IBM Power Systems E870 and E880 Technical Overview and Introduction

New modular architecture for increased reliability

Enterprise POWER8 processor-based servers

Exceptional memory and I/O bandwidth

Alexandre Bicas Caldeira
YoungHoon Cho
James Cruickshank
Bartłomiej Grabowski
Volker Haug
Andrew Laidlaw
Seulgi Yoppy Sung

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM Power Systems E870 and E880 Technical Overview and Introduction

December 2014

REDP-5137-00

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

First Edition (December 2014)

This edition applies to the IBM Power E870 (9119-MME) and Power E880 (9119-MHE) Power Systems servers.

This document was created or updated on July 25, 2017.

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 grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

Active Memory™
AIX®
DB2®
DS8000®
Easy Tier®
Electronic Service Agent™
EnergyScale™
eServer™
FlashSystem™
Focal Point™
IBM®
IBM FlashSystem™
Micro-Partitioning®
POWER®
Power Architecture®
POWER Hypervisor™
Power Systems™
Power Systems Software™
POWER6®
POWER6+™
POWER7®
POWER7+™
POWER8™
PowerHA®
PowerPC®
PowerVM®
PowerVP™
Rational®
Real-time Compression™
Redbooks®
Redpaper™
Redbooks (logo) ®
RS/6000®
Storwize®
System p®
System Storage®
System z®
SystemMirror®
Tivoli®
XIV®

The following terms are trademarks of other companies:

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.

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

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

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

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

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

Preface

This IBM® Redpaper™ publication is a comprehensive guide covering the IBM Power System E870 (9119-MME) and IBM Power System E880 (9119-MHE) servers that support IBM AIX®, IBM i, and Linux operating systems. The objective of this paper is to introduce the major innovative Power E870 and Power E880 offerings and their relevant functions:

- The new IBM POWER8™ processor, available at frequencies of 4.024 GHz, 4.190 GHz, and 4.350 GHz
- Up to 16 TB of memory in the E870 and 32 TB in the E880
- Significantly strengthened cores and larger caches
- Two integrated memory controllers with improved latency and bandwidth
- Integrated I/O subsystem and hot-pluggable PCIe Gen3 I/O slots
- I/O drawer expansion options offers greater flexibility
- Improved reliability, serviceability, and availability (RAS) functions
- IBM EnergyScale™ technology that provides features such as power trending, power-saving, capping of power, and thermal measurement

This publication is for professionals who want to acquire a better understanding of IBM Power Systems™ products. The intended audience includes the following roles:

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

This paper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power E870 and Power E880 systems.

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

Authors

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

Alexandre Bicas Caldeira is a Certified IT Specialist and is a member of the Power Systems Advanced Technical Sales Support team for IBM Brazil. He holds a degree in Computer Science from the Universidade Estadual Paulista (UNESP) and an MBA in Marketing. His major areas of focus are competition, sales, and technical sales support. Alexandre has more than 14 years of experience working on IBM Systems & Technology Group Solutions and has worked also as an IBM Business Partner on Power Systems hardware, AIX, and IBM PowerVM® virtualization products.

YoungHoon Cho is a Power Systems Top Gun with the post-sales Technical Support Team for IBM in Korea. He has over 10 years of experience working on IBM RS/6000®, IBM System p®, and Power Systems products. He provides second-line technical support to Field Engineers working on Power Systems and system management.

James Cruickshank works in the Power System Client Technical Specialist team for IBM in the UK. He holds an Honors degree in Mathematics from the University of Leeds. James has over 13 years experience working with RS/6000, pSeries, System p and Power Systems products. James supports customers in the financial services sector in the UK.

Bartłomiej Grabowski is a Principal System Support Specialist in DHL IT Services in the Czech Republic. He holds a Bachelor’s degree in Computer Science from the Academy of Computer Science and Management in Bielsko-Biala. His areas of expertise include IBM i, IBM PowerHA® solutions that are based on hardware and software replication, Power Systems hardware, and PowerVM. He is an IBM Certified Systems Expert and a coauthor of several IBM Redbooks® publications.

