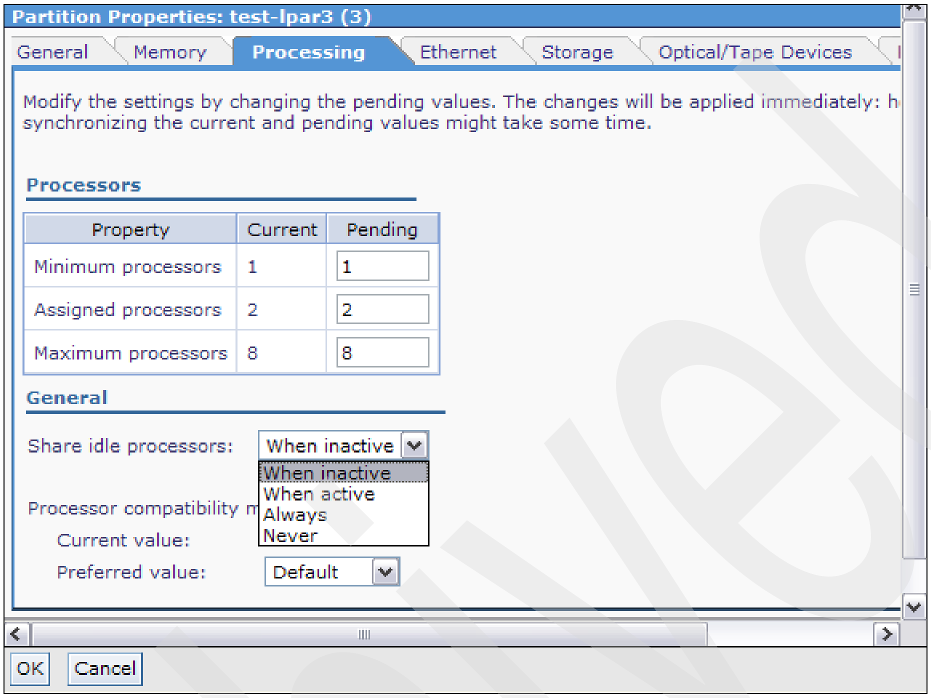
Dedicated mode

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 as the operating system is executing (for AIX, use the **smtctl** command, and for Linux, use the **ppc64_cpu** 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-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 use without compromising the computing power for critical workloads in a dedicated processor.

Figure 3-5 shows how the dedicated shared processor mode can be configured.



Modify the settings by changing the pending values. The changes will be applied immediately: h
synchronizing the current and pending values might take some time.

Processors

Property	Current	Pending
Minimum processors	1	1
Assigned processors	2	2
Maximum processors	8	8

General

Share idle processors: When inactive
When inactive
When active
Always
Never

Processor compatibility m
Current value:
Preferred value: Default

Figure 3-5 IVM console shows how to configure dedicated shared processor

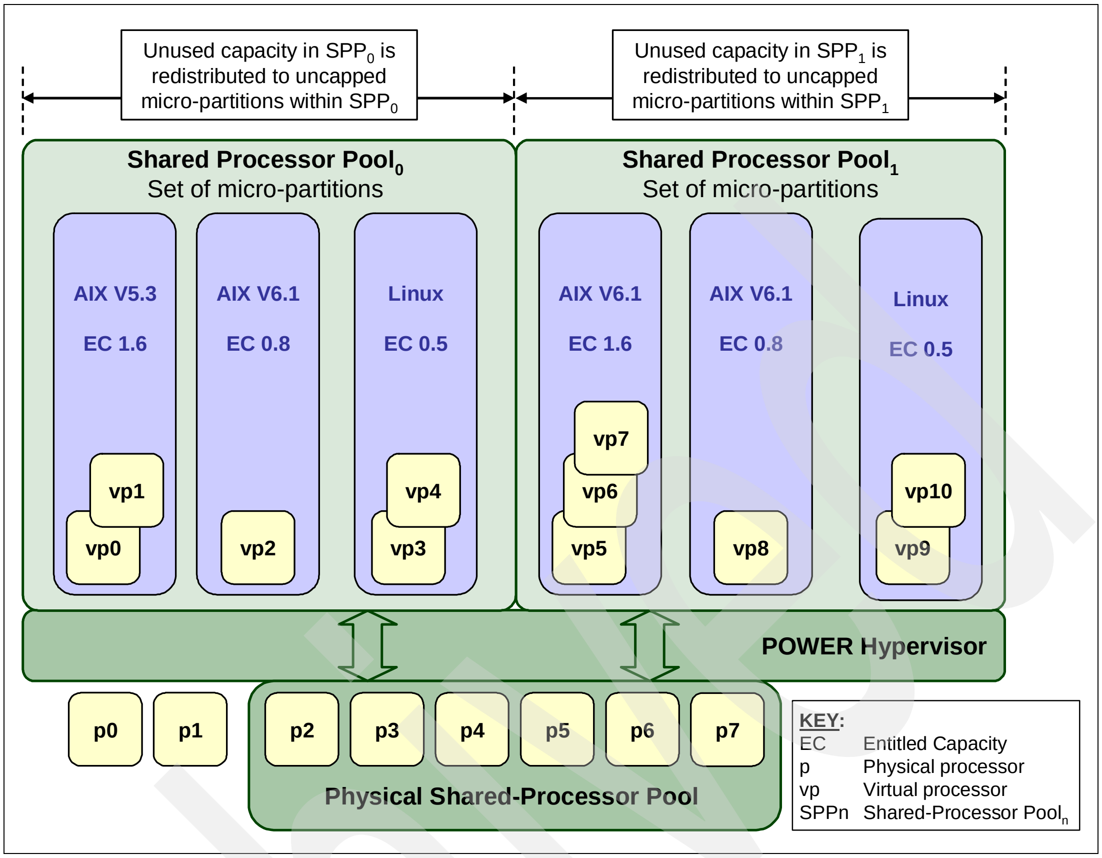
Shared mode

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 a 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 processors exists and the number of logical processors depends on whether the simultaneous multithreading is turned on or off.

3.4.3 Multiple Shared-Processor Pools

One of the functions now available on POWER6- and POWER7-based blades managed by an SDMC is Multiple Shared-Processor Pools (MSPPs). This capability allows a system administrator to create a set of micro-partitions with the purpose of controlling the processor capacity that can be consumed from the physical shared-processor pool.

To implement MSPPs, there is a set of underlying techniques and technologies. An overview of the architecture of Multiple Shared-Processor Pools can be seen in Figure 3-6.



Unused capacity in SPP_0 is redistributed to uncapped micro-partitions within SPP_0
Unused capacity in SPP_1 is redistributed to uncapped micro-partitions within SPP_1
Shared Processor $Pool_0$
Set of micro-partitions
Shared Processor $Pool_1$
Set of micro-partitions
AIX V5.3
EC 1.6
AIX V6.1
EC 0.8
Linux
EC 0.5
AIX V6.1
EC 1.6
AIX V6.1
EC 0.8
Linux
EC 0.5
vp1
vp0
vp2
vp4
vp3
vp7
vp6
vp5
vp8
vp10
vp9
POWER Hypervisor
p0
p1
p2
p3
p4
p5
p6
p7
Physical Shared-Processor Pool
KEY:
EC Entitled Capacity
p Physical processor
vp Virtual processor
SPPn Shared-Processor $Pool_n$

Figure 3-6 Overview of the architecture of Multiple Shared-Processor Pools

Micro-partitions are created and then identified as members of either the default Shared-Processor $Pool_0$ or a user-defined Shared-Processor $Pool_n$. The virtual processors that exist within the set of 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 can be allocated from the Reserved Pool Capacity if the micro-partition is uncapped.

If certain 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 (SPP_0)

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 SPP_0. The default Shared-Processor Pool has the same attributes as a user-defined Shared-Processor Pool except that these attributes are not directly under the control of the system administrator; they have fixed values (Table 3-5).

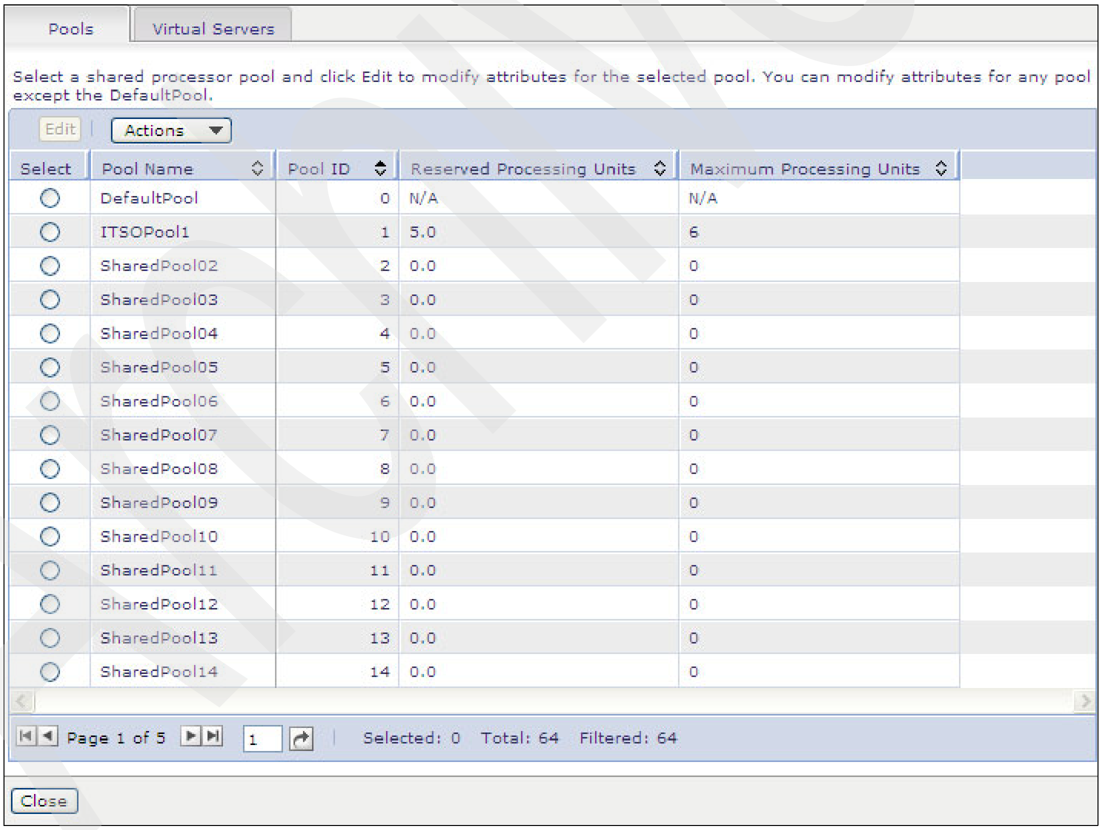
Table 3-5   *Attribute values for the default Shared-Processor Pool (SPP_0)*

SPP_0 attribute	Value
Shared-Processor Pool ID	0
Maximum Pool Capacity	The value is equal to the capacity in the physical shared-processor pool.
Reserved Pool Capacity	0
Entitled Pool Capacity	Sum (total) of the entitled capacities of the micro-partitions in the default Shared-Processor Pool.

Creating Multiple Shared-Processor Pools

The default Shared-Processor Pool (SPP_0) is automatically activated by the system and is always present.

All other Shared-Processor Pools exist, but by default, are inactive. By changing the Maximum Pool Capacity of a Shared-Processor Pool to a value greater than zero, it becomes active and can accept micro-partitions (either transferred from SPP_0 or newly created). Figure 3-7 shows the creation of a processor pool named ITSOPool1 and how it is displayed using an SDMC.



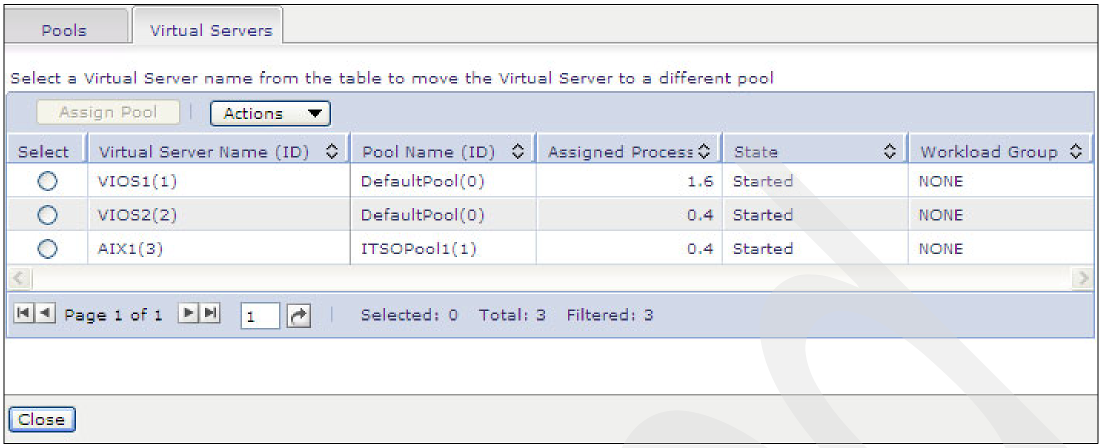
Select a shared processor pool and click Edit to modify attributes for the selected pool. You can modify attributes for any pool except the DefaultPool.

Select	Pool Name	Pool ID	Reserved Processing Units	Maximum Processing Units
○	DefaultPool	0	N/A	N/A
○	ITSOPool1	1	5.0	6
○	SharedPool02	2	0.0	0
○	SharedPool03	3	0.0	0
○	SharedPool04	4	0.0	0
○	SharedPool05	5	0.0	0
○	SharedPool06	6	0.0	0
○	SharedPool07	7	0.0	0
○	SharedPool08	8	0.0	0
○	SharedPool09	9	0.0	0
○	SharedPool10	10	0.0	0
○	SharedPool11	11	0.0	0
○	SharedPool12	12	0.0	0
○	SharedPool13	13	0.0	0
○	SharedPool14	14	0.0	0

Page 1 of 5   1   Selected: 0   Total: 64   Filtered: 64

Figure 3-7   *POWER-based blade processor pools shown using an SDMC*

Figure 3-8 shows the virtual server AIX1 assigned to processor pool ITSOPool1.



Select a Virtual Server name from the table to move the Virtual Server to a different pool

Select	Virtual Server Name (ID)	Pool Name (ID)	Assigned Process	State	Workload Group
○	VIOS1(1)	DefaultPool(0)	1.6	Started	NONE
○	VIOS2(2)	DefaultPool(0)	0.4	Started	NONE
○	AIX1(3)	ITSOPool1(1)	0.4	Started	NONE

Page 1 of 1 | 1 | Selected: 0 Total: 3 Filtered: 3

Figure 3-8 Virtual server assignments to processor pools shown by an SDMC

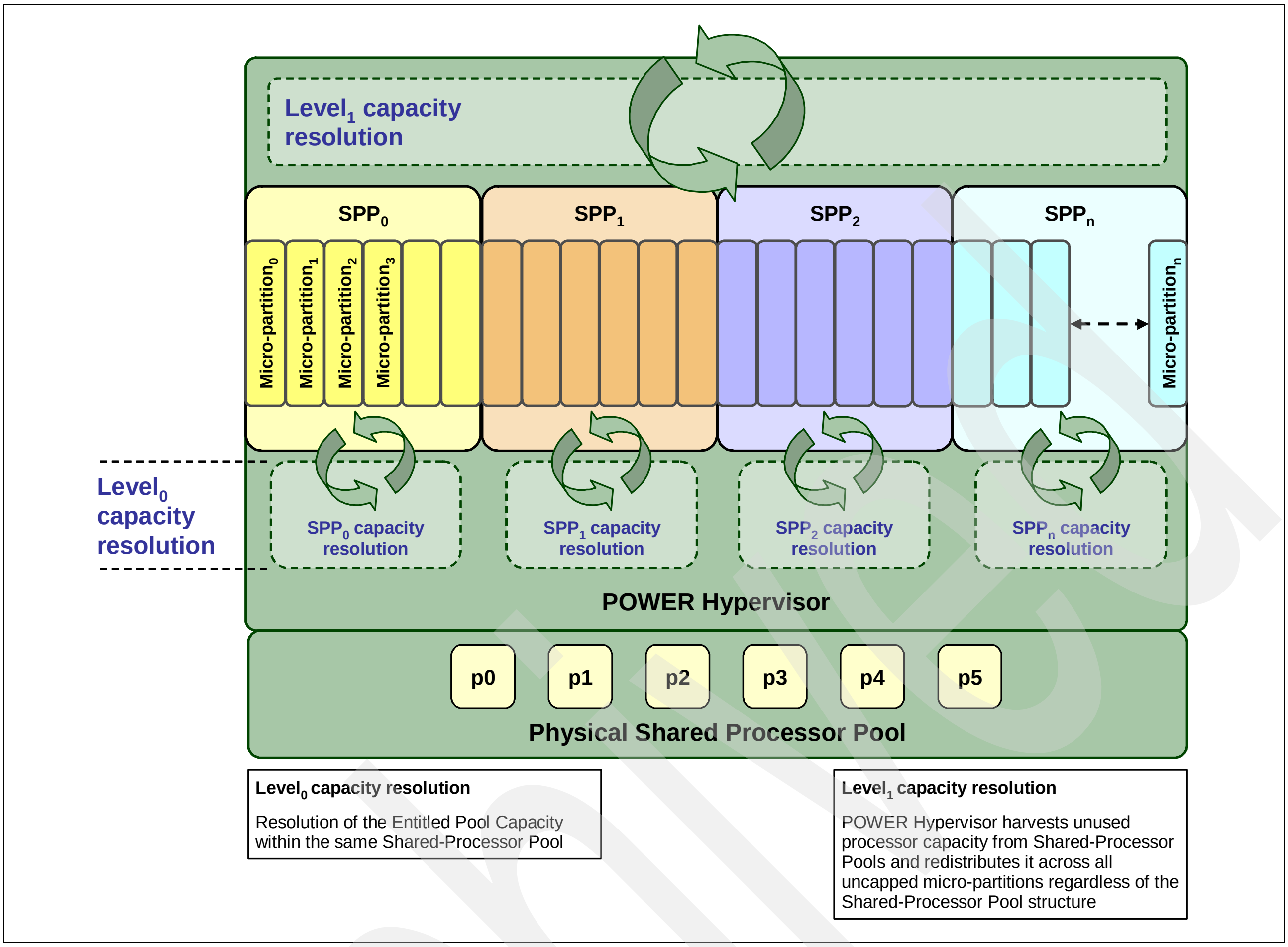
Levels of processor capacity resolution

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

Level$_0$ The first level, Level$_0$, is the resolution of capacity within the same Shared-Processor Pool. Unused processor cycles from within a Shared-Processor Pool are harvested and then redistributed to any eligible micro-partition within the same Shared-Processor Pool.

Level$_1$ When all Level$_0$ 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 on page 106.


Level$_1$ capacity resolution
SPP$_0$
SPP$_1$
SPP$_2$
SPP$_n$
Micro-partition$_0$
Micro-partition$_1$
Micro-partition$_2$
Micro-partition$_3$
Micro-partition$_n$
Level$_0$ capacity resolution
SPP$_0$ capacity resolution
SPP$_1$ capacity resolution
SPP$_2$ capacity resolution
SPP$_n$ capacity resolution
POWER Hypervisor
p0
p1
p2
p3
p4
p5
Physical Shared Processor Pool

Level$_0$ capacity resolution

Resolution of the Entitled Pool Capacity within the same Shared-Processor Pool

Level$_1$ capacity resolution

POWER Hypervisor harvests unused processor capacity from Shared-Processor Pools and redistributes it across all uncapped micro-partitions regardless of the Shared-Processor Pool structure

Figure 3-9 The two levels of unused capacity redistribution

Capacity allocation above the Entitled Pool Capacity (Level$_1$)

The POWER Hypervisor initially manages the Entitled Pool Capacity at the Shared-Processor Pool level. This is where unused processor capacity within a Shared-Processor Pool is harvested and then redistributed to uncapped micro-partitions within the same Shared-Processor Pool. This level of processor capacity management is sometimes referred to as Level$_0$ capacity resolution.

At a higher level, the POWER Hypervisor harvests unused processor capacity from the Multiple Shared-Processor Pools that do not consume all of their Entitled Pool Capacity. If a particular Shared-Processor Pool is heavily loaded and some 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 Level$_1$ capacity resolution.

Important: For Level$_1$ capacity resolution, when allocating additional processor capacity in excess of the Entitled Pool Capacity of the Shared-Processor Pool, the POWER Hypervisor takes into account the uncapped weights of *all micro-partitions in the system, 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 will allocate additional cycles to micro-partitions if all of the following conditions are met:

- 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 such 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.

Dynamic adjustment of Maximum Pool Capacity

The Maximum Pool Capacity of a Shared-Processor Pool, other than the default Shared-Processor Pool$_0$, can be adjusted dynamically from the HMC using either the GUI or CLI.

Dynamic adjustment of Reserve Pool Capacity

The Reserved Pool Capacity of a Shared-Processor Pool, other than the default Shared-Processor Pool$_0$, can be adjusted dynamically from the HMC using either the GUI or CLI.

Dynamic movement between Shared-Processor Pools

A micro-partition can be moved dynamically from one Shared-Processor Pool to another using the HMC and either the GUI or CLI. 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 will reduce the Entitled Pool Capacity for that Shared-Processor Pool. Similarly, the Entitled Pool Capacity of the Shared-Processor Pool that the micro-partition joins will increase.

Deleting a Shared-Processor Pool

Shared-Processor Pools cannot be deleted from the system, but they can be deactivated by setting the Maximum Pool Capacity and the Reserved Pool Capacity to zero. By doing so, the Shared-Processor Pool will still exist but will not be active. You can use the HMC interface to deactivate a Shared-Processor Pool. Be aware that 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 can leave a Shared-Processor Pool due to PowerVM Live Partition Mobility. Similarly, a micro-partition can join a Shared-Processor Pool in the same way. When performing PowerVM Live Partition Mobility, you are given the opportunity to designate a destination Shared-Processor Pool on the target server to receive and host the migrating 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 VIOS

The VIOS is part of all PowerVM Editions. The Virtual I/O partition allows the sharing of physical resources between logical partitions to allow more efficient use. The VIOS owns the physical resources (SCSI, Fibre Channel, network adapters, and optical devices) and allows client partitions to share access to them, minimizing the number of physical adapters in the system. The VIOS eliminates the requirement that every partition owns a dedicated network adapter, disk adapter, and disk drive. The VIOS 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 VIOS configuration utilizing virtual SCSI and virtual Ethernet.

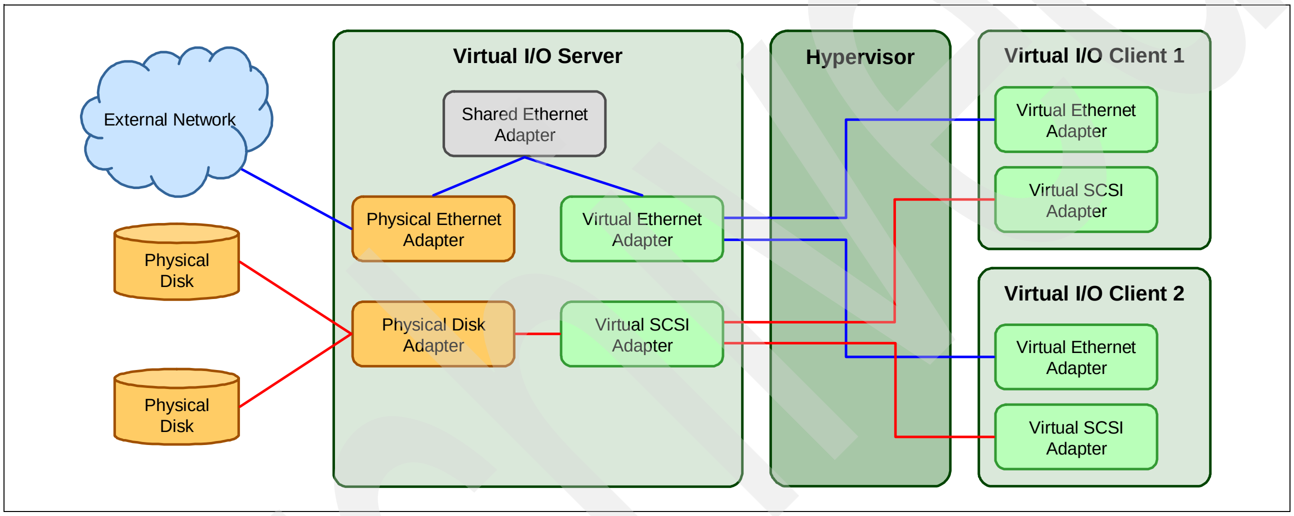


Figure 3-10 Architectural view of the VIOS

Because the VIOS is an operating system-based appliance server, redundancy for physical devices attached to the VIOS can be provided by using capabilities such as Multipath I/O and IEEE 802.3ad Link Aggregation.

Installation of the VIOS 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 VIOS (and IVM in case it is used) and is only supported in special VIOS partitions. Three major virtual devices are supported by the Virtual I/O Server:

- Shared Ethernet Adapter
- Virtual SCSI
- Virtual Fibre Channel adapter

The Virtual Fibre Channel adapter is used with the NPIV feature, described in 3.4.9, “N_Port ID Virtualization (NPIV)” on page 116.

Shared Ethernet Adapter

A Shared Ethernet Adapter (SEA) can be used to connect a physical Ethernet network to a virtual Ethernet network. The SEA provides this access by connecting the internal Hypervisor VLANs with the VLANs on the external switches. Because the SEA processes packets at layer 2, the original MAC address and VLAN tags of the packet are visible to other systems on the physical network. IEEE 802.1 VLAN tagging is supported.

The SEA also provides the ability for several client partitions to share a physical adapter. Using an SEA, you can connect internal and external VLANs using a physical adapter. The SEA service can only be hosted in the VIOS, 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 VIOS.

Tip: A Linux partition can provide bridging function as well, by using the **`brctl`** command.

Figure 3-11 shows a configuration example of an SEA with one physical and two virtual Ethernet adapters. An SEA can include up to 16 virtual Ethernet adapters on the VIOS that share the same physical access.

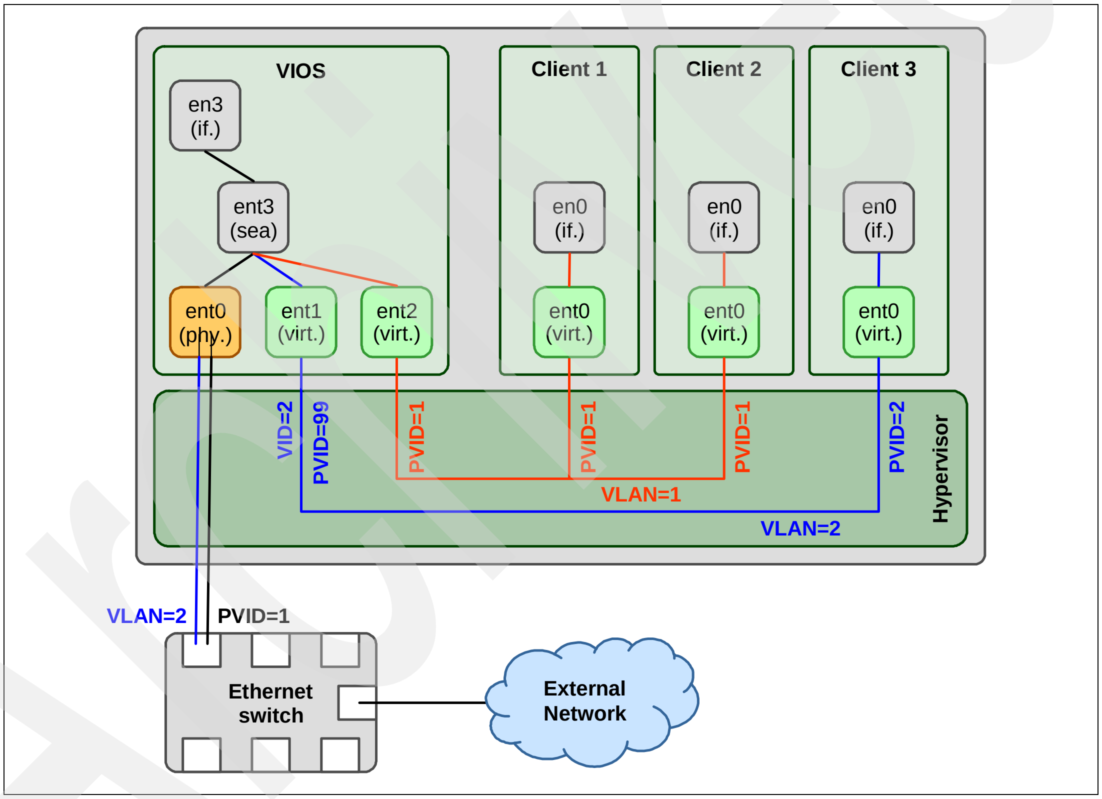


Figure 3-11 Architectural view of a SEA

A single SEA setup can have up to 16 Virtual Ethernet trunk adapters and each virtual Ethernet trunk adapter can support up to 21 VLANs (20 VIDs and 1 PVID). The number of SEAs that can be set up in a VIOS partition is limited only by the resource availability because there are no configuration limits.

Unicast, broadcast, and multicast are supported, so protocols that rely on broadcast or multicast, such as Address Resolution Protocol (ARP), Dynamic Host Configuration Protocol (DHCP), Boot Protocol (BOOTP), and Neighbor Discovery Protocol (NDP) can work across an SEA.

Note: An SEA does not need to have an IP address configured to perform the Ethernet bridging functionality. Configuring an IP address is required to access the IVM.

For a more detailed discussion about virtual networking, see the following web page:

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

Virtual SCSI

Virtual SCSI is provided by the POWER Hypervisor and provides a virtualized implementation of the SCSI protocol. This implementation provides secure communication between the Client partition (AIX, Linux, or IBM i) and the VIOS in a client/sever relationship. The VIOS logical partition or virtual server owns the physical resources and acts as server for them or, in SCSI terms, target devices. The client logical partitions access the virtual SCSI backing storage devices provided by the VIOS as clients.

The virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured using an SDMC or through IVM on POWER-based blade servers. The virtual SCSI server (target) adapter is responsible for executing any SCSI commands it receives. It is owned by the VIOS 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 VIOS.

Physical disks presented to the VIOS can be assigned to a client partition in a number of ways:

- The entire disk is presented to the client partition.
- The disk is divided into several logical volumes, which can be presented to a single client or multiple clients as virtual disks.
- 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 111 shows an example where one physical disk is divided into two logical volumes by the VIOS. 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 client partition, the disk is seen as a normal hdisk.

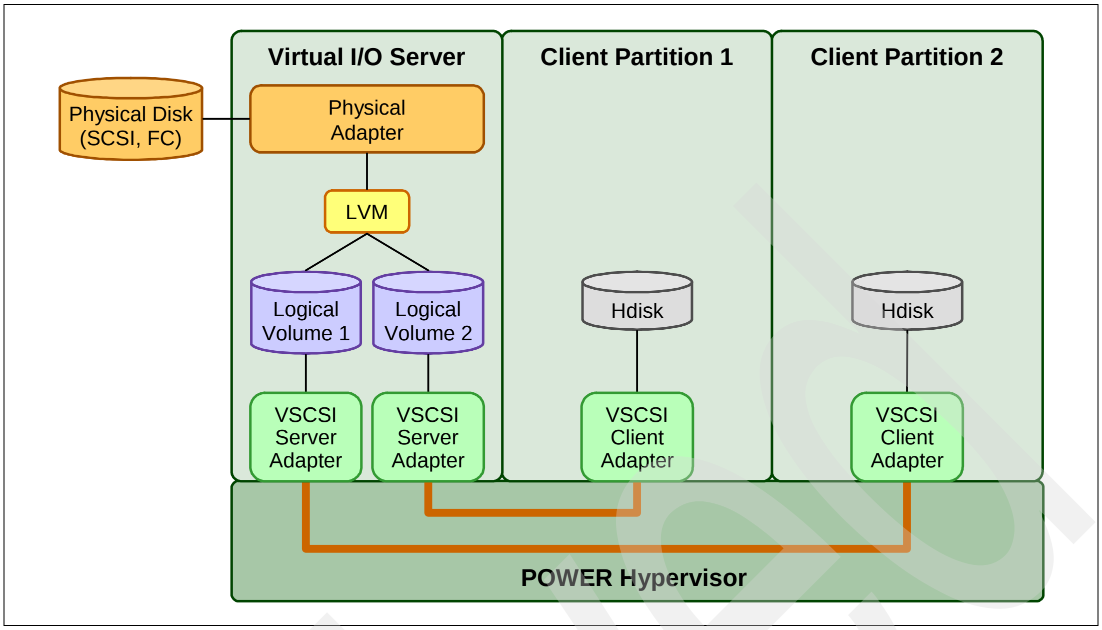


Figure 3-12 Architectural view of virtual SCSI

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 VIOS, see the following web page:

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

VIOS functions

VIOS includes a number of features, including monitoring solutions, as follows:

- Support for Live Partition Mobility on POWER6 and POWER7 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 112.
- 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.
- VIOS 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).
- The Workload Estimator, designed to ease the deployment of a virtualized infrastructure.
- IBM Systems Director agent and a number of IBM Tivoli Management agents are included, such as Tivoli Identity Manager (TIM) and Tivoli Application Dependency Discovery Manager (ADDM).
- vSCSI eRAS.

For more information about the Virtual I/O Server and its implementation, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, available from the following web page:

http://www.redbooks.ibm.com/abstracts/sg247940.html

3.4.5 PowerVM Lx86

The PowerVM Editions hardware feature includes PowerVM Lx86. Lx86 is a dynamic, binary translator that allows Linux applications (compiled for Linux on Intel architectures) to run without change alongside local Linux on POWER applications. Lx86 makes this possible by dynamically translating x86 instructions to POWER and caching them to enhance translation performance. In addition, Lx86 maps Linux on Intel architecture system calls to Linux on POWER architecture system calls. No modifications or recompilations of the x86 Linux applications are needed.

PowerVM Lx86 creates a virtual x86 environment in which the Linux on Intel applications can run. Currently, a virtual Lx86 environment supports SUSE Linux or Red Hat Linux x86 distributions. The translator and the virtual environment run strictly within user space. No modifications to the POWER kernel are required. PowerVM Lx86 does not run the x86 kernel on the POWER system and is not a virtual machine. Instead, x86 applications are encapsulated so that the operating environment appears to be Linux on x86, even though the underlying system is a Linux on POWER system.

PowerVM Lx86 is included in the PowerVM Express Edition, PowerVM Standard Edition, and in the PowerVM Enterprise Edition. More information about PowerVM Lx86 can be found at the following web page:

http://www.ibm.com/systems/power/software/virtualization/editions/lx86/

3.4.6 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility allows you to move a running logical partition or virtual server, including its operating system and running applications, from one system to another without disrupting the infrastructure services. The migration transfers the entire system environment, including processor state, memory, attached virtual devices, and connected users. Inactive partition mobility allows you to move a powered-off logical partition or virtual server from one hardware platform to another.

Note: Partition Mobility is only available with the Enterprise PowerVM Edition.

Partition mobility provides systems management flexibility and improves system availability, as follows:

- Avoid planned outages for server upgrade, hardware, or firmware maintenance. Move logical partitions or virtual servers to another server and perform the maintenance. Live Partition Mobility can help lead to zero downtime maintenance because you can use it to work around scheduled maintenance activities.
- Meet stringent service level agreements. You can proactively move the running partition and the applications from one server to another.
- Balance workload and resources. Should a key application's resource requirements peak unexpectedly to a point where there is contention for server resources, you might move it to a larger server or move other, less critical, partitions to separate servers, and use the freed-up resources to absorb the peak.

- Optimize the server. Consolidate workloads running on several small, under-used servers onto a single large server.

System requirements for Partition Mobility

Both source and destination systems must have the PowerVM Enterprise Edition license code installed. The source partition must be a virtual client and should have only 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.

An SDMC-managed blade requires the mover service attribute be checked for the VIOS virtual server. An IVM-managed blade has this attribute set by default. The source and the target VIOS can communicate over the network. The Virtual Asynchronous Services Interface (VASI) device provides communication between the mover service partition and the POWER Hypervisor.

Note: When you move an active logical partition between servers with different processor types (such as POWER 6 and POWER7), both current and preferred compatibility modes of the logical partition must be supported by the destination server.

To migrate partitions 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.

To move an active logical partition or virtual server 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 following these steps:

1. Set the 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/iphc3pcmcombosact.htm`

For more information about Live Partition Mobility and how to implement it, see *IBM PowerVM Live Partition Mobility*, SG24-7460, available from the following web page:

`http://www.redbooks.ibm.com/abstracts/sg247460.html`

The SDMC or IVM is used to configure, validate, and orchestrate Live Partition Mobility on POWER-based blades. Use the SDMC to enable mover service function for Virtual I/O Server; an SDMC or IVM wizard validates your configuration and identifies issues that will cause the migration to fail. During the migration, the SDMC or IVM controls all phases of the process.

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 PowerVM Enterprise edition.

The physical memory of an IBM POWER6 or POWER7 system can be assigned to multiple partitions in either a dedicated or a shared mode. The system administrator has the capability to assign physical memory to a partition and physical memory to a pool that is shared by other partitions. A single partition can have either dedicated or shared memory.

In a dedicated memory model, the system administrator's task is to optimize available memory distribution among partitions. When a partition has performance degradation due to memory constraints and other partitions have unused memory, the administrator can allocate memory by doing a DLPAR operation.

With a shared memory model, it is the system (PowerVM Hypervisor) that automatically decides the optimal distribution of the physical memory to partitions and adjusts the memory assignment based on partition load.

Active Memory Sharing can be exploited to increase memory use on the system either by decreasing the system 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 systems. For example, AIX partitions can take advantage of Active Memory Expansion while other operating systems take advantage of Active Memory Sharing.

For additional information regarding Active Memory Sharing see *PowerVM Virtualization Active Memory Sharing*, REDP-4470, available from:

http://www.redbooks.ibm.com/abstracts/redp4470.html

Figure 3-13 on page 115 shows each logical partition can be configured to have shared memory using Active Memory Sharing or dedicated memory.

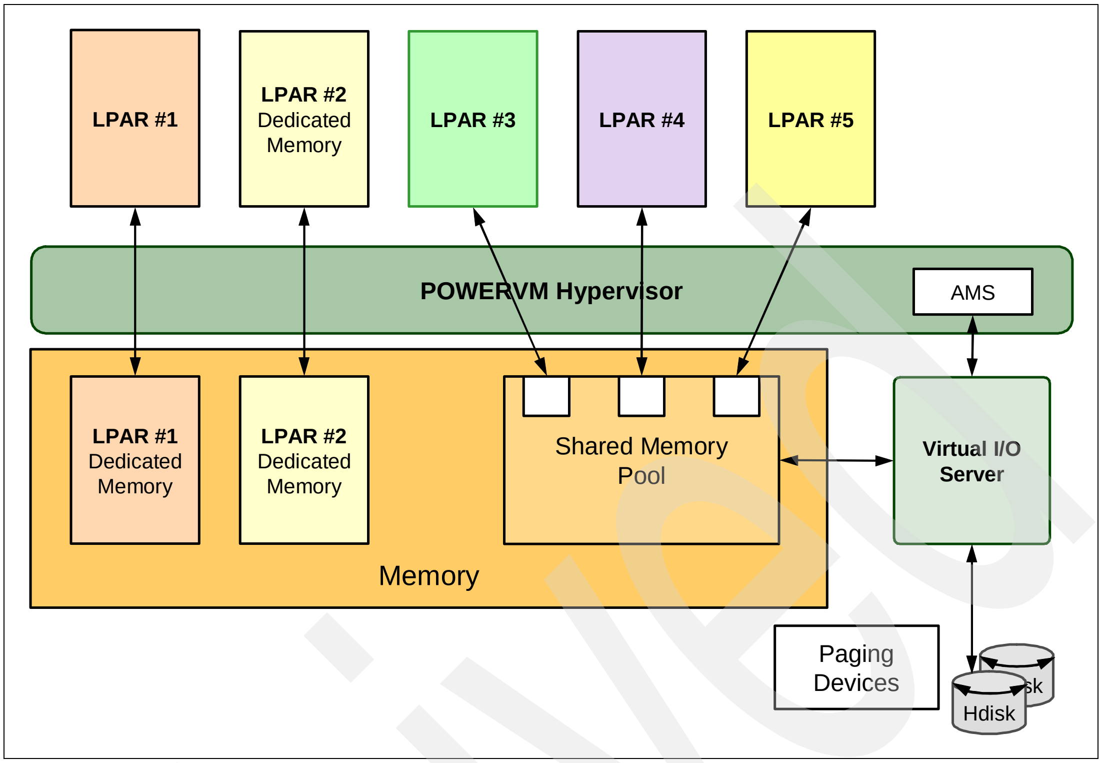


Figure 3-13 Active Memory Sharing block diagram with shared and dedicated memory

3.4.8 Suspend/Resume

Suspend/Resume partition or virtual server can provide long-term suspension of partitions. Partition state (memory, NVRAM, and VSP state) is saved on persistent storage, freeing server resources that were in use by that partition. A suspended partition can be resumed from a suspended state at a later time on the same system or migrated to another server and resumed.

Key requirements for Suspend/Resume are:

- VIOS 2.2 or later
- AIX6.1 TL6 SP1 or AIX7.1 SP1
- No Physical devices assigned to partition or virtual server
- No BSR or Huge Pages
- POWER6/6+ or POWER7 compatibility modes
- Dedicated or AMS memory
- Requires VIOS Paging Space Devices

Additional reference and implementation information can be found in *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940 and *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

3.4.9 N_Port ID Virtualization (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. NPIV provides direct access to Fibre Channel adapters from multiple client partitions, simplifying the Fibre Channel SAN environment. This adapter is attached to a VIOS 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.

Table 3-6 shows the NPIV compatibility matrix for AIX and Linux clients with feature codes.

Table 3-6 NPIV compatibility matrix for AIX and Linux clients

I/O modules in the BladeCenter chassis	**Expansion Card on the blade servers**		
	QLogic 8Gb FC CIOv (#8242)[a]	**QLogic 8Gb FC CFFh (#8271)**[a]	**Emulex 8Gb FC CIOv (#8240)**
QLogic 4 Gb Switch Module (#3243, #3244)[b]	No	No	No
QLogic 8 Gb Switch Module (#3284)[c]	Yes	Yes	No
Brocade 4 Gb Switch Module (#3206, #3207)	No	No	Yes[d]
Brocade 8 Gb Switch Module (#5045, # 5869)	Yes	Yes	Yes
Cisco 4Gb Switch Module (#3242)	Yes[e] [f]	Yes[e] [f]	Yes[f] [g]
QLogic 8Gb Internal Passthrough Module (#5449)	Yes[h]	Yes[h]	No

a. Requires firmware v5.02.01 or later

b. Requires firmware v6.5.022.00 or later

c. Requires firmware v7.10.1.04 or later

d. Requires AIX6.1 TL5 and AIX5.3 TL12, Emulex adapter firmware v1.11A8 and Brocade switch firmware v6.2.1b or later

e. Requires VIOS2.1.2, AIX5.3 TL12, AIX6.1 TL5 4Gb Cisco SM firmware v4.2.3, QLogic 8Gb firmware 0314050309, or later

f. Linux not supported for this combination

g. Requires AIX6.1 TL5 or AIX 5.3 TL11 or later levels, Emulex adapter firmware 1.11A8 and Cisco SM firmware v4.2.3 or later

h. Requires VIOS 2.1.3, AIX6.1 TL6, AIX7.1, 8Gb firmware v7.10.1.4.0 or later

Table 3-7 on page 117 shows the NPIV compatibility matrix for IBM i clients with feature codes.

Table 3-7 NPIV Compatibility Matrix for IBM i Client

I/O modules in the BladeCenter chassis	Expansion Card on the blade servers		
	QLogic 8Gb FC CIOv (#8242)[a]	**QLogic 8Gb FC CFFh (#8271)[a]**	**Emulex 8Gb FC CIOv (#8240)**
QLogic 4 Gb Switch Module (#3243, #3244)[b]	No	No	No
QLogic 8 Gb Switch Module (#3284)[c]	Yes[d]	Yes[d]	No
Brocade 4 gb Switch Module (#3206, #3207)	No	No	Yes[e]
Brocade 8 gb Switch Module (#5045, # 5869)	Yes[d]	Yes[d]	Yes[e]
Cisco 4Gb Switch Module (#3242	Yes[f]	Yes[f]	Yes[e]
QLogic 8Gb Int. Passthrough Module (#5449)	No	No	No
10 GbE Passthrough Module (#5412)	Not applicable	Not applicable	Not applicable

[a] Requires firmware v5.02.01 or later
[b] Requires Firmware level 6.5.0.22 or later
[c] Requires Firmware level 7.10.1.4 or later
[d] DS8000 support requires QLogic 8Gb FC adapter firmware 0314050309 or later
[e] Virtual tape only with VIOS 2.1.2 or later
[f] Requires VIOS 2.2.0, IBM i 6.1.1 or 7.1, Cisco SM firmware v3.3.2 or later

For additional information about NPIV, see the following resources:

- *PowerVM Migration from Physical to Virtual Storage*, SG24-7825

  http://www.redbooks.ibm.com/abstracts/sg247825.html

- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

  http://www.redbooks.ibm.com/abstracts/sg247590.html

NPIV is supported in PowerVM Express, Standard, and Enterprise Editions on the IBM POWER7 processor-based systems.

3.4.10 Supported PowerVM features by operating system

Table 3-8 summarizes the PowerVM features that are supported by the operating systems and that are compatible with the POWER7 processor-based blade servers.

Table 3-8 PowerVM features supported on AIX, IBM i, and Linux operating systems

Feature	AIX V5.3	AIX V6.1	AIX V7.1	IBM i 6.1.1	IBM i 7.1	SLES 10 SP3	SLES 11 sp1	RHEL 5.6	RHEL 6.0
Dynamic simultaneous multithreading (SMT)	Yes[a]	Yes[b]	Yes	Yes[c]	Yes	Yes	Yes[a]	Yes	Yes
DLPAR I/O adapter add/remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR processor add/remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR memory add	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR memory remove	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Micro-Partitioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared Dedicated Capacity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Feature	AIX V5.3	AIX V6.1	AIX V7.1	IBM i 6.1.1	IBM i 7.1	SLES 10 SP3	SLES 11 sp1	RHEL 5.6	RHEL 6.0
Multiple Shared Processor Pools	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual I/O Server	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IVM	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SDMC	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual SCSI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Ethernet	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
NPIV	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Live Partition Mobility	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes
Workload Partitions	No	Yes	Yes	No	No	No	No	No	No
Active Memory Sharing	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Active Memory Expansion	No	Yes	Yes	No	No	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

> **Note:** Most of the features listed in Table 3-8 require that VIOS/IVM or VIOS managed by an SDMC be used to configure features. Native OS does not support the functions described in Table 3-8.

For more information see "Supported features for Linux on Power Systems servers" in the Linux on Power Infocenter:

http://publib.boulder.ibm.com/infocenter/lnxinfo/v3r0m0/index.jsp?topic=/liaam/supportedfeaturesforlinuxonpowersystemsservers.htm

4

Continuous availability and manageability

This chapter provides information about IBM reliability, availability, and serviceability (RAS) design and features. This set of technologies, implemented on IBM Power Systems servers, provides the possibility to improve your architecture’s total cost of ownership (TCO) by reducing unplanned down time.

RAS can be described as follows:

- Reliability

  Reliability indicates how infrequently a defect or fault in a server manifests itself.

- Availability

  Availability indicates how infrequently the functionality of a system or application is impacted by a fault or defect.

- Serviceability

  Serviceability indicates how well faults and their effects are communicated to users and services and how efficiently and nondisruptively the faults are repaired.

This chapter covers the following topics:

4.1 Introduction

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, ease administrative burden, and increase system use.

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 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. This includes selective redundancy to tolerate certain faults without requiring an outage or parts replacement.

The PS703 and PS704 blades are used with a BladeCenter chassis and the various components that make up the BladeCenter infrastructure. In general, the BladeCenter infrastructure RAS is outside the scope of this chapter. However, when appropriate, the BladeCenter features that enable, complement, or enhance RAS functionality on the PS703 and PS704 blades are discussed.

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

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

Design and manufacturing verification and integration, along with field support feedback, informs and motivates continued improvement on the final products.

This chapter 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 varies.

4.2 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.

4.2.1 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 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. This reduces the total number of system components and reduces the total amount of heat that is generated in the design. This results 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.

4.2.2 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 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 the blades so they receive fresh air during operation.

4.2.3 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 with operation.

Redundant parts allow the system to remain operational. In POWER processor-based systems this includes:

- POWER7 cores with redundant bits in L1-I, L1-D, L2 caches, and L2 and L3 directories
- Redundant and hot-swap cooling fans in the BladeCenter chassis
- Redundant and hot-swap power supplies in the BladeCenter chassis
- Redundant integrated Ethernet ports on the blade with separate paths to independent I/O module bays in the BladeCenter
- Redundant paths for I/O expansion cards through the BladeCenter midplane to independent I/O module bays in the BladeCenter

For maximum availability, a strong recommendation is to connect power cords from the BladeCenter to two separate Power Distribution Units (PDUs) in the rack, and to connect each PDU to independent power sources.

4.3 Availability

The IBM hardware and microcode ability to monitor execution of hardware functions is generally described as the process of first-failure data capture (FFDC). This process includes predictive failure analysis. Predictive failure analysis 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 recreate 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 perform the following functions:

- Self-diagnose and self-correct during run time
- Automatically reconfigure to mitigate potential problems from suspect hardware
- Self-heal or substitute good components for failing components automatically

> **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 4.4.1, "Detecting" on page 131.

4.3.1 Partition availability priority

Also available is the ability to assign availability priorities to partitions. If an alternate processor recovery event requires spare processor resources and there are no other means of obtaining the spare resources, the system determines which partition has the lowest priority and attempts to claim the needed resource. On a properly configured POWER processor-based server, this approach allows that capacity to be first obtained from a low priority partition instead of a high priority partition.

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

Partition-availability priority is assigned to partitions by using a weight value or integer rating. The lowest priority partition is rated at 0 (zero) and the highest priority partition is valued at 255. The default value is set at 127 for standard partitions and 192 for Virtual I/O Server (VIOS) partitions. You can vary the priority of individual partitions.

> **Note:** On IVM-managed systems the partition availability priority is changed by using the `chsycfg` command with the lpar_avail_priority flag. SDMC-managed systems can change the virtual server priority from the Power Systems Resources view by right-clicking the server name and selecting **Virtual Server Availability Priority**.

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.

The partition specifications for minimum, desired, and maximum capacity are 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 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.

4.3.2 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 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 on both 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 the following elements:

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

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 certain logic functions, the POWER7 processor automatically retries the instruction. If the source of the error was truly transient, the instruction succeeds and the system continues as before.

On IBM systems prior to POWER6, this error 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 continues 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, which 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 effectively puts a CoD processor into operation after it has been determined that an activated processor is no longer operational. In this way the server remains with its total processor power.

If there are no CoD processor cores available, system-wide total processor capacity is lowered beneath the licensed number of cores.

Single processor checkstop

As in POWER6, POWER7 provides single processor check stopping for certain processor logic or command or control errors that cannot be handled by the availability enhancements mentioned previously.

This 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, errors might result on a system-wide outage.

4.3.3 Memory protection

A memory protection architecture that provides good error resilience for a relatively small L1 cache might be inadequate for protecting the much larger system main store. Therefore, a variety of protection methods are used in POWER processor-based systems to avoid uncorrectable errors in memory.

Memory protection plans must take into account many factors, including:

- 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. This includes the following capabilities:

- 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 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.

- Hardware-assisted memory scrubbing

  Memory 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 retune bus parameters dynamically when a fault occurs. In addition, the memory bus has spare capacity to substitute a spare data bit-line that 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 affects one bit in each word at most. The system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced, assuming no additional single bit errors.

POWER7 memory subsystem

The POWER7 chip contains two memory controllers with four channels per memory controller. The implementation on the PS703 and PS704 blades uses a single memory controller per processor chip and four advanced memory buffer chips. Each memory buffer chip connects to two memory DIMMs, 8 total per processor chip.

The bus transferring data between the processor and the memory uses CRC error detection with a failed operation retry mechanism and the ability to retune bus parameters dynamically when a fault occurs. In addition, the memory bus has spare capacity to substitute a spare data bit-line for that which is determined to be faulty.

Figure 4-1 shows a POWER7 chip as implemented on a PS703 or PS704 blade, with its memory interface comprised of one controller and four advanced memory buffers. Advanced memory buffer chips are exclusive to IBM. They help to increase performance acting as read/write buffers. The four advanced memory buffer chips are on the system planar and support two DIMMs each.