Volker Haug is an Open Group Certified IT Specialist within IBM Systems & Technology Group in Germany supporting Power Systems clients and Business Partners. He holds a Diploma degree in Business Management from the University of Applied Studies in Stuttgart. His career includes more than 28 years of experience with Power Systems, AIX, and PowerVM virtualization. He has written several IBM Redbooks publications about Power Systems and PowerVM. Volker is an IBM POWER8 Champion and a member of the German Technical Expert Council, which is an affiliate of the IBM Academy of Technology.

Andrew Laidlaw is a Client Technical Specialist for IBM working in the UK. He supports Service Provider clients within the UK and Ireland, focusing primarily on Power Systems running AIX and Linux workloads. His expertise extends to open source software package including the KVM hypervisor and various management tools. Andrew holds an Honors degree in Mathematics from the University of Leeds, which includes credits from the University of California in Berkeley.

Seulgi Yoppy Sung is a very passionate Engineer, supporting multi-platform systems as a System Services Representative almost three year, include Power System hardware, AIX, high-end and low-end storage DS8000 and V7000. She is very positive and enthusiastic about Power Systems.

The project that produced this publication was managed by:

Scott Vetter
Executive Project Manager, PMP

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

Tamikia Barrow, Ella Buslovich, Mark Clark, Gareth Coates, Tim Damron, Dan DeLapp, David Dilling, Nigel Griffiths, Stacy L Haugen, Daniel Henderson, Eddie Hoyal, Tenley Jackson, Stephanie Jensen, Roxette Johnson, Caroyln K Jones, Bob Kovacs, Stephen Lutz, Jai Lei Ma, Cesar D Maciel, Chris Mann, Dwayne Moore, Michael J Mueller, Mark Olson, Michael Poli, Michael Quaranta, Marc Rauzier, Monica Sanchez, Bill Starke, Dawn C Stokes, Jeff Stuecheli, Doug Szerdi, Kristin Thomas, Jacobo Vargas, Diane Wallace, Steve Will, Gerlinde Wochele, Doug Yakesch
IBM

Louis Bellanger
Bull

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

General description

The IBM Power System E870 (9119-MME) and IBM Power System E880 (9119-MHE) servers use the latest POWER8 processor technology that is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads.

The Power E870 is a highly-scalable rack-mount system optimized for running AIX, IBM i, and Linux workloads. The Power E870 is a modular-built system and uses one or two system nodes together with a system control unit. Each system node is five EIA-units tall (5U) and the system control unit is two EIA-units (2U) tall. The Power E870 is housed in a 19-inch rack.

The Power E880 is a highly-scalable rack-mount system optimized for running AIX, IBM i, and Linux workloads. The Power E880 is a modular-built system. The system is built of one, two, three, or four system nodes together with a system control unit. Each system node is five EIA-units tall (5U) and the system control unit is two EIA-units (2U) tall. The Power E880 is housed in a 19-inch rack.

1.1 Systems overview

The following sections provide detailed information about the Power E870 and Power E880 systems.

1.1.1 Power E870 server

The Power E870 (9119-MME) server is a powerful POWER8-based system that scales up to eight sockets. Each socket contains a single 8-core or 10-core POWER8 processor. Thus, a fully-configured Power E870 can scale up to 64 or 80 cores.

The Power E870 is a modular system that is built from a combination of the following four building blocks:

- System control unit
- System node
- PCIe Gen3 I/O expansion drawer
- EXP24S SFF Gen2-bay drawer

The system control unit provides redundant system master clock and redundant system master Flexible Service Processor (FSP) and support for the Operator Panel, the system VPD, and the base DVD. The system control unit provides clock signals to the system nodes with semi-flexible connectors. A system control unit is required for every E870 system.

Each system node provides four processor sockets for POWER8 processors and 32 CDIMM slots supporting a maximum of 8 memory features. Using the 1024 GB memory features, the system node can support a maximum of 8 TB of RAM. Thus a fully-configured Power E870 can support up to 16 TB of RAM. Each system node provides eight PCIe Gen3 x16 slots. One or two system nodes can be included in an E870 system. All of the system nodes in the server must be the same gigahertz and feature number.