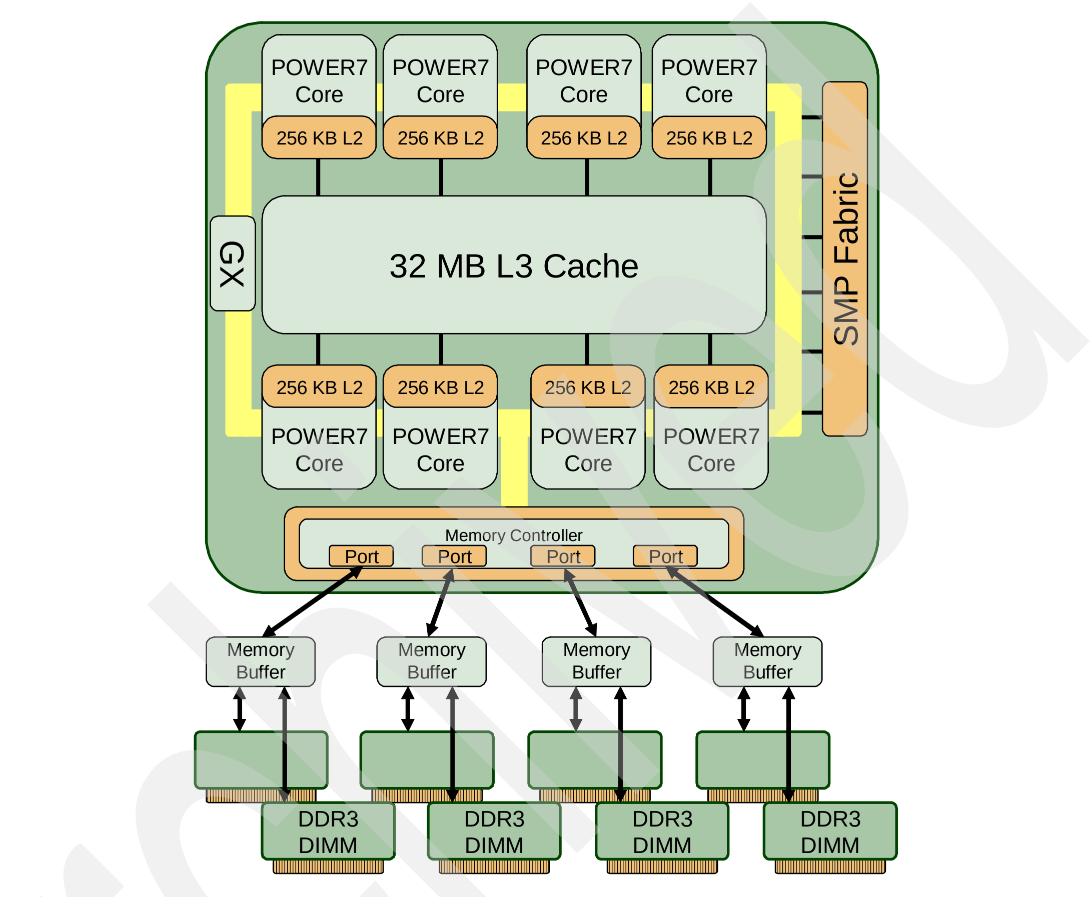


Figure 4-1 PS702 and PS703 memory subsystem

Memory page deallocation

Although coincident cell errors in separate memory chips are a statistical 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 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.

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 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 when waiting for parts replacement.

> **Note:** Although memory page deallocation handles single cell failures, because of the sheer size of data in a data bit line, it might be inadequate for dealing with more catastrophic failures. Redundant bit steering continues to be the preferred method for dealing with these types of problems.

Memory 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 used on subsequent reboots.

Upon reboot, if not enough memory is available to meet minimum partition requirements, the POWER Hypervisor reduces the capacity of one or more partitions.

Depending on the configuration of the system, the IVM Electronic Service Agent™, OS Service Focal Point, SDMC Service and Support Manager, or BladeCenter Advanced Management Module Service Advisor receives a notification of the failed component, and triggers a service call.

4.3.4 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. 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 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 checking and correcting code (ECC). Single-bit errors are corrected before forwarding to the processor, and subsequently written back to L2 and L3.

In addition, the caches maintain a cache-line delete capability. A threshold of correctable errors detected on a cache line can result in the data in the cache line being purged and the cache line removed from further operation without requiring a reboot. An ECC uncorrectable error detected in the cache can also trigger a purge and delete of the cache line. This does not result in a loss of operation because an unmodified copy of the data can be held on system memory to reload the cache line from main memory. Modified data would be handled through Special Uncorrectable Error handling.

L2 and L3 deleted cache lines are marked for persistent deconfiguration on subsequent system reboots until they can be replaced.

4.3.5 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. Consider the following examples:

- Data in the instruction L1 cache is never modified within the cache itself. Therefore, an uncorrectable error discovered in the cache is treated as 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 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. Rather, the system tags the data and determines whether it will ever be used again. Note the following information:

- If the error is irrelevant, it does not force a check stop.
- If the data is used, termination can be limited to the program or kernel, or hypervisor owning the data. Also possible is the 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 notified, and takes appropriate actions. When running AIX (since V5.2 and later) or Linux, and a process attempts to use the data, the operating system is informed of the error and might terminate, or might only terminate a specific process associated with the corrupt data. This depends 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 must the entire system must be rebooted, thereby preserving overall system integrity.

Depending on system configuration and source of the data, errors encountered during I/O operations might not result in a machine check. Instead, the incorrect data is handled by the processor host bridge (PHB) chip. When the PHB chip detects a problem it rejects the data, preventing data being written to the I/O device.

The PHB enters a freeze mode that halts normal operations. Depending on the model and type of I/O being used, the freeze might include the entire PHB chip, or 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 is performed. 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 can enable the system to recover transparently, without partition loss.

4.3.6 PCI extended 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 extended error handling (EEH) enabled adapters respond to a special data packet that is generated from the affected PCI slot hardware by calling system firmware (that examines the affected bus), allowing 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 might not provide native EEH support.

To detect and correct PCIe bus errors, POWER7 processor-based systems use CRC detection and instruction retry correction.

4.4 Serviceability

IBM Power Systems design enables IBM to be responsive to 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. Such an environment includes the following elements:

- 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.

The term *servicer*, when used in the context of this discussion, denotes the person tasked with performing service-related actions on a system. For an item designated as a Customer Replaceable Unit (CRU), the servicer might be the client. In other cases, for Field Replaceable Unit (FRU) items, the servicer might be an IBM representative or an authorized warranty service provider.

Service can be divided into three main categories:

- **Service Components:** The basic service-related building blocks
- **Service Functions:** Service procedures or processes containing one or more service components
- **Service Operating Environment:** The specific system operating environment, which specifies how service functions are provided by the various service components

The basic component of service is a *Serviceable Event*.

Serviceable events are platform, regional, and local error occurrences that require a service action (repair). This action can include a *call home* to report the problem so that the repair can be assessed by a trained service representative. In all cases, the client is notified of the event. Event notification includes a clear indication of when servicer intervention is required to rectify the problem. The intervention might be a service action that the client can perform or it might require a service provider.

Serviceable events are classified as follows:

1. Recoverable: This is a correctable resource or function failure. The server remains available, but there might be some decrease in operational performance available for client’s workload (applications).
2. Unrecoverable: This is an uncorrectable resource or function failure. In this instance, there is potential degradation in availability and performance, or loss of function to the client’s workload.
3. Predictable (using thresholds in support of Predictive Failure Analysis): This is a determination that continued recovery of a resource or function might lead to degradation of performance or failure of the client’s workload. Although the server remains fully available, if the condition is not corrected, an unrecoverable error might occur.
4. Informational: This is notification that a resource or function:
   - Is *out-of* or *returned-to* specification and might require user intervention.
   - Requires user intervention to complete one or more system tasks.

Platform errors are faults that affect all partitions in various ways. They are detected in the blade by the Service Processor, the System Power Control Network, or the Power Hypervisor. When a failure occurs in these components, the POWER Hypervisor notifies each partition’s operating system to execute any required precautionary actions or recovery methods. The OS is required to report these kinds of errors as serviceable events to the Service Focal Point application because, by definition, they affect the partition.

Platform errors are faults related to:

- The sysplanar: that part of the server composed of the central processor units, memory, storage controls, and the I/O hubs
- The power and cooling subsystems
- The firmware used to initialize the system and diagnose errors

Regional errors are faults that affect some, but not all partitions. They are detected by the POWER Hypervisor or the Service Processor.

Local errors are faults detected in a partition (by the partition firmware or the operating system) for resources owned only by that partition. The POWER Hypervisor and Service Processor are not aware of these errors. Local errors might include “secondary effects” that result from platform errors preventing partitions from accessing partition-owned resources. Examples include PCI adapters or devices assigned to a single partition. If a failure occurs to one of these resources, only a single operating system partition need be informed.

This section provides an overview of the progressive steps of error detection, analysis, reporting, notifying, and repairing that are found in all POWER processor-based systems.

4.4.1 Detecting

The first and most crucial component of a solid serviceability strategy is the ability to detect errors accurately and effectively 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 separate microprocessor from the main instruction processing complex. The service processor provides the capabilities for the following elements:

- POWER Hypervisor (system firmware), IVM, Service and Support Module (SSM) under the SDMC, and BladeCenter Advanced Management Module (AMM) coordination
- Remote power control options
- Reset and boot features
- Environmental monitoring

  The service processor monitors the server’s built-in temperature sensors and sends this information to the BladeCenter AMM. The AMM can send instructions to the BladeCenter fans to increase rotational speed when the ambient temperature is beyond the normal operating range. Using an architected operating system interface, the service processor notifies the operating system of potential environmental 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 in the following circumstances:

  – The operating temperature exceeds the critical level (for example, failure of air conditioning or air circulation around the system)
  – The system fan speed is out of operational specification (for example, because of multiple fan failures)

  - The server input voltages are out of operational specification

  The service processor can immediately shut down a system in the following circumstances:

  - Temperature exceeds the critical level or if the temperature remains beyond the warning level for too long
  - Internal component temperatures reach critical levels

- 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 by the BladeCenter AMM, can reboot the system automatically following AC power failure.

- Fault monitoring

  The 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.

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 runtime detection of each unique failure.
- Ability to initiate a wide variety of recovery mechanisms designed to correct the problem. The POWER processor-based systems include extensive hardware and firmware recovery logic.

Fault isolation registers

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, runtime error diagnostics can be deterministic so that for every check station, the unique error domain for that checker is defined and documented. Ultimately, the error domain becomes the field-replaceable unit (FRU) call, and manual interpretation of the data is not normally required.

First-failure data capture (FFDC)

First-failure data capture (FFDC) is an error isolation technique which ensures that when a fault is detected in a system through error checkers or other types of detection methods, the root cause of the fault is captured without the need to recreate the problem or run an extended tracing or diagnostics program.

For the vast majority of faults, a good FFDC design means that the root cause is detected automatically without intervention by a service representative. Pertinent error data related to the fault is captured and saved for analysis. In hardware, FFDC data is collected from the fault isolation registers and from the associated logic. In firmware, this data consists of return codes, function calls, and so forth.

FFDC check stations are carefully positioned within the server logic and data paths to ensure 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 4-2 shows a schematic of a fault isolation register implementation.

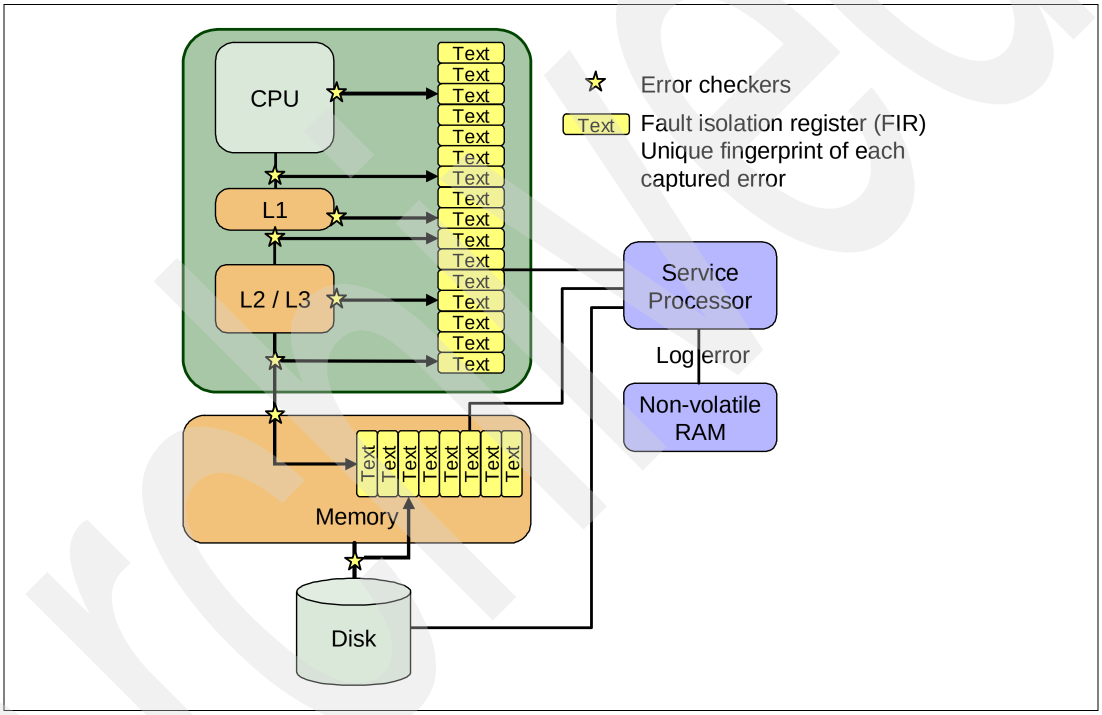


Figure 4-2 Schematic of a FIR implementation

Fault isolation

The service processor interprets error data captured by the FFDC checkers (saved in the FIRs or other firmware-related data capture methods) 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 might be incremented. No specific service action is necessary when the event is recoverable.

When the event requires a service action, additional required information is collected to service the fault. For unrecoverable errors or for recoverable events that meet or exceed their service threshold (meaning that a service action point has been reached) a request for service is initiated through an error logging component.

4.4.2 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 the following elements:

- 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 a processor-based power-on self-test (POST), or through diagnostics.
- Wire-tests discover and precisely identify connection faults between components such as processors, memory, or I/O hub chips.
- Initialization of components such as ECC memory, typically by writing patterns of data and allowing the server to store valid ECC data for each location, can help isolate errors.

To minimize boot time, the system determines which of the diagnostics are required to be started to ensure correct operation based on the way the system was powered off, or through 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 capability originates with the service processor. Extensive diagnostic and fault analysis routines have been developed and improved over many generations of POWER processor-based servers. They 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 to predict when corrective actions must be automatically initiated by the system. This includes the following actions:

- Requests for a part to be replaced
- Dynamic invocation of built-in redundancy for automatic replacement of a failing part
- Dynamic deallocation of failing components so that system availability is maintained

Device drivers

In certain cases, diagnostics are best performed by operating system-specific drivers, most notably I/O devices that are owned directly by a logical partition. In these cases, the operating system device driver often works in conjunction with I/O device microcode to isolate and recover from problems. Potential problems are reported to an operating system device driver, which logs the error. I/O devices can also include specific exercisers that can be invoked by the diagnostic facilities for problem recreation if required by service procedures.

4.4.3 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 send an alert automatically or contact service in the event of a critical system failure. This can be done either from the system itself with Electronic Service Agent (ESA), or from the system in conjunction with the BladeCenter AMM, or from Service and Support Manager (SSM) under SDMC. A hardware fault also illuminates the amber system fault LED (located on the front panel of the blade) to alert the user of an internal hardware problem.

On POWER7 processor-based servers, hardware and software failures are recorded in the system log. An ELA routine analyzes the error, forwards the event to the IVM Service Focal Point (SFP) application (which is either part of Electronic Service Agent running on the blade, or Service and Support Manager under SDMC), 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 or SSM and the BladeCenter AMM, and notifies the system administrator.

After the information is logged in the SFP application or SSM and AMM event log, a call-home service request is initiated. 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, is 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:

- An error code uniquely describing the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data such as engineering and manufacturing levels
- Return codes
- Resource identifiers
- First-failure data capture data

Information about the effect that the repair will have on the system is also included. Error log routines in the operating system can use this information and decide whether to contact service and support, send a notification message, or continue without an alert.

Service Focal Point and Service and Support Manager

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, under the Service Focal Point section of the IVM user interface, is responsible for aggregating duplicate error reports, and ensures that all errors are recorded for review and management on the single blade IVM is controlling.

A similar process is performed by the SDMC. The Problem Analysis component of the SDMC handles the detection and analysis of serviceable events. The Problem Analysis resides within the SSM. The SSM receives the errors directly from the FSP of the managed system and its virtual servers. SDMC handles the detection of duplicate serviceable events and determines if a given condition is a duplicate event.

The first occurrence of each failure type is recorded in **Manage Serviceable Events** in IVM or the **Problems** page under **System Status and Health** on the SDMC. These tasks filter and maintain a history of duplicate reports from other logical partitions or the service processor. They look at all active service event requests, analyzes the failure to ascertain the root cause and, if enabled, initiate a call for service. This methodology ensures that all platform errors are reported through at least one functional path, resulting in a single notification for a single problem.

Note: Because errors are sent to both the Service Focal Point on the blade and to the BladeCenter AMM, more than one call on the same problem can be generated.

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 collected is dependent on the invocation method but includes information such as firmware levels, operating system levels, additional fault isolation register values, recoverable error threshold register values, system status, and any other pertinent data. The data is formatted and prepared for transmission back to IBM to assist with preparing a service action plan for the service representative or for additional analysis.

System dump handling

In certain circumstances, an error might require a dump to be automatically or manually created. A manual dump creation can be done through the AMM Blade Service Data window. A service processor, platform, or partition dump can be initiated for a blade from the AMM. Service processor and platform dumps can be managed and downloaded to a workstation from the IVM Mange Dumps window, under the Service Focal Point.

4.4.4 Notifying the client

After a Power Systems server has detected, diagnosed, and reported an error to an appropriate aggregation point, it notifies 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 dispatched to the client site with the correct replacement part.

Client Notify events

When an event is important enough to report, but does not indicate the need for a repair action or the need to call 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. This includes the following examples:

- Network events such as the loss of contact over a Local Area Network (LAN)
- Environmental events such as ambient temperature warnings
- Events that need further examination by the client, 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 that requires client awareness in the event they want to take further action. These events can be reported back to IBM at the client's discretion.

Call home

A correctly configured POWER processor-based system, SDMC, or BladeCenter AMM can initiate a call from a client location to the IBM service and support organization with error data, server status, or other service-related information. A call home invokes the service organization for the appropriate service action to begin, automatically opening a problem report and, in certain cases, dispatching field support. This automated reporting provides faster and potentially more accurate transmittal of error information. Although configuring a call home is optional, you are strongly encouraged to configure this feature to obtain the full value of IBM service enhancements.

> **Note:** Call home is used generically to indicate automatically contacting IBM service. The actual method is through an Internet connection. BladeCenter AMM, SDMC, and individual blades do not have modem capability.

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 to assist in keeping the firmware and software on the server up-to-date.

The BladeCenter AMM also collects VPD on the individual blades and the components of the BladeCenter chassis. This information is used by support representatives to understand the complete BladeCenter/blade environment.

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.

4.4.5 Locating and servicing parts requiring service

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Packaging for service

The following service enhancements are included in the physical packaging of the systems to facilitate service:

- Color coding (touch points)
  - Terracotta colored touch points indicate that a component (FRU/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

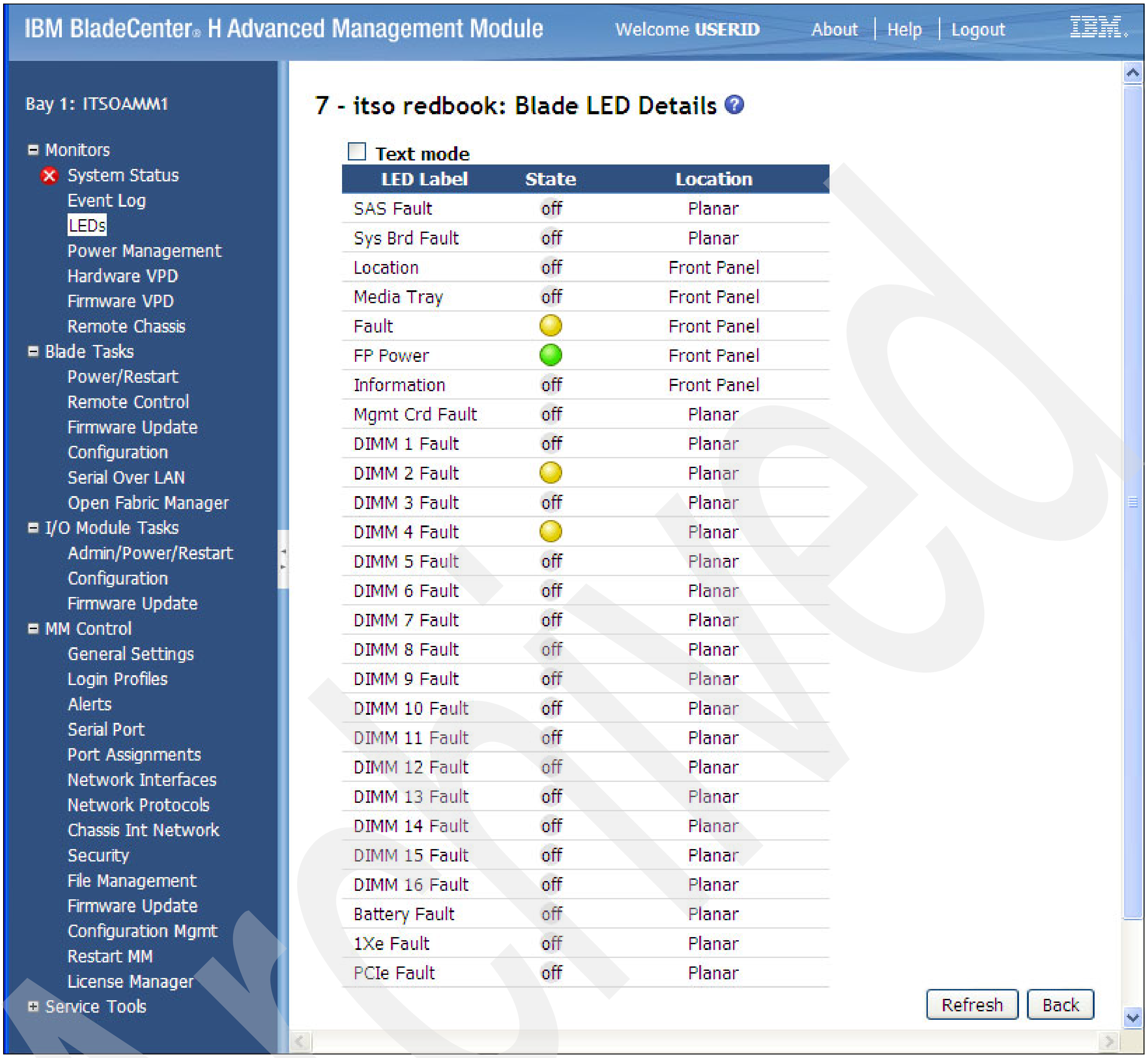
  Selected IBM systems support tool-less or simple tool designs. These designs require no tools or simple tools such as flathead screwdrivers to service the hardware components.

- Positive retention

  Positive retention mechanisms 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 (such as latches, levers, thumb-screws, pop Nylatches (U-clips), and cables) are included to help prevent loose connections and aid in installing (seating) parts correctly. These positive retention items do not require tools.

Light Path

The Light Path LED feature is used for the PS703 and PS704 blades. In the Light Path LED implementation, when a fault condition is detected on the POWER7 processor-based system, a FRU fault LED is illuminated, which is rolled up to the system fault LED. The Light Path system pinpoints the exact part by turning on the FRU fault LED associated with the part to be replaced. The Light Path diagnostic FRU fault LEDs can be reviewed from the AMM (as shown in Figure 4-3) or reviewed directly on the blade after removal from the BladeCenter chassis using the Light Path diagnostic switch on the blade planar.



7 - itso redbook: Blade LED Details

☐ Text mode

LED Label	State	Location
SAS Fault	off	Planar
Sys Brd Fault	off	Planar
Location	off	Front Panel
Media Tray	off	Front Panel
Fault	🟡	Front Panel
FP Power	🟢	Front Panel
Information	off	Front Panel
Mgmt Crd Fault	off	Planar
DIMM 1 Fault	off	Planar
DIMM 2 Fault	🟡	Planar
DIMM 3 Fault	off	Planar
DIMM 4 Fault	🟡	Planar
DIMM 5 Fault	off	Planar
DIMM 6 Fault	off	Planar
DIMM 7 Fault	off	Planar
DIMM 8 Fault	off	Planar
DIMM 9 Fault	off	Planar
DIMM 10 Fault	off	Planar
DIMM 11 Fault	off	Planar
DIMM 12 Fault	off	Planar
DIMM 13 Fault	off	Planar
DIMM 14 Fault	off	Planar
DIMM 15 Fault	off	Planar
DIMM 16 Fault	off	Planar
Battery Fault	off	Planar
1Xe Fault	off	Planar
PCIe Fault	off	Planar

Figure 4-3 AMM blade LED details

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 if the problem is fixed.

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. The following list details several of these service labels and the purpose of each:

- 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 and replace certain serviceable hardware components.
- Numbered arrows are used to indicate the order of operation and the 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 front panel

The front panel LEDs on the PS703 and PS704 blades indicate power status, error and informational states, disk and network activity, and physical location within a BladeCenter chassis.

Concurrent maintenance

The BladeCenter supporting infrastructure is designed with the understanding that certain components have higher intrinsic failure rates than others. The movement of fans, and power supplies make them more susceptible to wearing down or burning out. Other devices (such as I/O modules) might begin to experience wear on mechanical connectors from repeated plugging and unplugging for example, or other unexpected failure. For this reason, these devices are designed to be concurrently maintainable when properly configured.

Live Partition Mobility (LPM) provides the ability to move workload off the PS703 and PS704 blades, allowing uninterrupted service when performing *scheduled* maintenance on a blade. LPM is not considered and should not be used as a high availability method.

Firmware updates

In a BladeCenter/Blade environment there are multiple areas to consider when looking at firmware updates. In some cases, BladeCenter and infrastructure components can be updated concurrently without disrupting blade operations.

POWER processor-based blades require a supported operating system running on the blade to update the system firmware. Starting with the POWER6 generation blades and continuing with the POWER7-based blades, LPM can be used to avoid the disruptive nature of blade firmware updates. Firmware updates can provide fixes to previous versions and can enable new functions. Blade system firmware typically has a prerequisite AMM firmware level.

A regular program of reviewing current firmware levels of the BladeCenter components and the blades should be in place to ensure the best availability.

Firmware updates for the AMM, I/O modules, and blades can be obtained from the IBM Fix Central web page:

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

Repair and verify system

Repair and verify (R&V) is a system used to guide a service provider 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.

The following repair scenarios are covered by R&V:

- 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

R&V 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 R&V procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event, to ensure that the guided maintenance procedures are performed 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. A CD-ROM-based version is also available.

4.5 Manageability

Several functions and tools help manageability, and can allow you to efficiently and effectively manage your system.

4.5.1 Service user interfaces

The Service Interface allows support personnel or the client to communicate with the service support applications in a server and BladeCenter AMM using a console or user interface. 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 following list details the primary service interfaces:

- Light Path
- Blade LED Details
- Service Processor
- Blade front panel
- Operating system service menu
- IVM Service Management
- SDMC Service and Support Manager
- BladeCenter event log
- BladeCenter Service Advisor and Blade LED details

Service processor

The service processor is a controller running its own operating system. It is a component on the blade planar. 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 the main system unit’s state.

The system unit can be in the following states:

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

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, 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 BladeCenter AMM, and enables complete system and partition management from IVM.

The service processor Ethernet port can be enabled by the AMM to an external network and used to access the ASMI. The ASMI can be accessed through an HTTP server that is integrated into the service processor operating environment. This port must also be enabled so that the SDMC can discover the PS703 and PS704 blades for management operations.

Note: The ASMI implementation in the PS703 and PS704 blades does not provide for administrator logins.

Operating system service menu

The system diagnostics consist of stand-alone diagnostics that are loaded from the DVD drive in the BladeCenter media tray, and online diagnostics that are available through the operating system.

Online diagnostics, when installed, are a part of the AIX or VIOS 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. The modes are as follows:

- Service mode

  This mode requires a service mode boot of the system and 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

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

- Maintenance mode

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

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

IVM Service Management

The following functions are available through the IVM Service Management:

- Electronic Service Agent or ESA (for more information see 4.5.3, “Electronic Service Agent and Service and Support Manager” on page 147)
- Service Focal Point
  - Managing serviceable events
  - Service Utilities
    - Create Serviceable Event
    - Manage Dumps
- Collect VPD Information
- Updates (adapter capability only)
- Backup/Restore
- Application Logs
- Monitor tasks
- Hardware Inventory

SDMC Service and Support Manager

The SSM working in conjunction with the SDMC provides a method to list problems and to drill down event status and additional details.

Starting at the highest level of **System Status and Health** → **Problems**

- General, overall problem summary information
- Service
  - Problem summary
  - Transmission summary
- Details
  - Problem number
  - System reference codes
  - Status
  - Reporting/failing MTMs
  - Duplicate count
  - Part number, FRU description, Location
- Recommendation, reference code information
- Support Files, dump files, and so forth, related to the problem

When working on a specific problem these additional actions can be performed:

- Delete
- Ignore
- Submit to IBM
- Repair

BladeCenter event log

The BladeCenter event log includes entries for events that are detected by the BladeCenter unit and installed components.

The following sources can generate events that are recorded in the event log:

- Blade service processor
- BladeCenter unit
- Blade device by bay number

BladeCenter Service Advisor

The BladeCenter Service Advisor provides a method to notify a service and support representative on selected issues. When a serviceable event that has been designated as a call home event is detected, a message is written in the event log and any configured alerts are sent. The information gathered by the service advisor is the same information that is available to the AMM.

4.5.2 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.

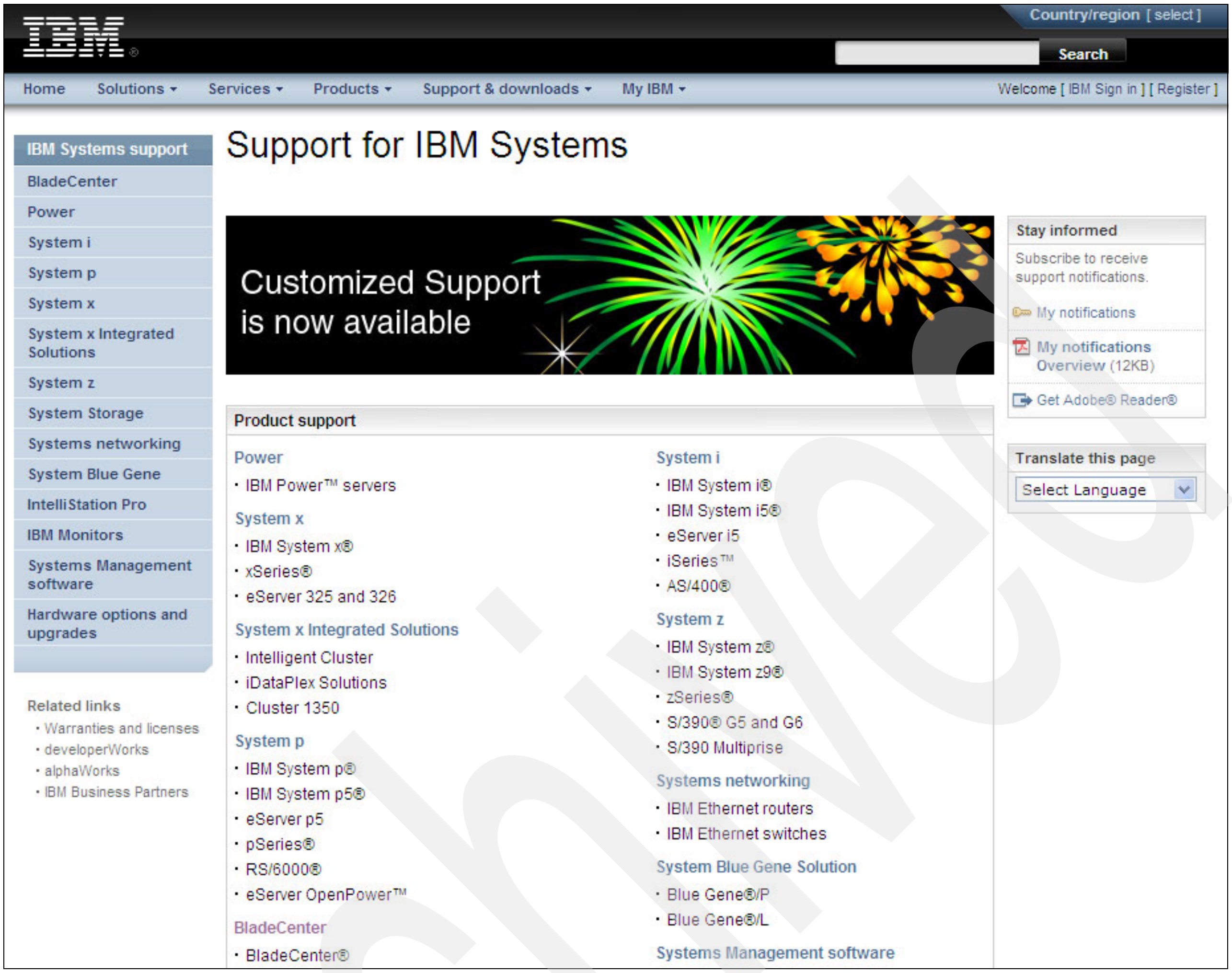
The system firmware consists of service processor microcode, Open Firmware microcode, SPCN microcode, and the POWER Hypervisor. The firmware can be installed from a supported and running operating system on the POWER-based blade or from the SDMC. The firmware can also be installed on a blade when booted from standalone AIX diagnostics.

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 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 to obtain new firmware, see the following web page:

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

For BladeCenter POWER-based blades and BladeCenter modules click the **BladeCenter** link shown Figure 4-4 on page 145.



Support for IBM Systems

Subscribe to receive support notifications.

Product support

Power
- IBM Power™ servers

System x
- IBM System x®
- xSeries®
- eServer 325 and 326

System x Integrated Solutions
- Intelligent Cluster
- iDataPlex Solutions
- Cluster 1350

System p
- IBM System p®
- IBM System p5®
- eServer p5
- pSeries®
- RS/6000®
- eServer OpenPower™

BladeCenter
- BladeCenter®

System i
- IBM System i®
- IBM System i5®
- eServer i5
- iSeries™
- AS/400®

System z
- IBM System z®
- IBM System z9®
- zSeries®
- S/390® G5 and G6
- S/390 Multiprise

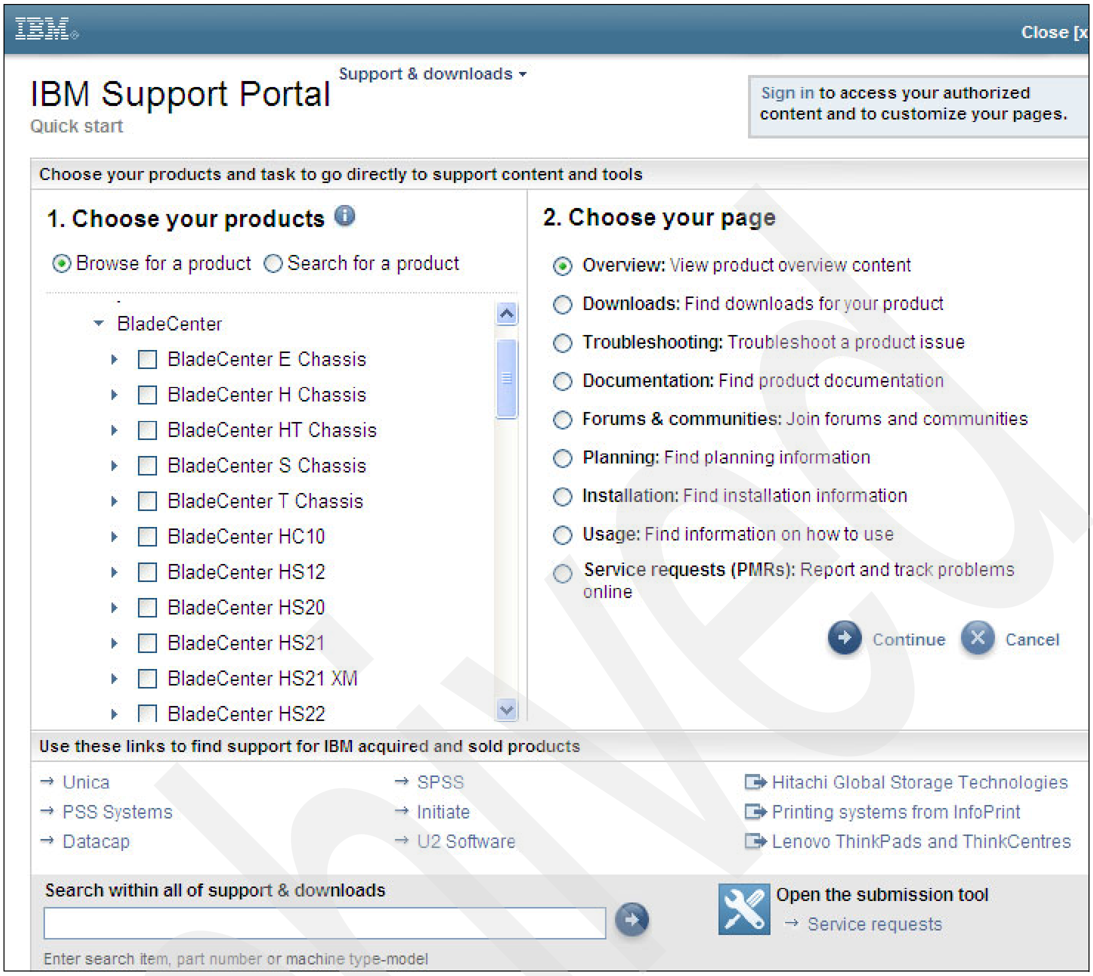
Systems networking
- IBM Ethernet routers
- IBM Ethernet switches

System Blue Gene Solution
- Blue Gene®/P
- Blue Gene®/L

Systems Management software

Figure 4-4 Support for IBM Systems web page

After selecting the BladeCenter link you will be directed the BladeCenter section of the IBM Support Portal (Figure 4-5 on page 146). From this page the specific BladeCenter chassis or blade type and help desired can be selected.



IBM Support Portal

Quick start

Sign in to access your authorized content and to customize your pages.

Choose your products and task to go directly to support content and tools

1. Choose your products

● Browse for a product ○ Search for a product

- BladeCenter
  - □ BladeCenter E Chassis
  - □ BladeCenter H Chassis
  - □ BladeCenter HT Chassis
  - □ BladeCenter S Chassis
  - □ BladeCenter T Chassis
  - □ BladeCenter HC10
  - □ BladeCenter HS12
  - □ BladeCenter HS20
  - □ BladeCenter HS21
  - □ BladeCenter HS21 XM
  - □ BladeCenter HS22

2. Choose your page

● **Overview:** View product overview content

○ **Downloads:** Find downloads for your product

○ **Troubleshooting:** Troubleshoot a product issue

○ **Documentation:** Find product documentation

○ **Forums & communities:** Join forums and communities

○ **Planning:** Find planning information

○ **Installation:** Find installation information

○ **Usage:** Find information on how to use

○ **Service requests (PMRs):** Report and track problems online

Use these links to find support for IBM acquired and sold products

→ Unica	→ SPSS	Hitachi Global Storage Technologies
→ PSS Systems	→ Initiate	Printing systems from InfoPrint
→ Datacap	→ U2 Software	Lenovo ThinkPads and ThinkCentres

Search within all of support & downloads

Enter search item, part number or machine type-model

Open the submission tool

→ Service requests

Figure 4-5 BladeCenter entry into IBM Support Portal

The current running level and boot side (A or B) of the firmware can be displayed from the AMM or the SDMC. The running, temporary, and permanent firmware version levels can also be obtained by using the `lsfware` command on the VIOS.

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 is 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.

For POWER-based blades the installation of system firmware is always disruptive, but the effects can be mitigated by used of Live Partition Mobility.

4.5.3 Electronic Service Agent and Service and Support Manager

IBM has transformed its delivery of hardware and software support services to help you achieve higher system availability. Electronic Services is a web-enabled solution that offers an exclusive, no-additional-charge enhancement to the service and support available for IBM servers. These services provide the opportunity for greater system availability with faster problem resolution and preemptive monitoring. The Electronic Services solution consists of several separate, but complementary, elements:

- Electronic Services news page

  The Electronic Services news page is a single Internet entry point that replaces the multiple entry points that are traditionally used to access IBM Internet services and support. The news page enables you to gain easier access to IBM resources for assistance in resolving technical problems.

- IBM Electronic Service Agent (ESA)

  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.

- Service and Support Manager (SSM)

  SSM is part of the System Director Management Console and available as a plug-in module for IBM Systems Director. ESA is integrated into SSM and provides event monitoring and call home function for SDMC and IBM Systems Director.

> **Note:** SSM installed on IBM Systems Director has more restrictive levels of eligible hardware and operating systems compared to SSM on SDMC. SSM on Systems Director cannot monitor AIX and Linux on POWER partitions managed by a Hardware Management Console (HMC) or IVM. In addition, SSM on Systems Director does not support JS and QS blades.

ESA is available on VIOS, AIX, IBM i, and Linux operating systems. This tool tracks and captures service information, hardware error logs, and performance information. It automatically reports hardware error information to IBM support as long as the system is under an IBM maintenance agreement or within the IBM warranty period. Service information and performance information reporting do not require an IBM maintenance agreement or do not need to be within the IBM warranty period to be reported. Information collected by the Electronic Service Agent tool is available to IBM service support representatives to help diagnose problems.

Early knowledge about potential problems enables IBM to deliver proactive service that can result in higher system availability and performance. In addition, information that is collected through the Service Agent is made available to IBM service support representatives when they help answer your questions or diagnose problems. Installation and use of IBM Electronic Service Agent for problem reporting enables IBM to provide better support and service for your IBM server.

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

`https://www.ibm.com/support/electronic/portal`

4.5.4 BladeCenter Service Advisor

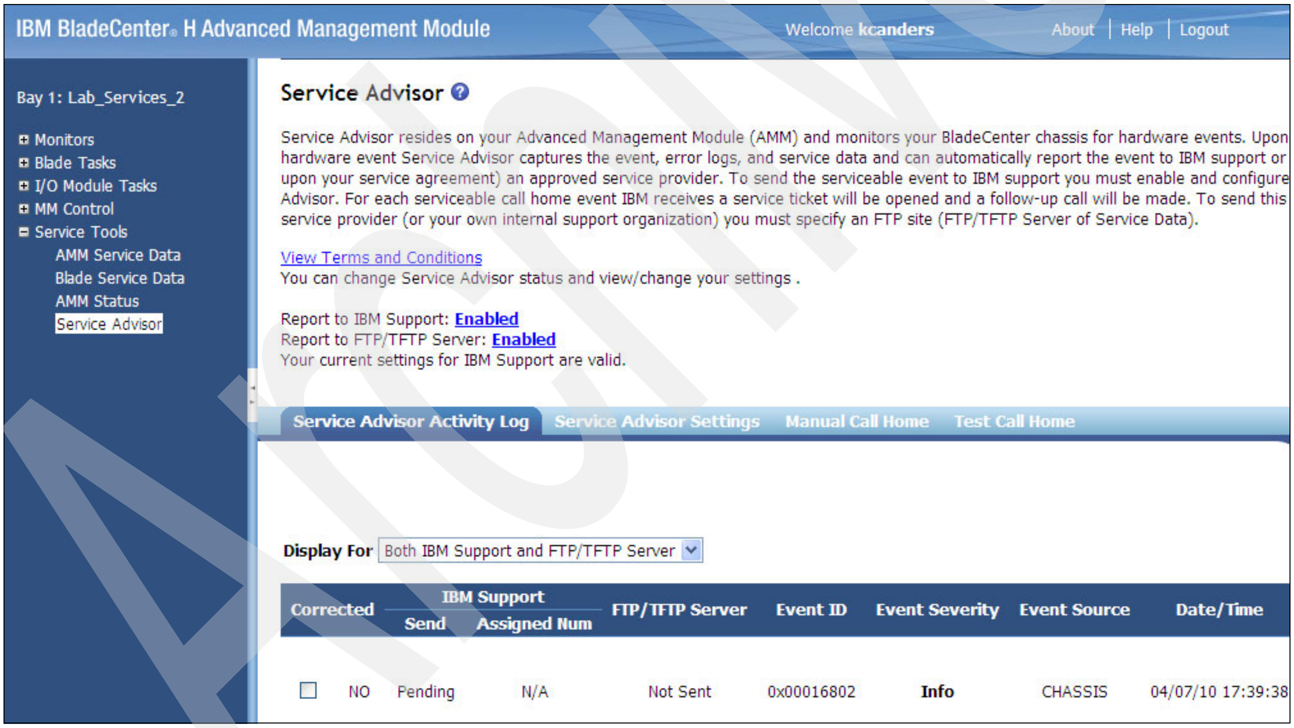
IBM BladeCenter Service Advisor comes standard in all BladeCenter chassis that have an AMM. After being configured and activated, a service event on the BladeCenter chassis can be reported to IBM Service & Support, or to a FTP/TFTP server, or to both.

IBM BladeCenter Service Advisor is built from the IBM Electronic Service Agent offering. There is no installation required for the service advisor, but it must be configured with customer information and enabled.

When a serviceable event designated as a call home event is detected, a message is written in the event log and any configured alerts sent. The information gathered by the service advisor is the same information that is available if you save service data from the advanced management module Web interface.

After gathering the information, the service advisor automatically initiates a call to IBM. Upon receipt of the information, IBM returns a service request ID, which is placed in the call home activity log.

Figure 4-6 shows BladeCenter Service Advisor enabled to send alerts to both IBM Support and a FTP/TFTP server. Service advisor can be tailored through the use of the Call Home Exclusion List to specify particular call home events not to be reported.



Service Advisor

Service Advisor resides on your Advanced Management Module (AMM) and monitors your BladeCenter chassis for hardware events. Upon hardware event Service Advisor captures the event, error logs, and service data and can automatically report the event to IBM support or upon your service agreement) an approved service provider. To send the serviceable event to IBM support you must enable and configure Advisor. For each serviceable call home event IBM receives a service ticket will be opened and a follow-up call will be made. To send this service provider (or your own internal support organization) you must specify an FTP site (FTP/TFTP Server of Service Data).

View Terms and Conditions
You can change Service Advisor status and view/change your settings .

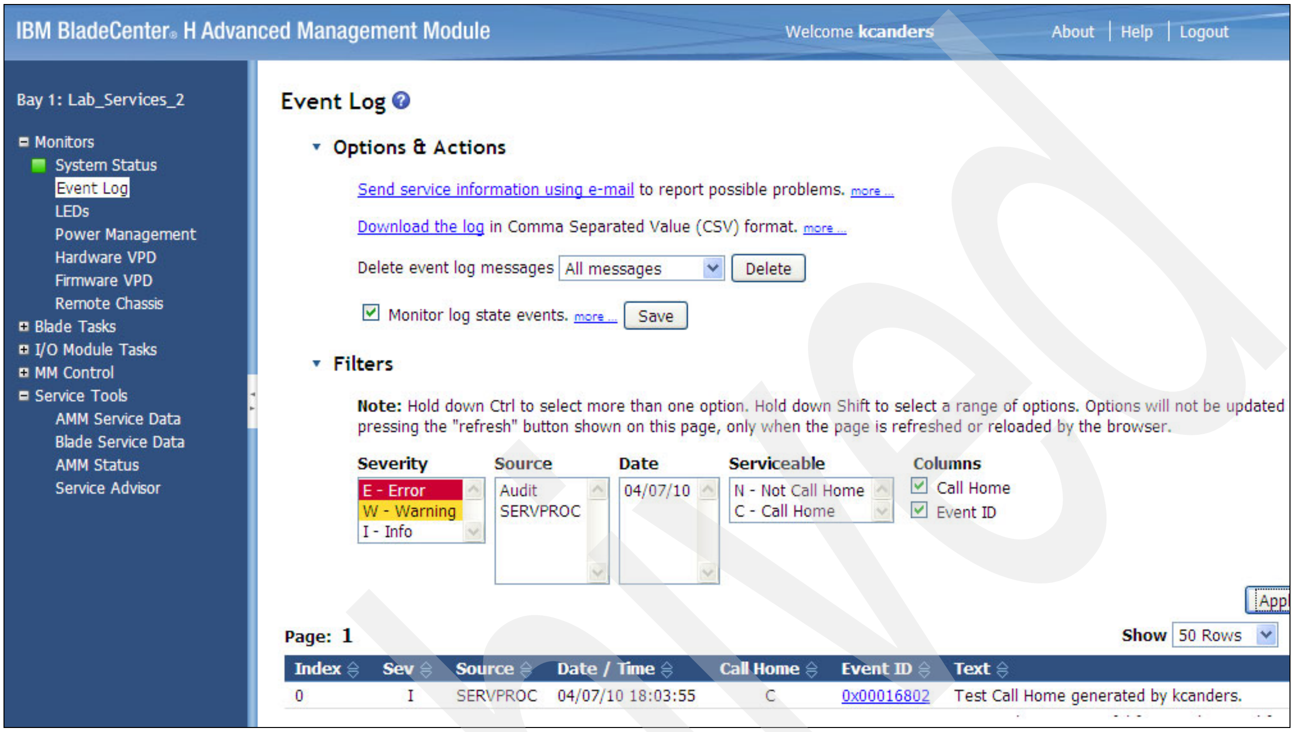
Report to IBM Support: **Enabled**
Report to FTP/TFTP Server: **Enabled**
Your current settings for IBM Support are valid.

Display For Both IBM Support and FTP/TFTP Server

Corrected	IBM Support		FTP/TFTP Server	Event ID	Event Severity	Event Source	Date/Time
	Send	Assigned Num					
☐ NO	Pending	N/A	Not Sent	0x00016802	**Info**	CHASSIS	04/07/10 17:39:38

Figure 4-6 BladeCenter Service Advisor

On the Event Log page of the advanced management module web interface, you can select the Display Call Home Flag checkbox. If you select the checkbox, events are marked with a **C** for call home events and an **N** for events that are not called home. In addition, you can filter the event log view based on this setting. Figure 4-7 shows the BladeCenter event log depicting a call home event.



IBM BladeCenter® H Advanced Management Module

Welcome **kcanders**

About | Help | Logout

Bay 1: Lab_Services_2

- ⊟ Monitors
  - 🟩 System Status
  - Event Log
  - LEDs
  - Power Management
  - Hardware VPD
  - Firmware VPD
  - Remote Chassis
- ⊞ Blade Tasks
- ⊞ I/O Module Tasks
- ⊞ MM Control
- ⊟ Service Tools
  - AMM Service Data
  - Blade Service Data
  - AMM Status
  - Service Advisor

Event Log

▾ Options & Actions

Send service information using e-mail to report possible problems. more ...

Download the log in Comma Separated Value (CSV) format. more ...

Delete event log messages [All messages] [Delete]

☑ Monitor log state events. more ... [Save]

▾ Filters

Note: Hold down Ctrl to select more than one option. Hold down Shift to select a range of options. Options will not be updated pressing the "refresh" button shown on this page, only when the page is refreshed or reloaded by the browser.

Severity	Source	Date	Serviceable	Columns
E - Error	Audit	04/07/10	N - Not Call Home	☑ Call Home
W - Warning	SERVPROC		C - Call Home	☑ Event ID
I - Info				

[Appl

Page: 1

Show [50 Rows]

Index	Sev	Source	Date / Time	Call Home	Event ID	Text
0	I	SERVPROC	04/07/10 18:03:55	C	0x00016802	Test Call Home generated by kcanders.

Figure 4-7 BladeCenter event log showing call home event

5

Systems Director Management Console

The Systems Director Management Console (SDMC) is the successor to the Hardware Management Console (HMC) and the Integrated Virtualization Manager (IVM).

This chapter highlights some of the new concepts, terminology, features, and functionality of the SDMC and how they relate to POWER6- and POWER7-based blades. Installation and setup of the SDMC is outside the scope of this paper and will not be covered in this chapter.

For a comprehensive view of the SDMC consult *Systems Director Management Console Introduction and Overview*, SG24-7860.

This chapter covers the following topics:

5.1 SDMC Introduction

The SDMC is an extension to IBM Systems Director. The SDMC is available in two appliance versions, hardware and software. The SDMC is designed to replace both the HMC and IVM for POWER-based systems management. POWER-based blades managed by SDMC have their features and functionality brought into line with other POWER-based systems that have been managed by HMC in the past. The use of a this new management console also brings a new set of terminology; these terms are covered in this section.

The SDMC is designed to be integrated into the administrative framework of IBM Systems Director and has the same look and feel. It provides a common interface for systems administration across the data center.

5.1.1 Hardware appliance

The hardware appliance is required for midrange and high-end Power Systems, but can also manage low-end systems.

The hardware appliance consists of a virtual image that resides on a host system running Red Hat Linux. Login to the host system or access to the hypervisor is not permitted. All communication is accomplished through a special channel from the guest to the host.

5.1.2 Software appliance

The SDMC software appliance can be installed in an existing x86 virtualized infrastructure. Virtualization hypervisor options include VMware (vSphere server and ESXi Version 4 or later; at least Version 4.1 is required if USB support is desired) and Red Hat Enterprise Linux (RHEL) Enterprise KVM Version 5.5 or later.

The supported host operating systems and hypervisors are:

- Red Hat Enterprise Linux 5.5 with KVM (kvm-83-164.el5) or later; requires at least one network bridge.
- VMware ESXi 4.01or later.
- VMware ESX 4 or later.