Each optional 19-inch PCIe Gen3 4U I/O Expansion Drawer provides 12 PCIe Gen 3 slots. The I/O expansion drawer connects to the system unit with a pair of PCIe x16 to optical CXP converter adapters housed in the system unit. Each system node can support up to four I/O expansion drawers. Thus, a fully configured Power E870 can support a maximum of eight I/O expansion drawers.

Each EXP24S SFF Gen2-bay Drawer provides 24 x 2.5-inch form-factor (SFF) SAS bays. The EXP24S connects to the Power E870 server using a SAS adapter in a PCIe slot in a system node or in a I/O expansion drawer.

Figure 1-1 shows a single system node Power E870 in a T42 rack with two PCIe I/O drawers and an EXP24S disk drawer.

Figure 1-1 A Power E880 in a T42 rack

1.1.2 Power E880 server

The Power E880 (9119-MHE) server is a powerful POWER8-based system that scales up to 16 sockets. Each socket contains a single 8-core or 12-core POWER8 processor. Thus, a fully-configured Power E880 can scale up to 128 or 192 cores.

The Power E880 is a modular system that is built from a combination of the following four building blocks:

- System control unit
- System node
- PCIe Gen3 I/O expansion drawer
- EXP24S SFF Gen2-bay drawer

The system control unit provides redundant system master clock and redundant system master Flexible Service Processor (FSP) and support for the Operator Panel, the system VPD, and the base DVD. The system control unit provides clock signals to the system nodes with semi-flexible connectors. A system control unit is required for every E880 system.

Each system node provides four processor sockets for POWER8 processors and 32 CDIMM slots supporting a maximum of 8 memory features. Using the 1024 GB memory features, each system node can support a maximum of 8 TB of RAM. Thus a fully configured four-node Power E880 can support up to 32 TB of memory. Each system node provides eight PCIe Gen3 x16 slots. One, two, three, or four system nodes per server are supported. All of the system nodes in the server must be the same gigahertz and feature number.

Each optional 19-inch PCIe Gen3 4U I/O Expansion Drawer provides twelve PCIe Gen 3 slots. The I/O expansion drawer connects to the system unit with a pair of PCIe x16 to Optical CXP converter adapters housed in the system unit. Each system node can support up to four I/O expansion drawers, for a total of 48 PCIe Gen 3 slots. A fully configured Power E880 can support a maximum of 16 I/O expansion drawers.

Each EXP24S SFF Gen2-bay Drawer provides 24 x 2.5-inch form-factor (SFF) SAS bays. The EXP24S is connected to the Power E880 server using a SAS adapter in a PCIe slot in a system node or in a I/O expansion drawer.

Figure 1-2 shows a four system node Power E880 with two PCIe I/O drawers and an EXP24S disk drawer.

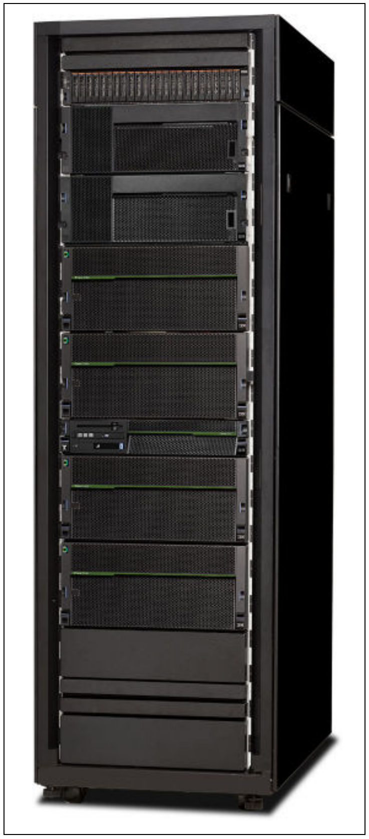


Figure 1-2 A Power E880 in a T42 rack

1.1.3 System control unit

The system control unit (SCU) in a Power E870 and E880 is a new innovation to increase the reliability, availability, and serviceability of the servers. The 2U unit provides redundant clock and service processor capability to Power E870 and E880 systems even if they have only one system unit installed. The SCU also provides a DVD option connected to a PCIe adapter with a USB cable.