The VMware hypervisor requires another machine to be configured with either:

- Windows XP with VMware Virtual Infrastructure Client or VMware OVF Tool installed
- Linux (preferably Red Hat Enterprise Linux 5.5) with VMware OVF Tool installed

5.1.3 IVM, HMC, and SDMC support

The SDMC can replace both HMC and IVM management methods, or it can work in conjunction with an HMC during a transition. Table 5-1 and Table 5-2 show the current system management types and the possible SDMC appliance type they can support. There is no SDMC support for POWER5 and earlier and there is currently no plan to support HMC or IVM on platforms beyond POWER7.

Table 5-1 POWER6 Support by SDMC

POWER6 models	Machine types	HMC or IVM	Systems Director Management Console
Rack systems			
595	9119-FHA	HMC	Hardware appliance only
575	9125-F2A	HMC	HMC Only
570	9117-MMA	HMC	Hardware appliance only
570	9406-MMA	HMC	Hardware appliance only
560	8234-EMA	HMC	Hardware appliance only
550	8204-E8A	HMC or IVM	Hardware or software appliance
550	9409-M50	HMC or IVM	Hardware or software appliance
520	8203-E4A	HMC or IVM	Hardware or software appliance
520	8203-E4A	HMC or IVM	Hardware or software appliance
520-SB	8261-E4A	HMC or IVM	Hardware or software appliance
520	9408-M25	HMC or IVM	Hardware or software appliance
520	9407-M15	HMC or IVM	Hardware or software appliance
Blade systems			
JS22	7998-61X	IVM	Hardware or software appliance
JS12	7998-60X	IVM	Hardware or software appliance
JS23/43	7778-23X	IVM	Hardware or software appliance

Table 5-2 POWER7 Support by SDMC

POWER7 models	Machine types	HMC or IVM	Systems Director Management Console
Rack systems			
795	9119-FHB	HMC	Hardware appliance only
780	9179-MHB	HMC	Hardware appliance only
770	9117-MMB	HMC	Hardware appliance only
755	8236-E8C	HMC	Hardware or software appliance
750	8233-E8B	HMC or IVM	Hardware or software appliance
720	8202-E4B	HMC or IVM	Hardware or software appliance
740	8205-E6B	HMC or IVM	Hardware or software appliance
710/730	8231-E2B	HMC or IVM	Hardware or software appliance
Blade systems			
PS700	8406-70Y	IVM	Hardware or software appliance
PS701/702	8406-71Y	IVM	Hardware or software appliance
PS703	7891-73X	IVM	Hardware or software appliance
PS704	7891-74X	IVM	Hardware or software appliance

5.1.4 Terminology

SDMC uses the terminology that has been established in IBM Systems Director. Many of the terms used in the HMC and IVM context have changed with SDMC. Table 5-3 correlates some of the more common HMC and IVM terms with those used for SDMC.

Table 5-3 SDMC Comparison of terms

HMC/IVM terminology	SDMC terminology	Additional notes/where used
Managed System	Server	-
LPAR/Logical Partition	Virtual Server	-
DLPAR	Manage Virtual Server	-
Systems Management	Power Systems Resources	Navigation Menu item
Servers	Hosts	Navigation Menu item
Server names	Server names	Navigation Menu item
System Plans	System Configuration → Manage System Plans	System plans are not supported on POWER based blades
HMC Management	Settings tab	-
Service Management	System Status and Health → Problems	Navigation Menu item
Updates	Release management → Power Firmware Management	right click server name, used for system firmware updates
Updates	Release management → Updates	Navigation Menu item
not applicable	Operating System	Systems Director function of managing an operating system
Frame / BPA	Power Unit	Does not apply to blades

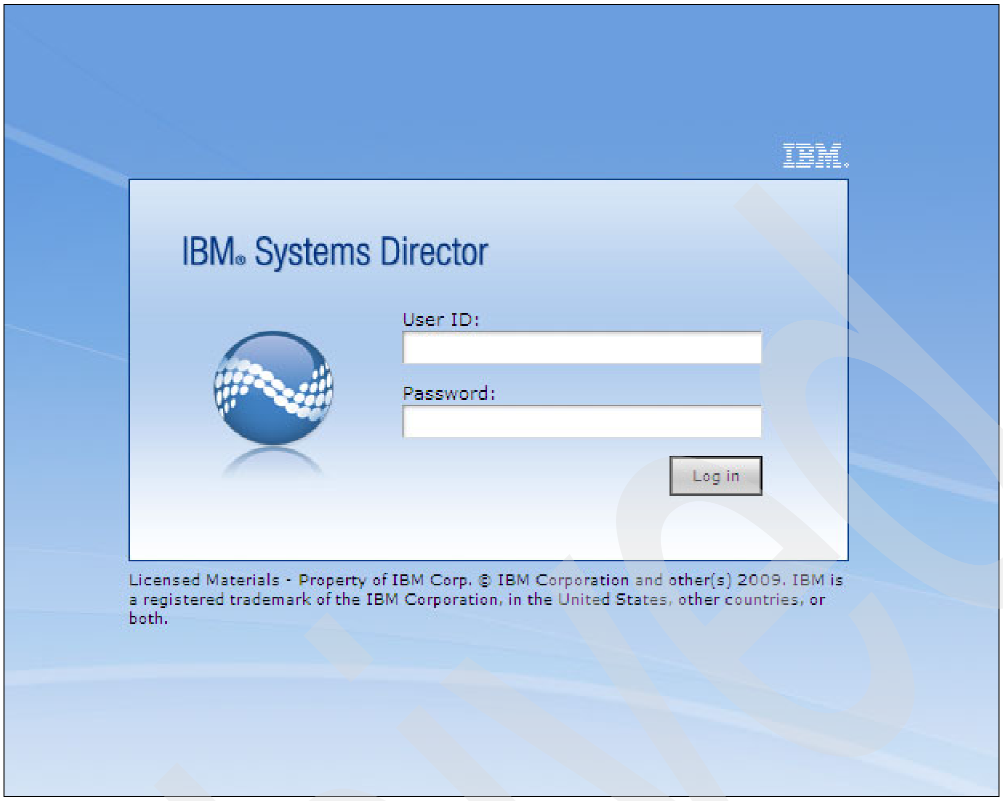
5.2 Using the web interface

SDMC provides a web interface similar to that of the HMC. Once the SDMC appliance is powered up, you should see the SDMC login window.

You can also access SDMC remotely using a browser. The supported browsers are the same as IBM Systems Director and are currently Firefox versions 3, 3.5, and 3.6, and Internet Explorer versions 7 and 8. For logging in remotely, open the browser and point the browser to the following URL (where `system_name` is the host name or IP address of the SDMC system):

```
https://system_name
```

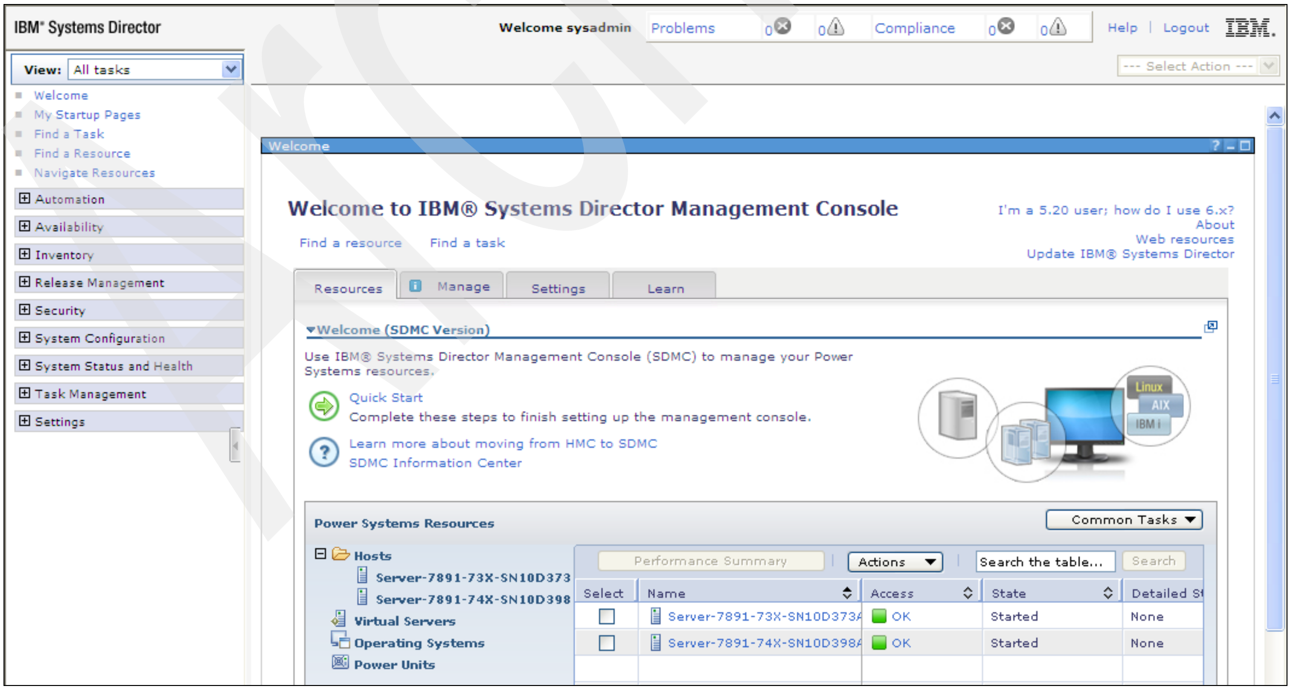
A login page opens as shown in Figure 5-1 on page 155.



Licensed Materials - Property of IBM Corp. © IBM Corporation and other(s) 2009. IBM is a registered trademark of the IBM Corporation, in the United States, other countries, or both.

Figure 5-1 SDMC login

Enter the user ID and password that corresponds to an authorized SDMC user and click **Log in**. The Welcome page shown in Figure 5-2 will display after logging in successfully to the SDMC.



Welcome to IBM® Systems Director Management Console

Welcome (SDMC Version)

Use IBM® Systems Director Management Console (SDMC) to manage your Power Systems resources.

Quick Start
Complete these steps to finish setting up the management console.

Learn more about moving from HMC to SDMC
SDMC Information Center

Power Systems Resources

- Hosts
  - Server-7891-73X-SN10D373
  - Server-7891-74X-SN10D398
- Virtual Servers
- Operating Systems
- Power Units

Select	Name	Access	State	Detailed St
□	Server-7891-73X-SN10D373A	🟩 OK	Started	None
□	Server-7891-74X-SN10D398A	🟩 OK	Started	None

Figure 5-2 SDMC Welcome page

SDMC has the same navigation method on the left side of the user interface as IBM Systems Director, and essentially the same base functionality.

After the login process, the SDMC begins with the SDMC **Welcome** page. This page has two additional tabs when compared to Systems Director, **Resources** and **Settings**.

5.2.1 Resources tab

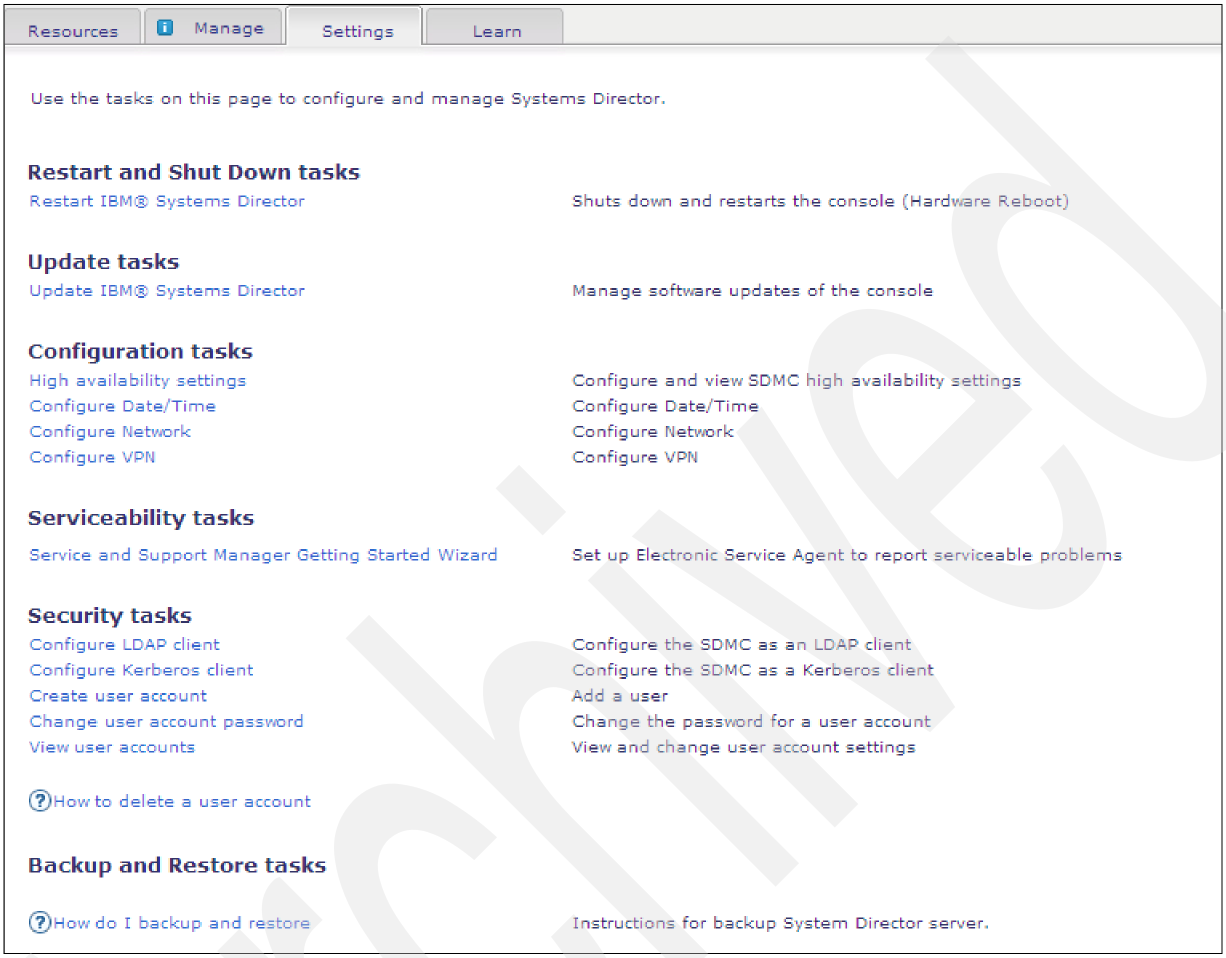
The Welcome option from the navigation area of the SDMC can be selected at any time and will return you to the Resources or SDMC main page as shown in Figure 5-2 on page 155.

The Resource tab provides a dashboard of all the servers, virtual servers, power units, and operating systems that your SDMC is currently managing. The properties for a selected resource are displayed and tasks can be selected using the context menu available by right-clicking the object. The Welcome page only shows the power resources that you will be managing using the SDMC.

There could be other non-Power Systems that are discovered and managed by SDMC. You can see these systems in the Navigate Resources page available from the navigation area. The Welcome page has certain columns that are not available in the Navigate Resources page. These columns hold data specific to Power resources that the SDMC is managing.

5.2.2 Settings tab

The Settings tab provides the interface to the tasks that are related to managing your SDMC appliance itself. These tasks are shown in Figure 5-3.



Use the tasks on this page to configure and manage Systems Director.

Restart and Shut Down tasks

Restart IBM® Systems Director	Shuts down and restarts the console (Hardware Reboot)

Update tasks

Update IBM® Systems Director	Manage software updates of the console

Configuration tasks

High availability settings	Configure and view SDMC high availability settings
Configure Date/Time	Configure Date/Time
Configure Network	Configure Network
Configure VPN	Configure VPN

Serviceability tasks

Service and Support Manager Getting Started Wizard	Set up Electronic Service Agent to report serviceable problems

Security tasks

Configure LDAP client	Configure the SDMC as an LDAP client
Configure Kerberos client	Configure the SDMC as a Kerberos client
Create user account	Add a user
Change user account password	Change the password for a user account
View user accounts	View and change user account settings

?How to delete a user account

Backup and Restore tasks

?How do I backup and restore	Instructions for backup System Director server.

Figure 5-3 Settings tab tasks

You can click any of the tasks, which opens a new page that will request input to change or manage the desired setting.

5.3 POWER-based blades management

The traditional way to manage POWER-based blade servers is through the use of Integrated Virtualization Manager (IVM). The introduction of the SDMC provides new capability to these blades. This section highlights some of these new features.

5.3.1 IVM characteristics

POWER-based blades have the option to either run a single operating system (native) or multiple operating systems in logical partitions or LPARs. Under IVM-managed blades, when multiple LPARs are desired, the Virtual I/O Server (VIOS) is required to be installed first. Part of the VIOS is the IVM. IVM provides the interface to the hypervisor in much the same way the Hardware Management Console (HMC) functions with a typical rack-based POWER system.

Characteristics of an IVM-managed system are:

- Runs on top of a Virtual I/O Server
- Administers POWER-based blade servers and entry POWER servers
- Administers only one managed system per IVM
- Allows for only one VIOS installed per system
- System firmware updates require installed operating system and are always disruptive
- No concept of LPAR profiles; single configuration for each LPAR
- Virtual console only to the VIOS or client LPARs on the same blade

The Advanced Management Module (AMM) of the BladeCenter chassis also plays a role in the basic control and management of a POWER-based blade. The typical AMM functions related to a POWER-based blade are:

- Blade power up/down/restart
- Select boot device order
- Select temp/perm firmware side for boot
- Console support through Serial Over LAN (SOL)
- Blade Service Data (SRC reporting)
- Service Advisor (problem reporting)

5.3.2 SDMC added capability for POWER-based blades

The SDMC can consolidate most of the AMM and IVM functions and adds new capability when used with POWER-based blades. The SDMC interfaces directly with the service processor and hypervisor on the blade.

New capabilities added by the SDMC in a POWER-based blade environment are:

- Management of multiple servers from a single point
- Dual VIOS servers
- Multiple profiles can be created for each virtual server
- System firmware management
- Active Memory Expansion
- Suspend/Resume virtual servers
- Multiple shared processor pools
- Centralized problem reporting from Service and Support Manager
- Open virtual console to any virtual server on any blade

The SDMC also consolidates or replaces the BladeCenter Advanced Management Module (AMM) functions to a single instance for the following features:

- SRC codes displayed by SDMC
- Better coordination between Service Advisor and Service and Support Manager for POWER based blades
- Server and virtual server power and start/restart control

5.4 IVM to SDMC transition

This section describes the transition of managed systems from the IVM environment to the SDMC environment. HMC to SDMC transitions can also be performed but are not discussed here.

5.4.1 IVM to SDMC transition process

The transition process is performed manually for the IVM to SDMC transition. The transition wizard available for HMC to SDMC moves cannot be used for IVM to SDMC transitions. Prior to the transition the managed system has to be in an IVM-managed state. After the transition is complete, you are not able to use the IVM user interface, because the Virtual Management Channel (VMC) is deactivated. The VMC is the device on the Virtual I/O Server that enables a direct hypervisor configuration. This device is activated only when the Virtual I/O Server detects that the environment has to be managed by IVM.

What information is transitioned

When you transition managed systems from IVM to SDMC, only the following information is transitioned:

- Managed system information.
- Virtual server information is automatically retrieved from the managed system after a request access to the managed system is successful.

How to transition

The following steps are used to transition a managed system from IVM to SDMC:

1. Discover the managed system.
2. Request access to the managed system.
3. Update the ASM passwords.

When the access state of the system changes to OK, additional tasks are now available on the context menu for managing the system. The virtual servers hosted by the system are retrieved and listed on the Welcome page. Information on the actual steps required to perform the transition can be found in 5.4.2, “Steps to SDMC management of POWER-based blades” on page 160.

This transition performs the following operations internally:

- Fetches the virtual server configuration information from the hypervisor.
- Updates the SDMC with the managed system information from the hypervisor.
- The virtual slots ranging from 2 to 10 are reserved in SDMC. The IVM does not have such restrictions, and it is possible that you were using some virtual slots in the reserved range in the IVM environment. The transition process looks for such virtual slots and dynamically adjusts the virtual slots range to a new available range within the maximum slots value specified.

> **Note:** If the automatic readjustment of the reserved virtual slot range is unsuccessful, then you have to shut down and activate the Virtual I/O Server again to use the advanced PowerVM features. The shutdown and reboot are needed only when you want to use the advanced PowerVM features; the virtual servers continue to run normally even without the shutdown and reboot.

- Creates default profiles based on virtual server current configuration read from PHYP because IVM does not support profiles.

Messages

By default, all success and failure messages are listed in the Event Log page. The Event Log page is available under the System Status and Health category in the navigation area. Check the Event Log page for the success or failure of the transition process. When the transition completes successfully, the following message is posted in the Event Log:

```
The IVM to SDMC transition completed successfully
```

The Status Manager displays the Alert and Resolution messages. Clicking the **Health and Summary** link under the System Status and Health category shows the messages in the Status Manager. An Alert message (error) is displayed first, followed by a Resolution event.

An alert message is displayed in the Status Manager to flag a problem. A Resolution event is received by the Status Manager when the problem is resolved. Thus, the Resolution event removes the corresponding alert that it has resolved from the Status Manager.

Table 5-4 shows the Alert and Resolution events related to the IVM to SDMC transition.

Table 5-4 Status Manager Error/Resolution messages

Alert	Resolution
Virtual I/O Server Slot Range Adjustment Failed (Error).	Virtual I/O Server Reactivated & Slot Adjustment Completed Successfully (Resolution).
The Virtual I/O Server Slot adjustment did not complete successfully. You have to reactivate the Virtual I/O Server virtual server before attempting LPM or AMS operations.	The Virtual I/O Server Slot Adjustment completed successfully.

5.4.2 Steps to SDMC management of POWER-based blades

The initial steps when setting up SDMC and POWER-based blades to work together are very similar to the way IBM Systems Director works with managed endpoints. One additional step on the BladeCenter AMM is required to allow the Flexible Service Processor (FSP) on the blade to communicate on the network.

The basic steps are:

1. Expose the POWER-based blade FSP to the network
2. Discovery of the server to be managed
3. Request access to the server FSP
4. Update access passwords
5. Collect inventory of the managed system

When these steps are completed the POWER-based blade or server will be ready for management by the SDMC.

These steps are described in detail in the following sections using a PS703 as an example.

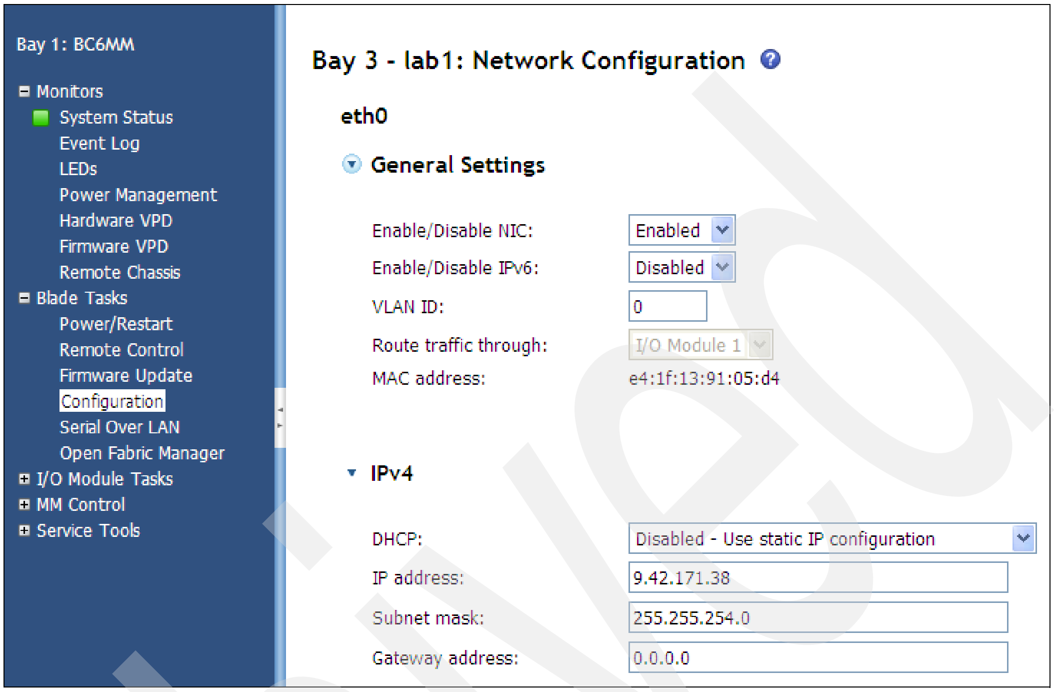
5.4.3 FSP access configuration from BladeCenter AMM

In the default configuration the flexible service processor or FSP in POWER-based blades does not have an IP address assigned, and it is not available on the network.

SDMC interfaces directly with the FSP and requires network connectivity. This connectivity is established by exposing the FSP on the network using configuration changes made on the Advanced Management Module (AMM).

1. From the AMM user interface select **Blade Tasks** → **Configuration** from the navigation area of the AMM, then select the **Management Network** tab.

2. Click the desired blade from the list displayed. Figure 5-4 shows the Network Configuration information and settings that relate to the FSP, allowing it to be present on the network.



Bay 3 - lab1: Network Configuration

eth0

General Settings

Enable/Disable NIC: Enabled
Enable/Disable IPv6: Disabled
VLAN ID: 0
Route traffic through: I/O Module 1
MAC address: e4:1f:13:91:05:d4

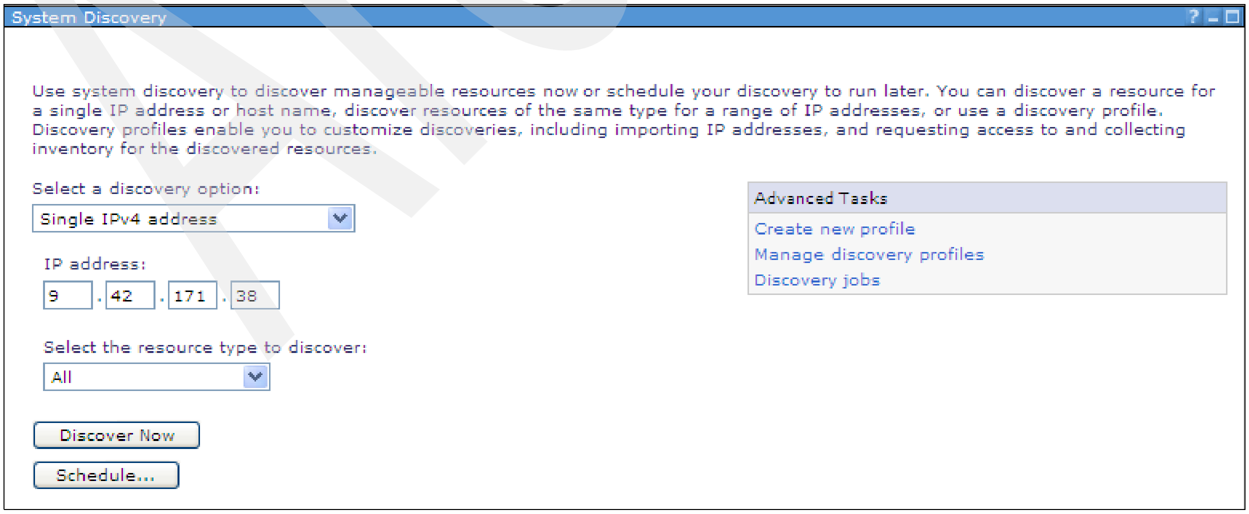
IPv4

DHCP: Disabled - Use static IP configuration
IP address: 9.42.171.38
Subnet mask: 255.255.254.0
Gateway address: 0.0.0.0

Figure 5-4 Setting FSP IP address

5.4.4 Discovery

The server discovery process used by SDMC is the same as Systems Director and is initiated from the navigation section by selecting **Inventory** → **System Discovery**. Figure 5-5 shows the user interface page used to discover a PS703 blade. The IP address that was assigned to the FSP from the AMM is used in the discovery process. Discovery can also use DNS names, ranges of IP addresses, or a discovery profile. These are all standard discovery methods from Systems Director.



Use system discovery to discover manageable resources now or schedule your discovery to run later. You can discover a resource for a single IP address or host name, discover resources of the same type for a range of IP addresses, or use a discovery profile. Discovery profiles enable you to customize discoveries, including importing IP addresses, and requesting access to and collecting inventory for the discovered resources.

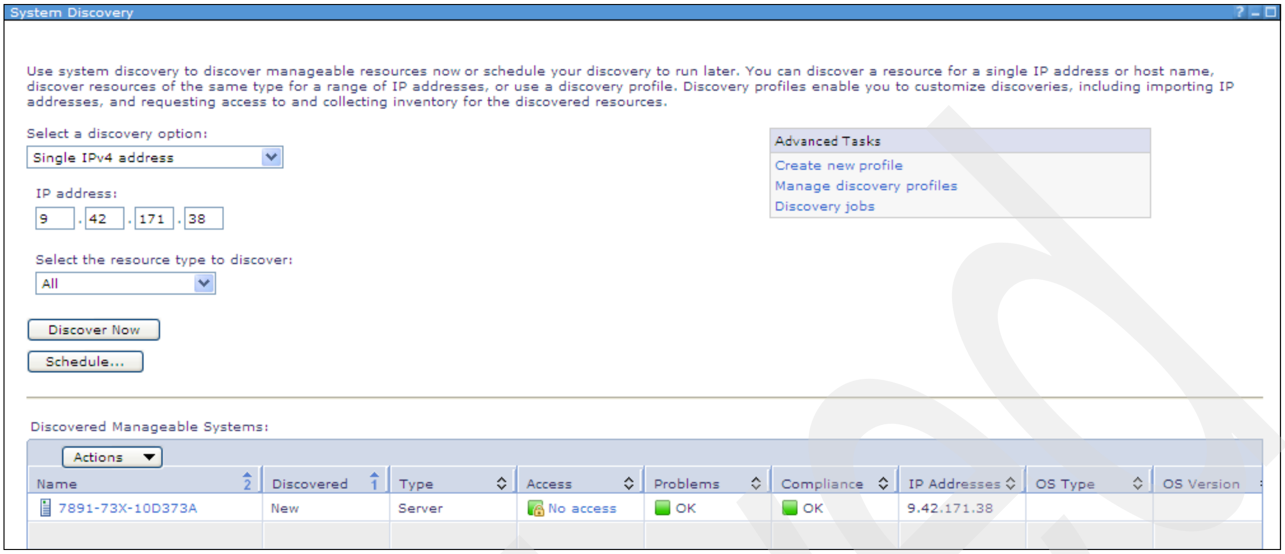
Select a discovery option: Single IPv4 address

IP address: 9 . 42 . 171 . 38

Select the resource type to discover: All

Figure 5-5 FSP discovery setup

Figure 5-6 shows the PS703 server after the discovery process has been completed. After discovery the next step is to request access.



Use system discovery to discover manageable resources now or schedule your discovery to run later. You can discover a resource for a single IP address or host name, discover resources of the same type for a range of IP addresses, or use a discovery profile. Discovery profiles enable you to customize discoveries, including importing IP addresses, and requesting access to and collecting inventory for the discovered resources.

Select a discovery option: Single IPv4 address

IP address: 9 . 42 . 171 . 38

Select the resource type to discover: All

Discovered Manageable Systems:

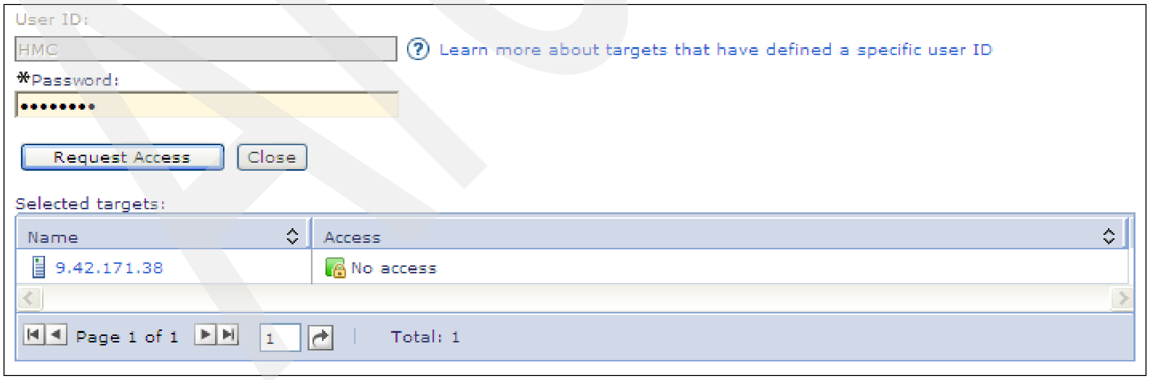
Name	Discovered	Type	Access	Problems	Compliance	IP Addresses	OS Type	OS Version
7891-73X-10D373A	New	Server	No access	🟩 OK	🟩 OK	9.42.171.38		

Figure 5-6 PS703 FSP after initial discovery

5.4.5 Request Access to Server

After discovery but before the SDMC can start managing the server, access must be requested. Initial access is requested by clicking the **No access** link shown in Figure 5-6. This link will take you to the page that allows you to enter an HMC password as shown in Figure 5-7. If you have not set a password previously, you can enter a new one here which will be used from now on.

Tip: Remember the password for the initial HMC access request! If a server is removed from the SDMC and later added back, the same password that was set during the initial access request must be reused.


User ID: HMC

*Password: ••••••••

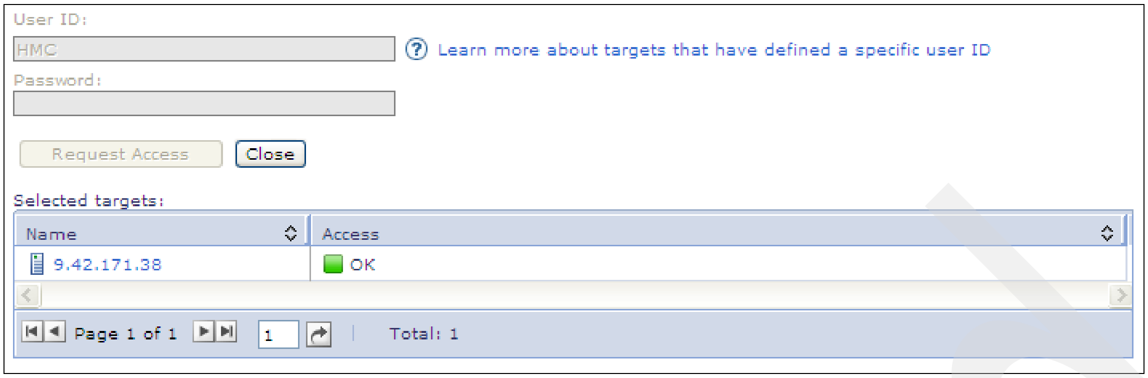
Learn more about targets that have defined a specific user ID
Request Access
Close

Selected targets:

Name	Access
9.42.171.38	No access

Figure 5-7 Initial access request to PS703 FSP

After requesting access you should see Access identified as **OK** as shown in Figure 5-8 on page 163.


User ID:

HMC

Learn more about targets that have defined a specific user ID

Password:

Selected targets:

Name	Access
9.42.171.38	🟩 OK

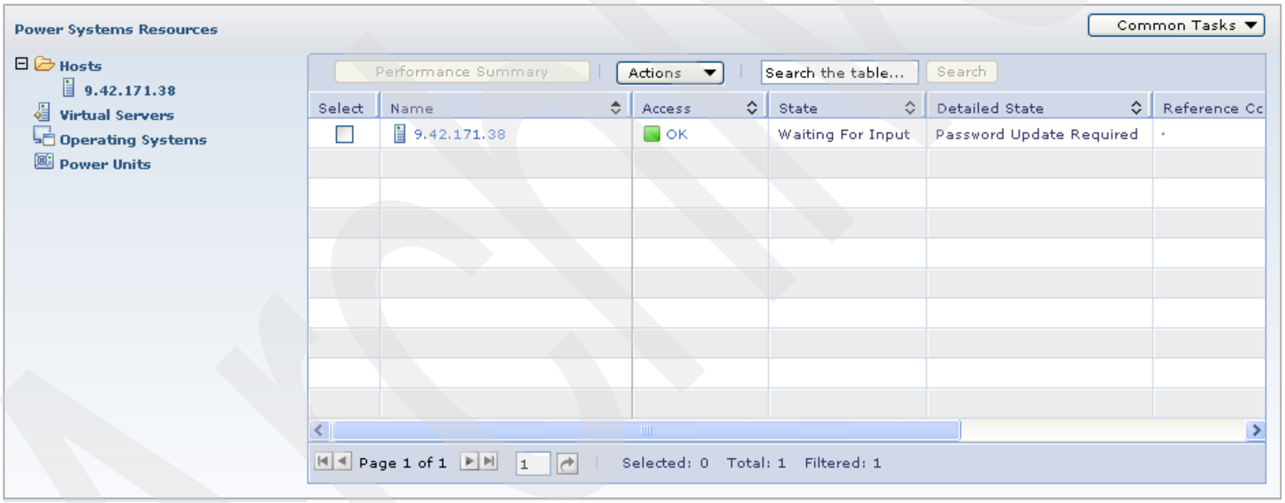
Page 1 of 1 | 1 | Total: 1

Figure 5-8 Successful access request to FSP on PS703

After access has been obtained, the request page and the discovery page can be closed.

5.4.6 Updating ASM passwords

After access has been requested you can return to the SDMC main view and review the newly added server as show in Figure 5-9. The name is the same as the discovery method, IP address in this case. Notice the **Detailed State** is Password Update Required.

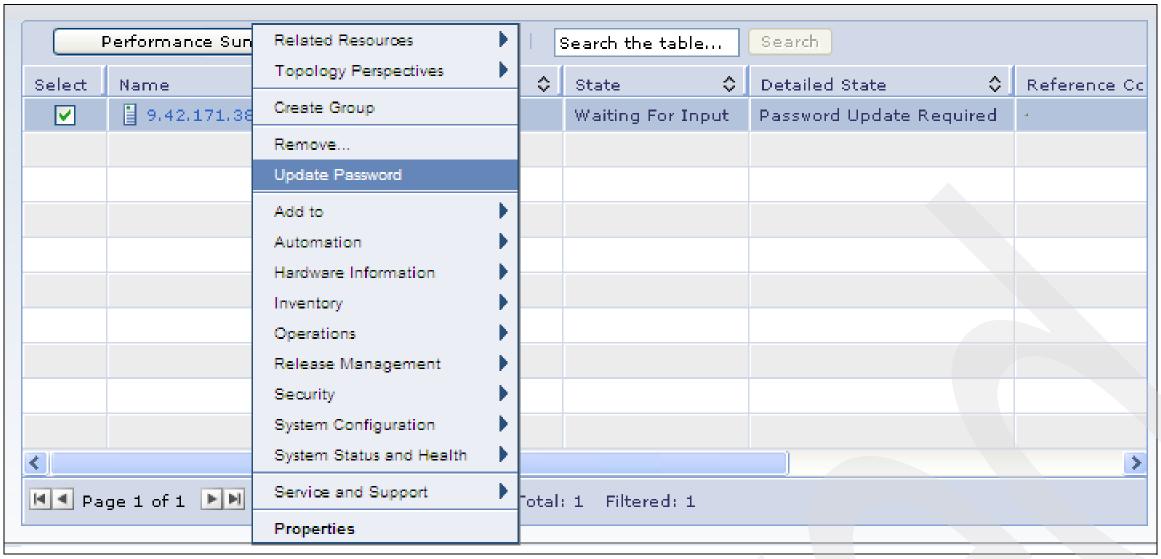


Select	Name	Access	State	Detailed State	Reference Cc
□	9.42.171.38	🟩 OK	Waiting For Input	Password Update Required	-

Page 1 of 1 | 1 | Selected: 0 Total: 1 Filtered: 1

Figure 5-9 New server displayed in SDMC

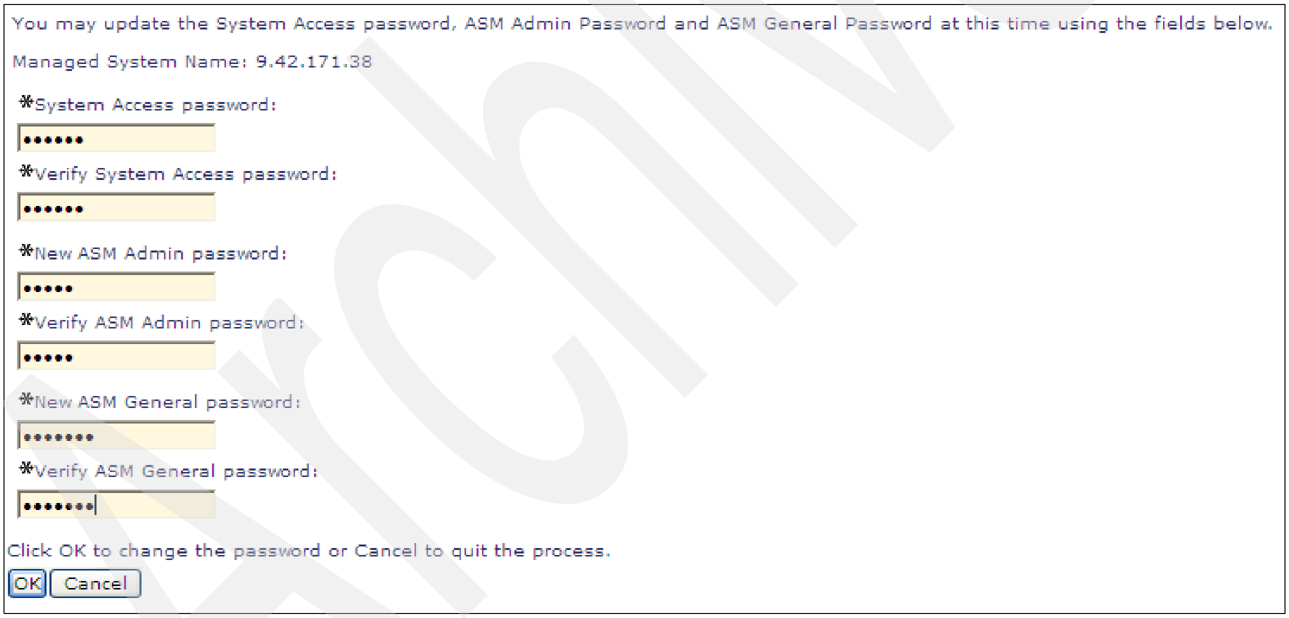
Before you can start managing a server the Advanced System Management (ASM) passwords must be updated. The update process is started by right-clicking the server name ant then selecting **Update Password** as shown in Figure 5-10 on page 164.



Select	Name	State	Detailed State	Reference Cc
☑	9.42.171.38	Waiting For Input	Password Update Required	-

Figure 5-10 Starting ASM/FSP password update process

Figure 5-11 show the page used to set/update the various ASM passwords. After the password values are entered click **OK**.



You may update the System Access password, ASM Admin Password and ASM General Password at this time using the fields below.

Managed System Name: 9.42.171.38

*System Access password: ••••••

*Verify System Access password: ••••••

*New ASM Admin password: •••••

*Verify ASM Admin password: •••••

*New ASM General password: •••••••

*Verify ASM General password: •••••••

Click OK to change the password or Cancel to quit the process.

Figure 5-11 Updating/Setting ASM passwords

The discovery and access steps are now complete.

5.4.7 Systems Director inventory collection

One of the most important tasks to be completed on a newly discovered and accessed object is the core Systems Director function of collecting inventory information on a new object. Inventory collection by SDMC can be initiated by right-clicking the server object displayed by the SDMC, as shown in Figure 5-12 on page 165. Also note the server object name is now represented by a Systems Director classification (Server), model and type (7891-73X), and serial number. The change in the name is the result of the previous step, which allows full access to the FSP.

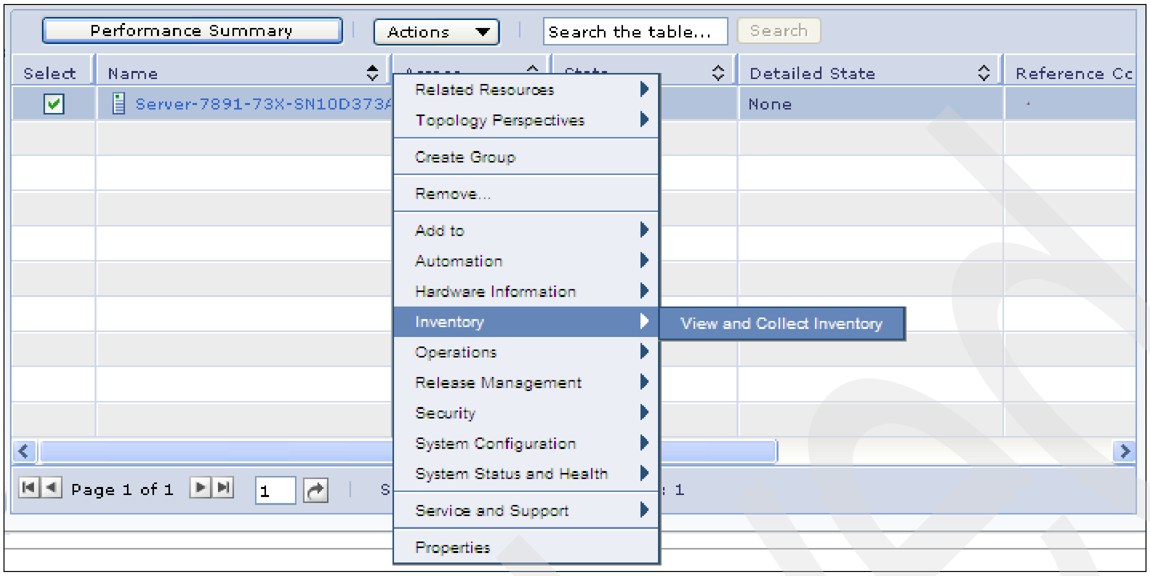


Figure 5-12 Starting inventory collection from the SDMC

Select **View and Collect Inventory** to access an additional page that will show you existing inventory and a option button to collect inventory.

Essentially all core Systems Director functions of SDMC are executed as jobs and can be run immediately or scheduled. Inventory collection for a Server object would typically be run immediately.

Inventory collection completes the preliminary operations of bringing a POWER- based blade into the SDMC environment. Configuration of virtual servers can now be started.

5.5 SDMC basic management of POWER-based blades

This section is only a brief introduction to the functionality of the SDMC in a POWER-based blade environment. No attempt has been made to be all-inclusive, but only to show the basics such as the virtual server creation and new functions now available to POWER-based blades when managed from an SDMC. These topics will likely be of interest primarily to users of the SDMC who have only used IVM in the past and have no HMC experience.

Topics covered in this section are:

- 5.5.1, “Virtual server creation”
- 5.5.2, “Multiple profiles for POWER-based blades” on page 171
- 5.5.3, “Dual VIOS on POWER-based blades” on page 173
- 5.5.4, “Virtual server Suspend and Resume” on page 175
- 5.5.5, “Active Memory Expansion (AME)” on page 176
- 5.5.6, “Virtual consoles from the SDMC” on page 177

The following sections continue to use a PS703 to illustrate the examples.

5.5.1 Virtual server creation

Previous to SDMC, Integrated Virtualization Manager was required as the management device for POWER-based blades. IVM would install as part of the Virtual I/O Server LPAR in position 1, then additional LPARs could be created and managed either from the CLI or the user interface.

The SDMC interfaces directly with the FSP on the POWER-based blades and allows virtual server creation even with the blade in a standby power state. The creation of virtual servers prior to the installation of the VIOS allows more flexibility and capability for this POWER platform.

Starting from the SDMC Resources view you can select a server to create a virtual server on. Right-click the server object, then select **System Configuration** → **Create Virtual Server** as shown in Figure 5-13.

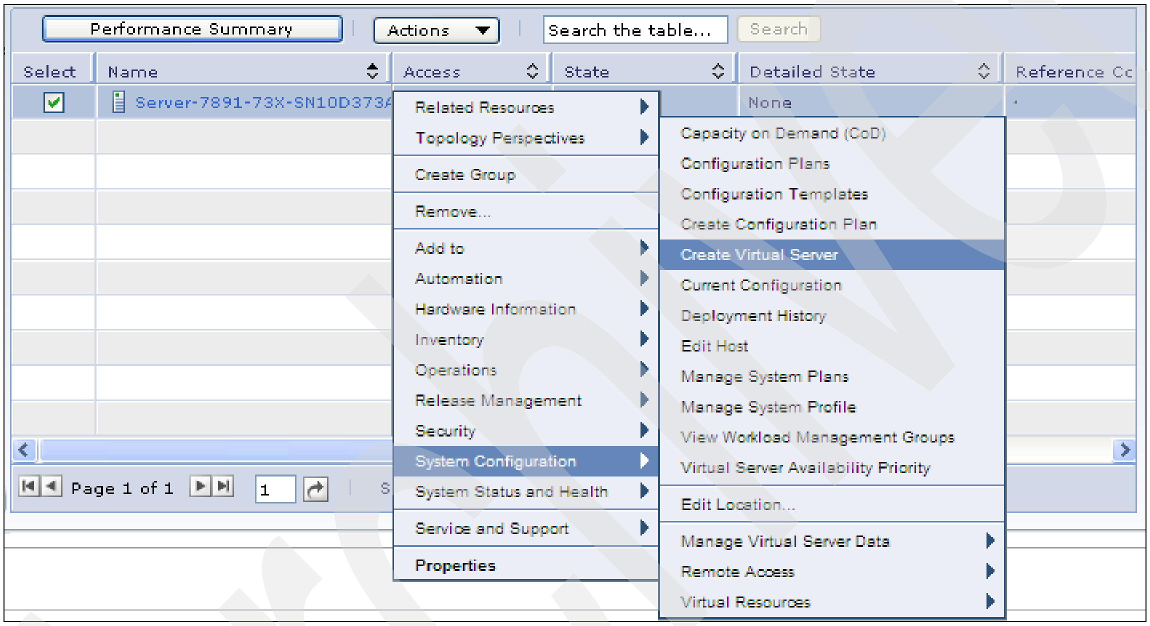


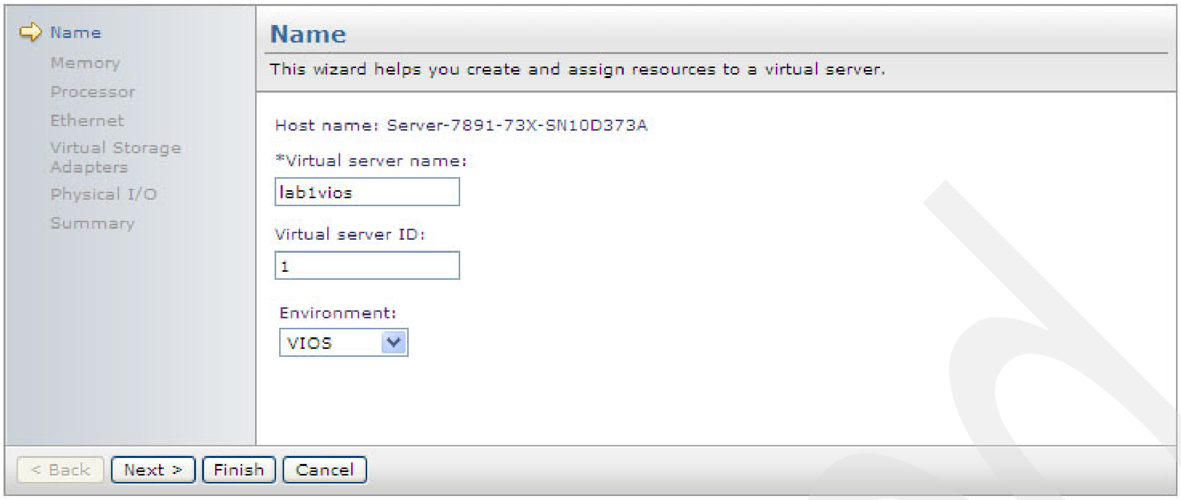
Figure 5-13 Starting the creation of a virtual server on a PS703 blade

With this option selected, a virtual server creation wizard is started to lead you through the remaining steps. These steps are very similar to the ones presented by the HMC or IVM user interfaces.

The wizard prompts you to specify a Virtual server name, Virtual server ID, and the type of Environment (Figure 5-14 on page 167). The environment choices are:

- AIX/Linux
- IBM i
- VIOS

In this example a VIOS environment was chosen.



Name

This wizard helps you create and assign resources to a virtual server.

Host name: Server-7891-73X-SN10D373A

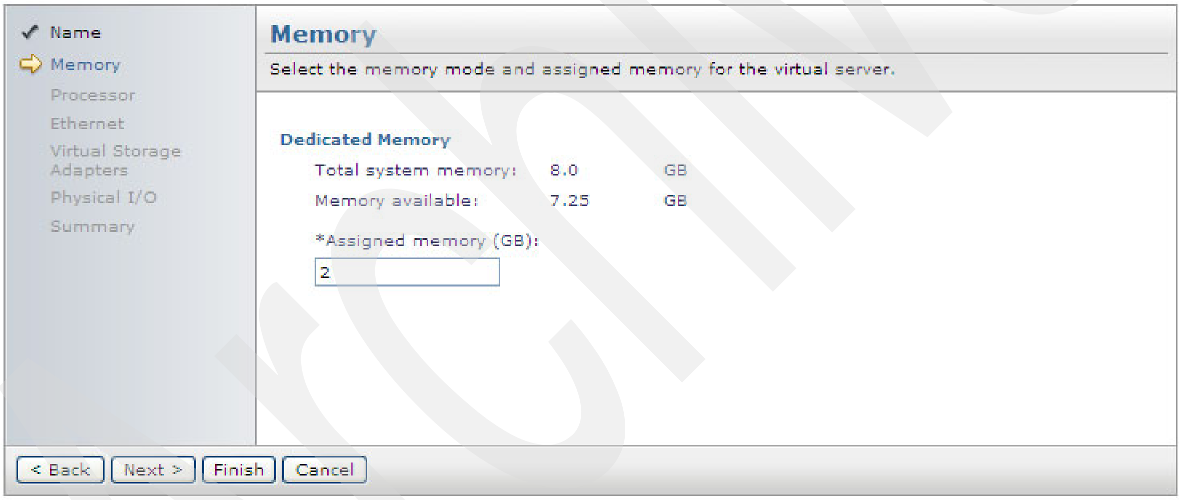
*Virtual server name: lab1vios

Virtual server ID: 1

Environment: VIOS

Figure 5-14 Virtual server creation wizard Step 1

Next you assign memory to the virtual server (Figure 5-15). Virtual servers for VIOS only have the option for dedicated memory. Virtual servers for IBM i or AIX/Linux can be assigned dedicated or shared memory if an Active Memory Sharing (AMS) pool has been previously created.



Memory

Select the memory mode and assigned memory for the virtual server.

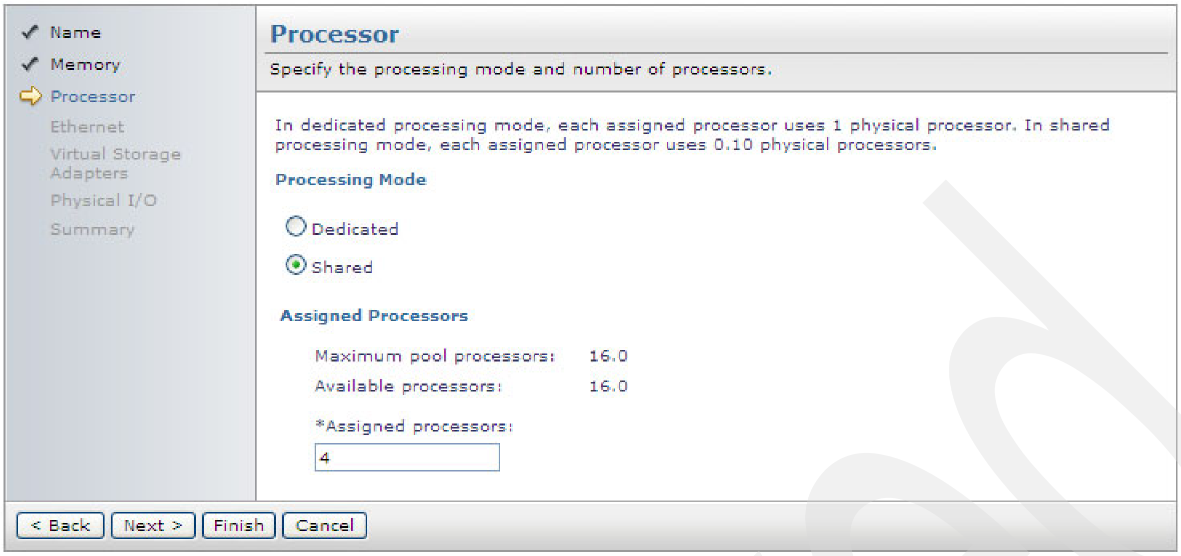
Dedicated Memory

Total system memory:	8.0	GB
Memory available:	7.25	GB

*Assigned memory (GB): 2

Figure 5-15 Virtual server creation wizard Step 2: Memory assignment

Processor allocation to a virtual server can either be dedicated or shared. SDMC can created multiple shared processor pools. Pool assignments can be made after a virtual server is created. Figure 5-16 on page 168 shows the assignment of 4 virtual processor.



Processor

Specify the processing mode and number of processors.

In dedicated processing mode, each assigned processor uses 1 physical processor. In shared processing mode, each assigned processor uses 0.10 physical processors.

Processing Mode

○ Dedicated

● Shared

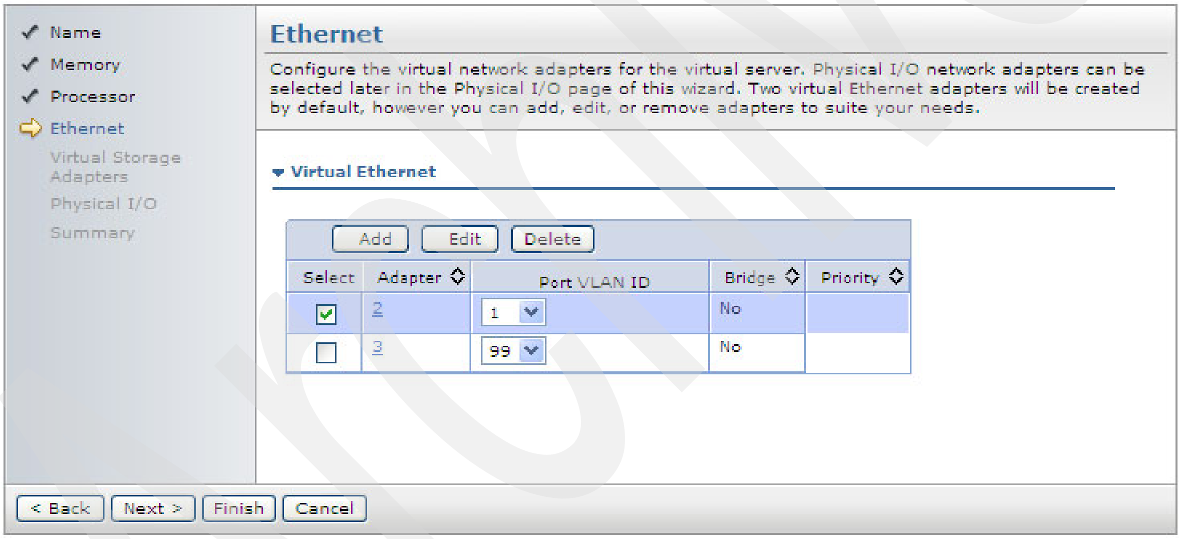
Assigned Processors

Maximum pool processors:	16.0
Available processors:	16.0

*Assigned processors: 4

Figure 5-16 Virtual server creation wizard Step 3: Processor assignment

Virtual Ethernet adapter assignment is the next step. Figure 5-17 shows the default number of virtual adapters is two.



Ethernet

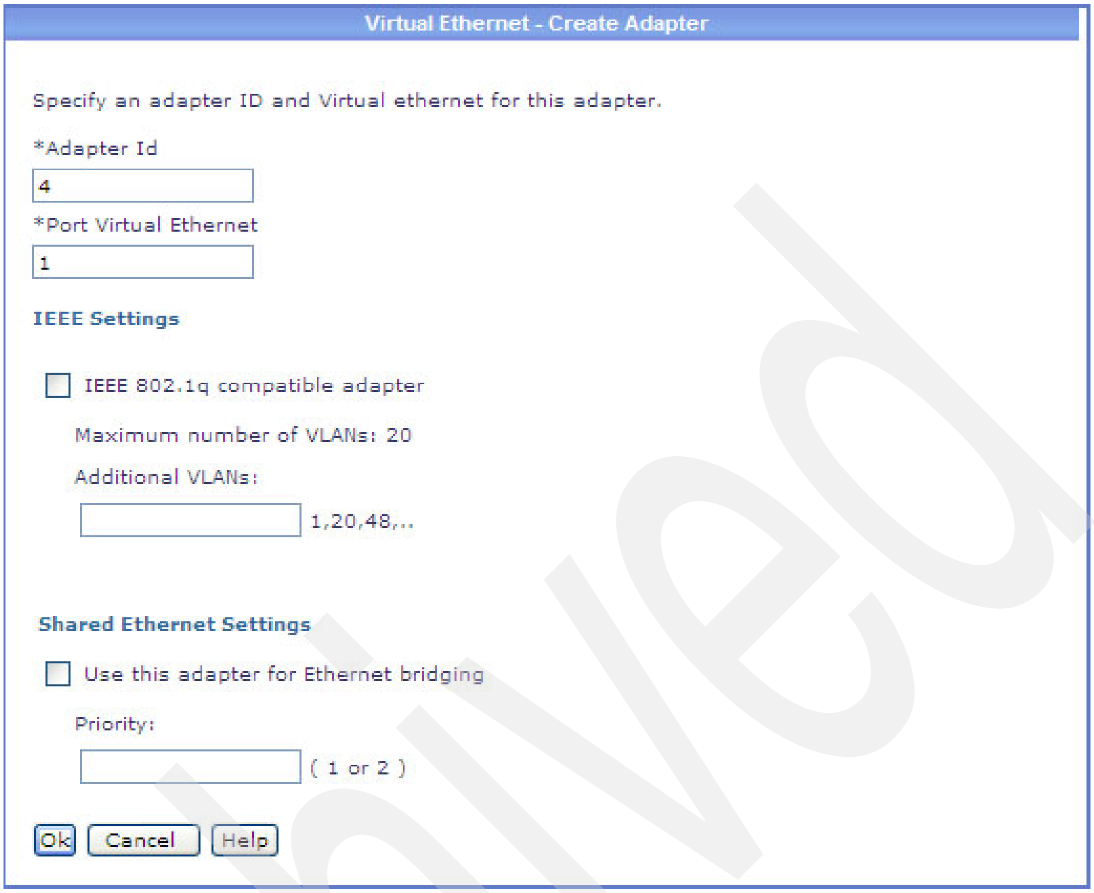
Configure the virtual network adapters for the virtual server. Physical I/O network adapters can be selected later in the Physical I/O page of this wizard. Two virtual Ethernet adapters will be created by default, however you can add, edit, or remove adapters to suite your needs.

Virtual Ethernet

Select	Adapter	Port VLAN ID	Bridge	Priority
☑	2	1	No	
□	3	99	No	

Figure 5-17 Virtual server creation wizard Step 4: Virtual Ethernet assignment

The **Add** or **Edit** button on the virtual Ethernet adapter page can be used to add additional adapters or modify existing adapters. The possible changes are shown in Figure 5-18 on page 169. Unused adapters can be removed with the **Delete** button.



Virtual Ethernet - Create Adapter

Specify an adapter ID and Virtual ethernet for this adapter.

*Adapter Id: 4

*Port Virtual Ethernet: 1

IEEE Settings

□ IEEE 802.1q compatible adapter

Maximum number of VLANs: 20

Additional VLANs: 1,20,48,..

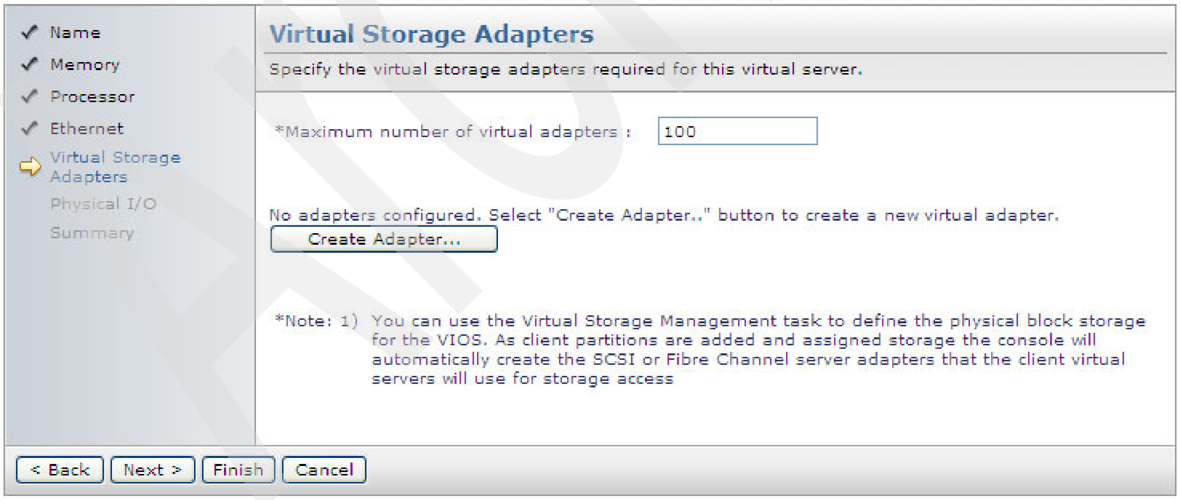
Shared Ethernet Settings

□ Use this adapter for Ethernet bridging

Priority: (1 or 2)

Figure 5-18 Editing a virtual Ethernet adapter

Virtual storage adapters can be added next, as shown in Figure 5-19.



Virtual Storage Adapters

Specify the virtual storage adapters required for this virtual server.

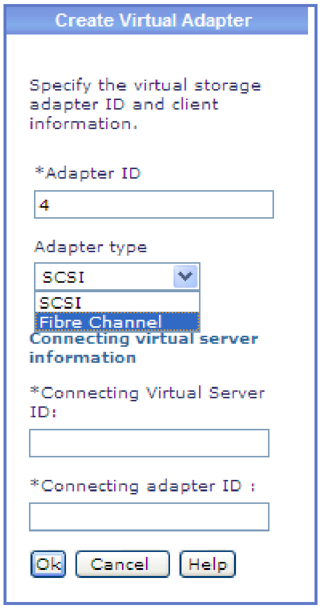
*Maximum number of virtual adapters : 100

No adapters configured. Select "Create Adapter.." button to create a new virtual adapter.

*Note: 1) You can use the Virtual Storage Management task to define the physical block storage for the VIOS. As client partitions are added and assigned storage the console will automatically create the SCSI or Fibre Channel server adapters that the client virtual servers will use for storage access

Figure 5-19 Virtual server creation wizard Step 5: Virtual storage adapter creation

The types of virtual storage adapters available are SCSI and Fibre Channel, as shown in Figure 5-20 on page 170. Connecting or partner adapter information is also specified on this panel.



Create Virtual Adapter

Specify the virtual storage adapter ID and client information.

*Adapter ID: 4

Adapter type: SCSI

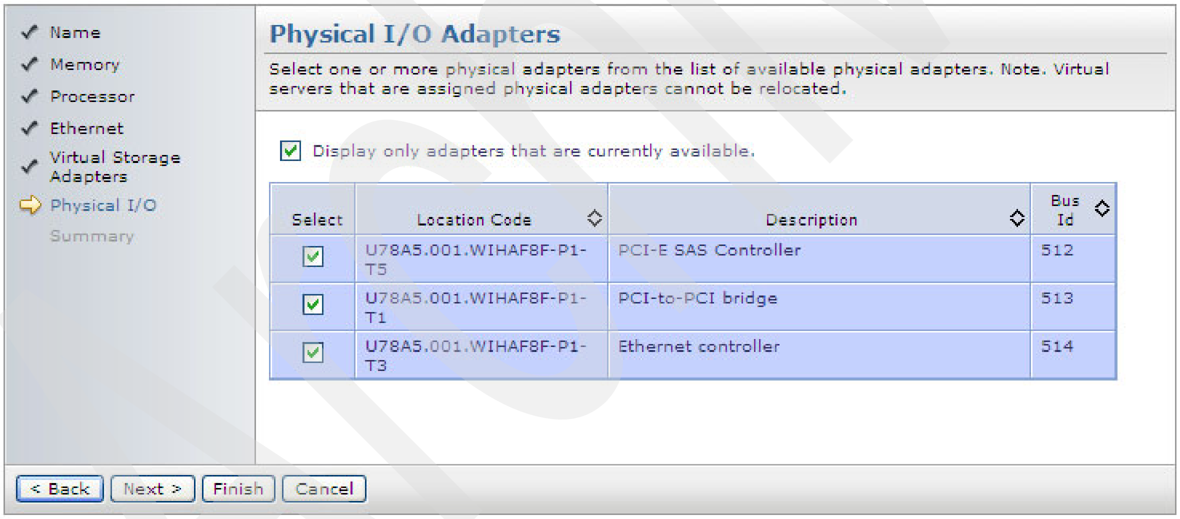
Connecting virtual server information

*Connecting Virtual Server ID:

*Connecting adapter ID :

Figure 5-20 Virtual server creation wizard Step 5: Virtual storage adapter type selection

Physical adapter selection is the next step in the wizard. All the physical adapters in the system can be displayed, or toggled to only show the available hardware. Figure 5-21 shows the check box marked to display only adapters that are currently available.



Physical I/O Adapters

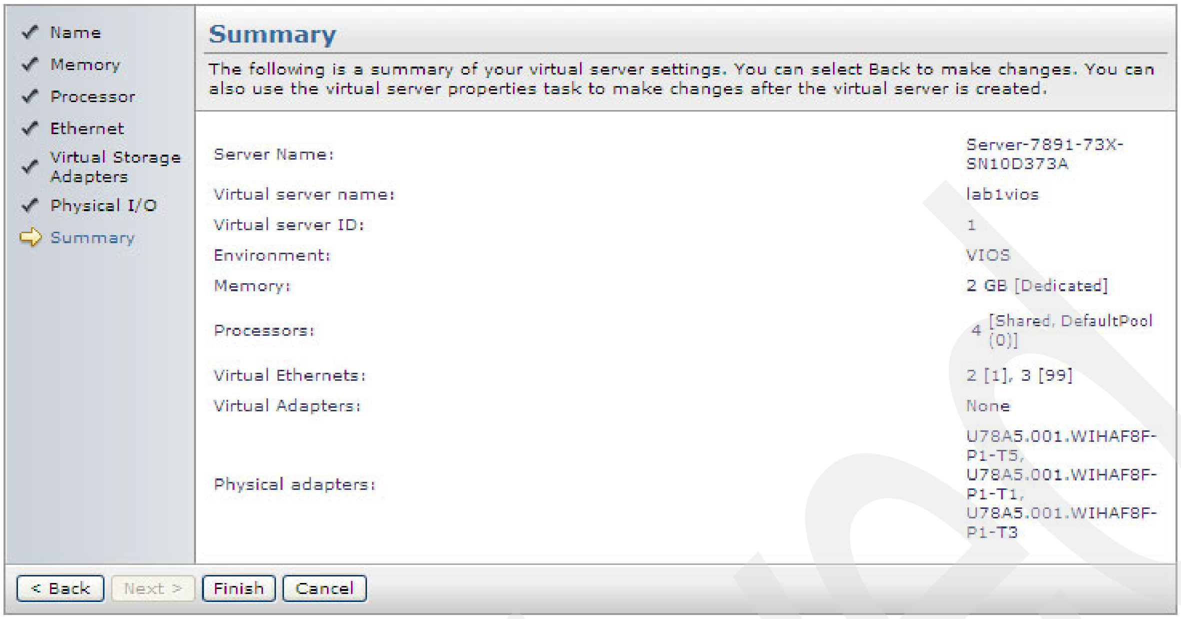
Select one or more physical adapters from the list of available physical adapters. Note. Virtual servers that are assigned physical adapters cannot be relocated.

☑ Display only adapters that are currently available.

Select	Location Code	Description	Bus Id
☑	U78A5.001.WIHAF8F-P1-T5	PCI-E SAS Controller	512
☑	U78A5.001.WIHAF8F-P1-T1	PCI-to-PCI bridge	513
☑	U78A5.001.WIHAF8F-P1-T3	Ethernet controller	514

Figure 5-21 Virtual server creation wizard Step 6: Physical I/O adapter selection

The last step is a summary page that lists all the selections that have been made in the previous steps, as shown in Figure 5-22 on page 171.



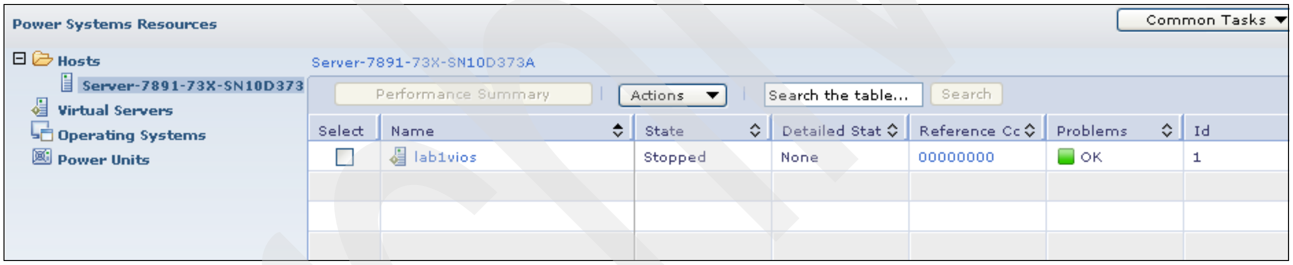
Summary

The following is a summary of your virtual server settings. You can select Back to make changes. You can also use the virtual server properties task to make changes after the virtual server is created.

Server Name:	Server-7891-73X-SN10D373A
Virtual server name:	lab1vios
Virtual server ID:	1
Environment:	VIOS
Memory:	2 GB [Dedicated]
Processors:	4 [Shared, DefaultPool (0)]
Virtual Ethernets:	2 [1], 3 [99]
Virtual Adapters:	None
Physical adapters:	U78A5.001.WIHAF8F-P1-T5, U78A5.001.WIHAF8F-P1-T1, U78A5.001.WIHAF8F-P1-T3

Figure 5-22 Virtual server creation wizard Step 7: Summary page

Click **Finish** on the summary page to view a panel displaying the new virtual server (Figure 5-23).



Select	Name	State	Detailed Stat	Reference Cc	Problems	Id
□	lab1vios	Stopped	None	00000000	🟩 OK	1

Figure 5-23 SDMC showing new virtual server and status information

A profile is created for the virtual server. This profile can be edited and additional profiles can be generated for the same virtual server.

5.5.2 Multiple profiles for POWER-based blades

IVM-managed systems had the limitation of a single configuration for each LPAR and essentially did not use the concept of profiles. POWER-based blades managed by SDMC can have multiple profiles for each virtual server.

Accessing the profiles for a virtual server can be done from the right-click context menus as shown Figure 5-24 on page 172.

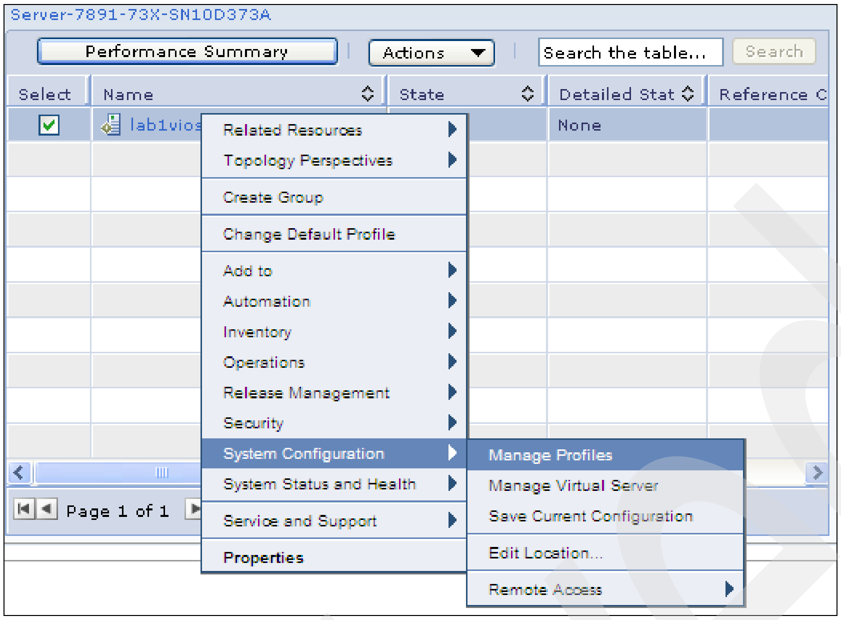
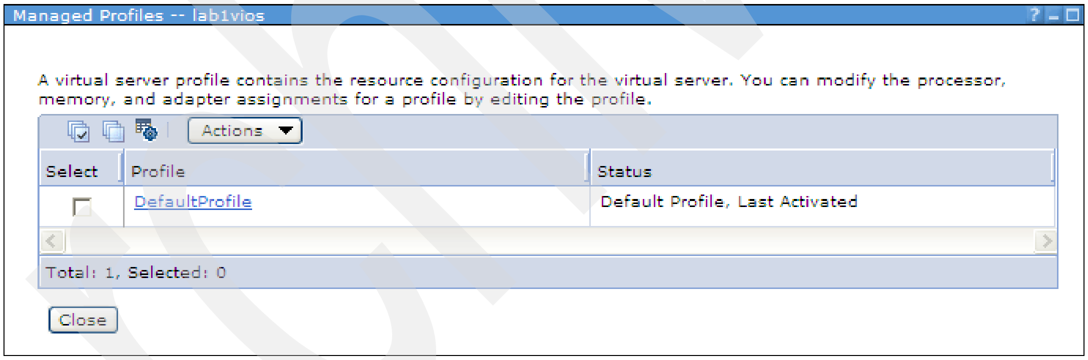


Figure 5-24 Accessing a virtual server profile

When the Managed Profiles page opens a list of profiles associated with the virtual server is shown. Figure 5-25 shows a single profile that has a status of being the default profile and the last activated.



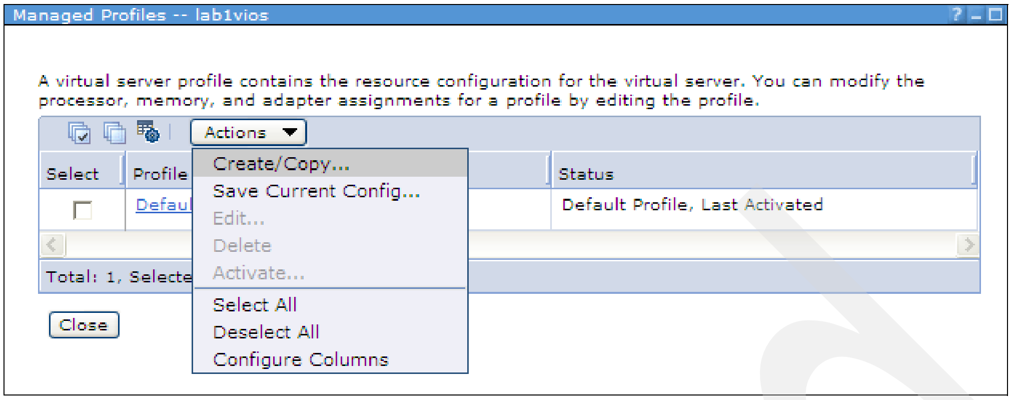
A virtual server profile contains the resource configuration for the virtual server. You can modify the processor, memory, and adapter assignments for a profile by editing the profile.

Select	Profile	Status
□	DefaultProfile	Default Profile, Last Activated

Total: 1, Selected: 0

Figure 5-25 Managed profiles page

The existing profile can be edited, used as a starting point for a new copy, or the current running configuration (which might include DLPAR or Managed Virtual Server operations) can be saved into a new profile. These options are shown in Figure 5-26 on page 173.



Managed Profiles -- lab1vios

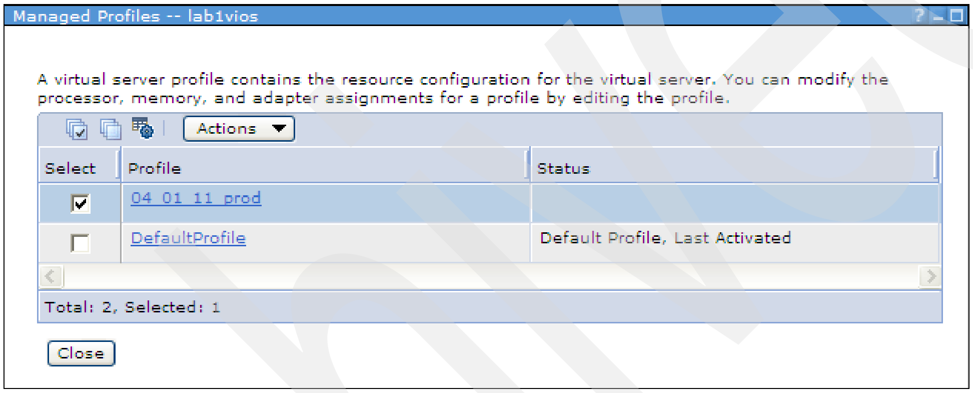
A virtual server profile contains the resource configuration for the virtual server. You can modify the processor, memory, and adapter assignments for a profile by editing the profile.

Select	Profile	Status
□	Defaul	Default Profile, Last Activated

Total: 1, Selecte

Figure 5-26 Editing and creating new virtual server profiles

After a new profile is created it will be included in the profile list as shown in Figure 5-27.



Managed Profiles -- lab1vios

A virtual server profile contains the resource configuration for the virtual server. You can modify the processor, memory, and adapter assignments for a profile by editing the profile.

Select	Profile	Status
☑	04_01_11_prod	
□	DefaultProfile	Default Profile, Last Activated

Total: 2, Selected: 1

Figure 5-27 List of available profiles for a virtual server

5.5.3 Dual VIOS on POWER-based blades

One of the new capabilities that has been added to POWER6- and POWER7-based blades managed by an SDMC is the ability to implement dual Virtual I/O Servers in the same way HMC-managed systems are able to.

Previously, under IVM-managed systems, the VIOS/IVM would install itself in LPAR number one, then additional LPARs could be created using IVM. No additional VIOS LPARs are allowed to be created under this management method.

With SDMC the creation of virtual servers and the type of operating system environment that they support can occur prior to any operating system installation. The only limitation from a dual VIOS perspective is the availability of disk and network physical resources. Physical resource assignment to an virtual server is made at the expansion card slot or controller slot (physical location code) level. Individual ports and internal disks cannot be individually assigned. This type of assignment is not unique to POWER-based blades and is common practice to all POWER platforms.

A minor exception is the 4Gb or 8Gb CFFh combo cards that have both two Ethernet and two Fibre Channel ports. The Ethernet section of the expansion card can be assigned to a different virtual server from the virtual server that is using the Fibre Channel section. However, a split assignment of the same type of port is still not allowed.

A dual VIOS environment setup requires the creation of the two virtual servers, both of which are set for a VIOS environment. After the virtual servers are created with the appropriate environment setting and physical resources assigned to support independent disk and network I/O, then the VIOS operating systems can be installed.

When planning for a dual VIOS environment on a blade, your hardware configuration will have to have take into account the need for two virtual servers that both must have physical Ethernet and disk resources available to each VIOS. The following examples describe some possible hardware configurations to support a dual VIOS environment, but are not intended to be all-inclusive.

PS703

A typical basic configuration would be 8 GB of memory, a single internal disk, and the two integrated Ethernet adapters. To support a dual VIOS environment the following additional options would be required:

- One CIOv Fibre Channel adapter (feature codes 8240, 8241, or 8242)
- One CFFh Fibre Channel and Ethernet combo card (feature codes 8252 or 8271)
- BladeCenter MSIM adapters to accommodate additional BladeCenter switch modules

The pair of internal Ethernet ports on the blade and Fibre Channel ports on the CFFh expansion would be assigned to the first VIOS. The Fibre Channel ports on the CIOv expansion card and the Ethernet ports on the CFFh combo card would be assigned to the second VIOS. Both VIO Servers in this example would boot from SAN. The SAS controller and internal drive can only be owned by one VIOS and in this example would not be used.

PS704

A typical basic configuration would be 16 GB of memory, two internal drives, and four integrated Ethernet adapters. To support a dual VIOS environment the following additional option would be required:

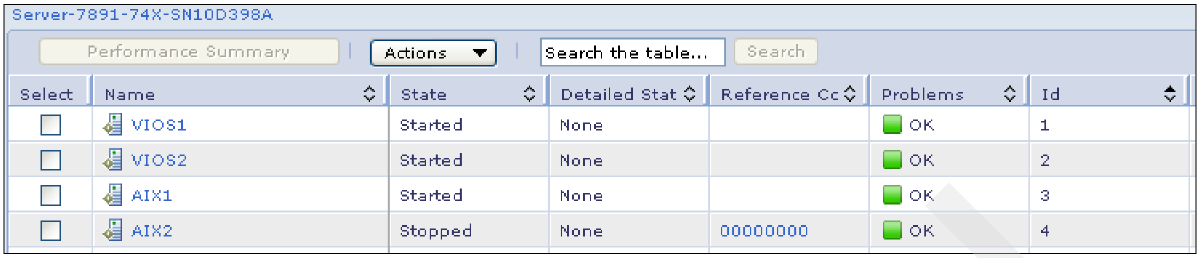
- Two CIOv Fibre Channel adapters (feature codes 8240, 8241, or 8242)

One of the two pairs of internal Ethernet ports in each half of the PS704 would be assigned to a VIOS. One CIOv expansion card in each half of the PS704 would be assigned to a VIOS. Again in this scenario both VIO Servers would be booting from SAN and the SAS controller and two internal disks would be unused.

Other variations of two CFFh Converged Network Adapters (feature code 8275) and FCoE capable switches or other CFFh combo expansion cards with BladeCenter MSIMs and switches could be used.

These examples for the PS703 and PS704, while not inclusive, provide the basics for a dual VIOS environment. Memory requirements for additional virtual servers beyond the base order amounts were not considered and should be evaluated prior to ordering either model.

The actual steps of creating a dual VIOS will not be covered here, however the end results of this type of configuration performed on a PS704 are show in Figure 5-28 on page 175.



Server-7891-74X-SN10D398A

Select	Name	State	Detailed Stat	Reference Cc	Problems	Id
□	VIOS1	Started	None		🟩 OK	1
□	VIOS2	Started	None		🟩 OK	2
□	AIX1	Started	None		🟩 OK	3
□	AIX2	Stopped	None	00000000	🟩 OK	4

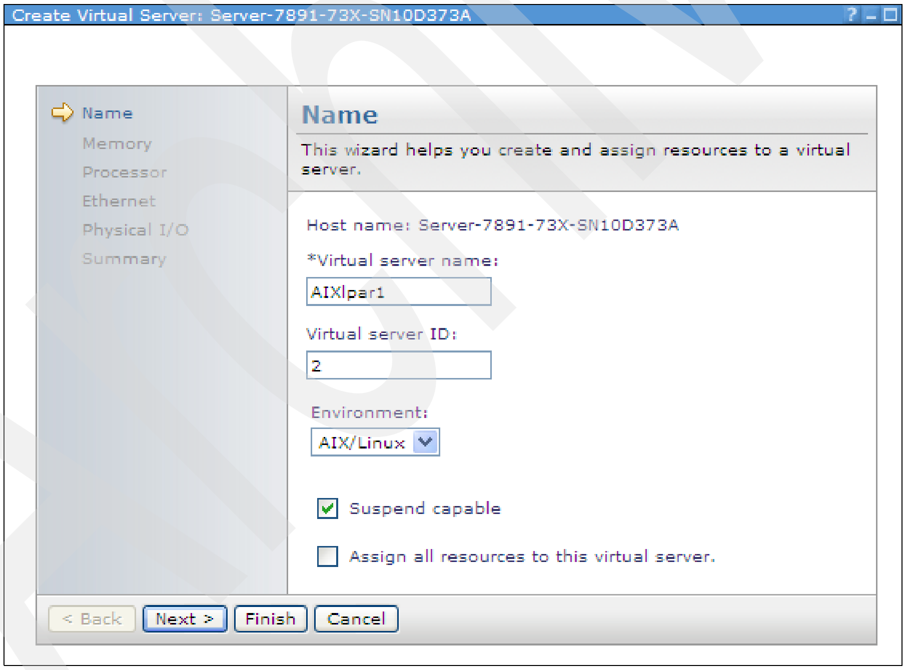
Figure 5-28   Dual VIOS configuration on a PS704

Once the two Virtual I/O Servers are installed, the normal methods of creating a Share Ethernet Adapter (SEA) failover for virtual networking, and redundant paths for the client virtual server disks (NPIV and vSCSI), can be configured.

5.5.4 Virtual server Suspend and Resume

A virtual server working in conjunction with a VIOS reserved storage device pool can suspend the operating system and applications. When the virtual server is resumed the prior running processes are also resumed. The processor and memory resources of a suspended virtual server can be re-assigned to other virtual servers if needed.

When using the SDMC to create a new AIX/Linux or IBM i virtual server a check box option lets you make the virtual server suspend capable (Figure 5-29).



Create Virtual Server: Server-7891-73X-SN10D373A

Name

This wizard helps you create and assign resources to a virtual server.

Host name: Server-7891-73X-SN10D373A

*Virtual server name: AIXlpar1

Virtual server ID: 2

Environment: AIX/Linux

☑ Suspend capable

□ Assign all resources to this virtual server.

Figure 5-29   Setting suspend/resume option

Initiating a suspend or resume operation is done by right-clicking the server of interest and making selections from the context menu. Figure 5-30 on page 176 shows how to start the suspend or resume operation.

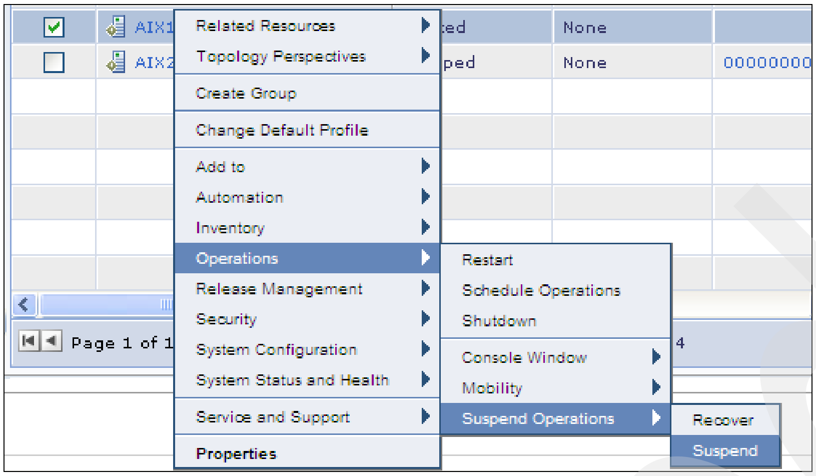
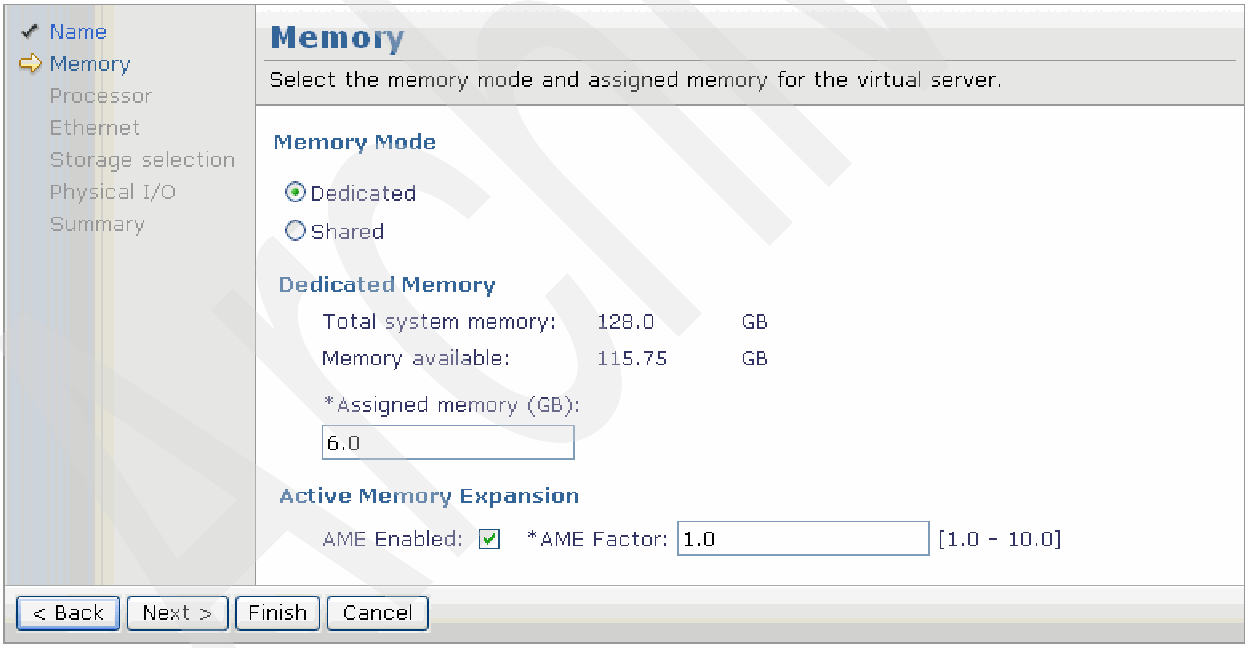


Figure 5-30 Starting the suspend operation on a virtual server

5.5.5 Active Memory Expansion (AME)

Active Memory Expansion is the ability to expand the memory available to an AIX virtual server beyond the amount of assigned physical memory. AME compresses memory pages using processor capacity to provide additional memory capacity for a virtual server. Figure 5-31 show how to select this option during the creation of a virtual server.



Memory

Select the memory mode and assigned memory for the virtual server.

Memory Mode

● Dedicated

○ Shared

Dedicated Memory

Total system memory: 128.0 GB

Memory available: 115.75 GB

*Assigned memory (GB): 6.0

Active Memory Expansion

AME Enabled: ☑ *AME Factor: 1.0 [1.0 - 10.0]

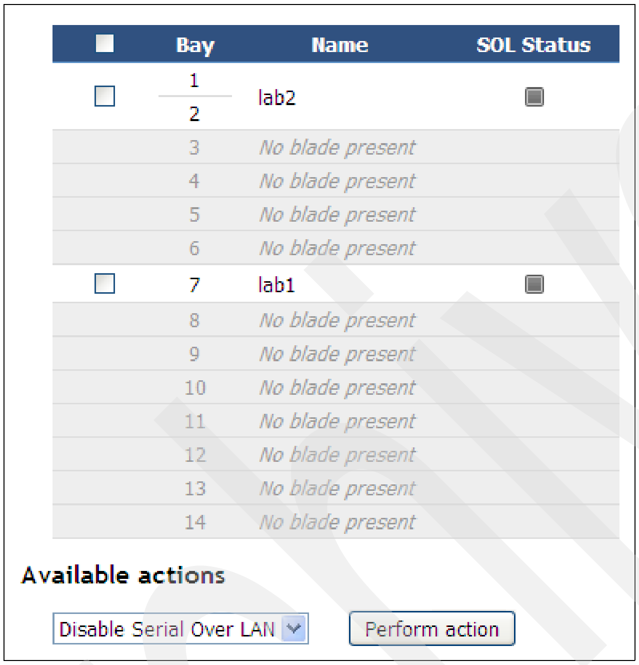
Figure 5-31 Selecting the AME option during virtual server creation

> **Note:** AME is not a PowerVM capability; it must be ordered as a separate feature code #4796. AIX 6.1 Technology Level 6 or later, or AIX 7.1 are the supported operating systems for enabled virtual servers.

5.5.6 Virtual consoles from the SDMC

IVM-managed POWER-based blades can have console sessions to the VIOS from either the VIOS virtual console function or Serial Over LAN (SOL) through the BladeCenter AMM. Client LPARs in a VIOS/IVM environment can have a single virtual console through the IVM function.

SDMC provides for a virtual console to any virtual server operating system. In order for the SDMC console to function to the same virtual server that owns the first Ethernet port, SOL must be disabled for that blade from the AMM by selecting **Blade Tasks** → **Configuration** → **Serial Over LAN**, as shown in Figure 5-32.



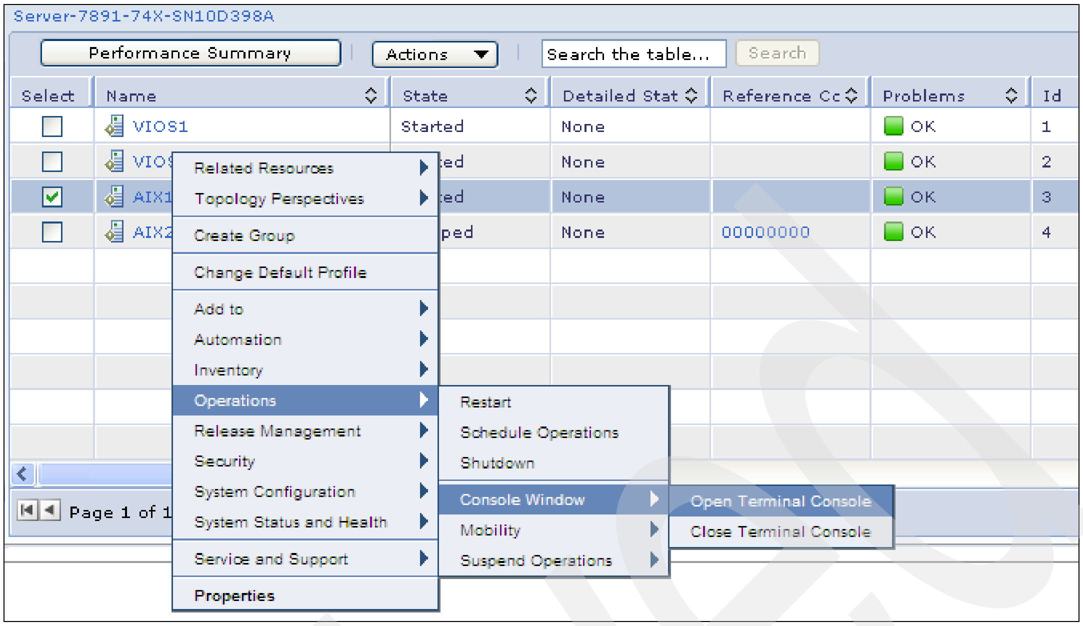
□	Bay	Name	SOL Status
□	1 2	lab2	■
	3	*No blade present*	
	4	*No blade present*	
	5	*No blade present*	
	6	*No blade present*	
□	7	lab1	■
	8	*No blade present*	
	9	*No blade present*	
	10	*No blade present*	
	11	*No blade present*	
	12	*No blade present*	
	13	*No blade present*	
	14	*No blade present*	

Available actions

Disable Serial Over LAN

Figure 5-32 SOL disabled from AMM

The console window to a virtual server can be opened by right-clicking the virtual server to access the context menu as shown in Figure 5-33 on page 178.



Select	Name	State	Detailed Stat	Reference Cc	Problems	Id
☐	VIOS1	Started	None		🟩 OK	1
☐	VIOS	ted	None		🟩 OK	2
☑	AIX1	ted	None		🟩 OK	3
☐	AIX2	ped	None	00000000	🟩 OK	4

Figure 5-33 Opening a console to a virtual server

When the operation completes a separate window will be opened for the console that will require the same password as the current ID login to the SDMC before the connection completes.

Abbreviations and acronyms

AC	alternating current
AMD	Advanced Micro Devices
AMM	Advanced Management Module
ARP	Address Resolution Protocol
ASIC	application-specific integrated circuit
ASMI	Advanced System Management Interface
BIOS	basic input output system
BIST	Built-in self-test
BOOTP	boot protocol
CD-ROM	compact disc read only memory
CEC	central electronics complex
CEE	Converged Enhanced Ethernet
CIF	Customer Installed Features
CLI	command-line interface
CMOS	complementary metal oxide semiconductor
CNA	Converged Network Adapter
CPU	central processing unit
CRC	cyclic redundancy check
CRU	customer replaceable units
CSU	Customer Set Up
DASD	direct access storage device
DC	domain controller
DDR	Double Data Rate
DHCP	Dynamic Host Configuration Protocol
DIMM	dual inline memory module
DLPAR	Dynamic Logical Partition
DPM	Distributed Power Management
DRAM	dynamic random access memory
DSM	disk storage module
ECC	error checking and correcting
EED	extended error data
EEH	extended error handling
ELA	error log analysis
ESA	Electronic Service Agent
ETSI	European Telecommunications Standard Industry
FC	Fibre Channel
FC-AL	Fibre Channel-arbitrated loop
FC-IP	Fibre Channel Internet Protocol
FDMI	Fibre Device Management Interface
FFDC	first-failure data capture
FIR	fault isolation registers
FRU	field replaceable unit
FSP	flexible service processor
GB	gigabyte
HBA	host bus adapter
HDD	hard disk drive
HEA	Host Ethernet Adapter
HMC	Hardware Management Console
HPC	high performance computing
HSSM	high speed switch module
HT	Hyper-Threading
HTTP	Hypertext Transfer Protocol
I/O	input/output
IBA	InfiniBand Architecture
IBM	International Business Machines
IBTA	InfiniBand Trade Association
ID	identifier
IEEE	Institute of Electrical and Electronics Engineers
IOC	IO controllers
IP	Internet Protocol
IPL	initial program load
IPTV	Internet Protocol Television
ISA	industry standard architecture
IT	information technology
ITSO	International Technical Support Organization
IVE	Integrated Virtual Ethernet
IVM	Integrated Virtualization Manager
KVM	keyboard video mouse
LAN	local area network
LDAP	Lightweight Directory Access Protocol
LED	light emitting diode
LMB	Logical Memory Block
LPAR	logical partitions
LPM	Live Partition Mobility

LUN	logical unit number
MAC	media access control
MES	Miscellaneous Equipment Specification
MPI	Message Passing Interface
MSIM	Multi-Switch Interconnect Module
MSP	mover service partition
MTU	maximum transmission unit
NASA	National Aeronautics and Space Administration
NDP	Neighbor Discovery Protocol
NEBS	Network Equipment Building System
NGN	next-generation network
NIM	Network Installation Management
NL	near line
NPIV	N_Port ID Virtualization
NVRAM	non-volatile random access memory
OS	operating system
PCI	Peripheral Component Interconnect
PDU	power distribution unit
PHB	processor host bridge
POST	power-on self test
PS	Personal System
PVID	Port VLAN Identifier
PXE	Preboot eXecution Environment
QDR	quad data rate
RAID	redundant array of independent disks
RAS	remote access services; row address strobe
RDIMM	registered DIMM
RHEL	Red Hat Enterprise Linux
RSA	Remote Supervisor Adapter
SAN	storage area network
SAS	Serial Attached SCSI
SATA	Serial ATA
SCM	Supply Chain Management
SCSI	Small Computer System Interface
SEA	Shared Ethernet Adapter
SER	soft error
SFP	small form-factor pluggable
SLB	Segment Lookaside Buffer
SLES	SUSE Linux Enterprise Server
SMP	symmetric multiprocessing
SMS	System Management Services
SMT	Simultaneous Multithread
SNMP	Simple Network Management Protocol
SOI	silicon-on-insulator
SOL	Serial over LAN
SPCN	System Power Control Network
SRAM	static RAM
SSA	serial storage architecture
SSD	solid state drive
SSH	Secure Shell
SSL	Secure Sockets Layer
SSP	Serial SCSI Protocol
SUE	Special Uncorrectable Error
SWMA	Software Maintenance Agreement
TB	terabyte
TCO	total cost of ownership
TL	technology level
TPMD	thermal and power management device
TTY	teletypewriter
USB	universal serial bus
VASI	Virtual Asynchronous Services Interface
VESA	Video Electronics Standards Association
VID	VLAN Identifier
VIOS	Virtual I/O Server
VLAN	virtual LAN
VLP	very low profile
VOIP	Voice over Internet Protocol
VPD	vital product data
WOL	Wake on LAN
WWPN	World Wide Port Name

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 documents

For information about ordering these publications, see “How to get Redbooks” on page 182. Note that the documents referenced here might be available in softcopy only.

- *An Introduction to Fibre Channel over Ethernet, and Fibre Channel over Convergence Enhanced Ethernet*, REDP-4493
- *Hardware Management Console V7 Handbook*, SG24-7491
- *HPC Clusters Using InfiniBand on IBM Power Systems Servers*, SG24-7767
- *IBM BladeCenter JS12 and JS22 Implementation Guide*, SG24-7655
- *IBM BladeCenter JS23 and JS43 Implementation Guide*, SG24-7740
- *IBM BladeCenter Products and Technology*, SG24-7523
- *IBM PowerVM Live Partition Mobility*, SG24-7460
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *Integrated Virtual Ethernet Adapter Technical Overview and Introduction*, REDP-4340
- *Integrated Virtualization Manager on IBM System p5*, REDP-4061
- *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- *PowerVM Virtualization Active Memory Sharing*, REDP-4470
- *PowerVM Virtualization on IBM System p: Introduction and Configuration Fourth Edition*, SG24-7940

Other publications

These publications are also relevant as further information sources, available from `http://http://ibm.com/systems/support` (click BladeCenter):

- *IBM BladeCenter PS703 and PS704 Installation and User's Guide*
- *IBM BladeCenter PS703 and PS704 Problem Determination and Service Guide*

Online resources

This Web site is also relevant as further information sources:

- IBM BladeCenter PS700, PS701, and PS702 Express home page

  http://ibm.com/systems/bladecenter/hardware/servers/ps700series

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®

IBM BladeCenter PS703 and PS704 Technical Overview and Introduction

Redpaper™

Features the POWER7 processor providing advanced multi-core technology

Details the follow-on to the BladeCenter PS700, PS701 and PS702

Describes management using the new Systems Director Management Console

The IBM BladeCenter PS703 and PS704 are premier blades for 64-bit applications. They are designed to minimize complexity, improve efficiency, automate processes, reduce energy consumption, and scale easily. These blade servers are based on the IBM POWER7 processor and support AIX, IBM i, and Linux operating systems. Their ability to coexist in the same chassis with other IBM BladeCenter blade servers enhances the ability to deliver the rapid return on investment demanded by clients and businesses.

This IBM Redpaper doocument is a comprehensive guide covering the IBM BladeCenter PS703 and PS704 servers. The goal of this paper is to introduce the offerings and their prominent features and functions.

REDP-4744-00