The SCU is powered from the system nodes using cables plugged into the first and second system nodes in a two-, three-, or four-system node server. The SCU is powered from the single system node in servers with only one.

The SCU provides redundant clock signalling to the system nodes using semi-flexible connectors. These connectors are located at the rear of the system and route up and down the middle of the system nodes. In this way they do not cause any restrictions on the allowed rack specification.

The SCU provides redundant service processor function to the system nodes using flexible service processor (FSP) cables. Each system node has two FSP connections to the system control unit.

The SCU provides redundant connections to one or two HMCs using 1 Gbps RJ45 Ethernet connections.

Figure 1-3 shows the front and rear view of a system control unit. The locations of the connectors and features are indicated.

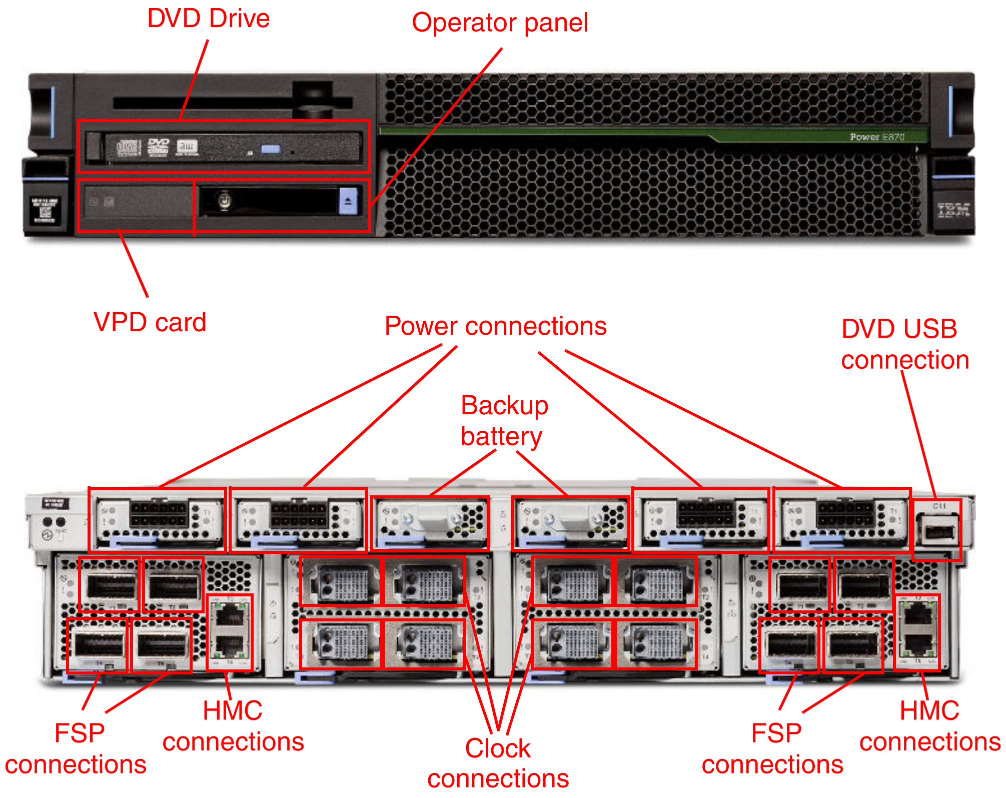


Figure 1-3 Front and rear view of the system control unit

1.1.4 System nodes

The system nodes in the Power E870 and E880 servers host the processors, memory, PCIe slots, and power supplies for the system. Each system node is 5U high and the first and second system nodes in a server connect to the system control unit with FSP, clock, and power connectors. The system node connects to other system nodes using SMP connectors.

Each system node in a Power E870 or E880 server provides four POWER8 processors, 32 CDIMM slots, and eight PCIe Gen3 x16 slots.

Figure 1-4 shows the front view of the system node.

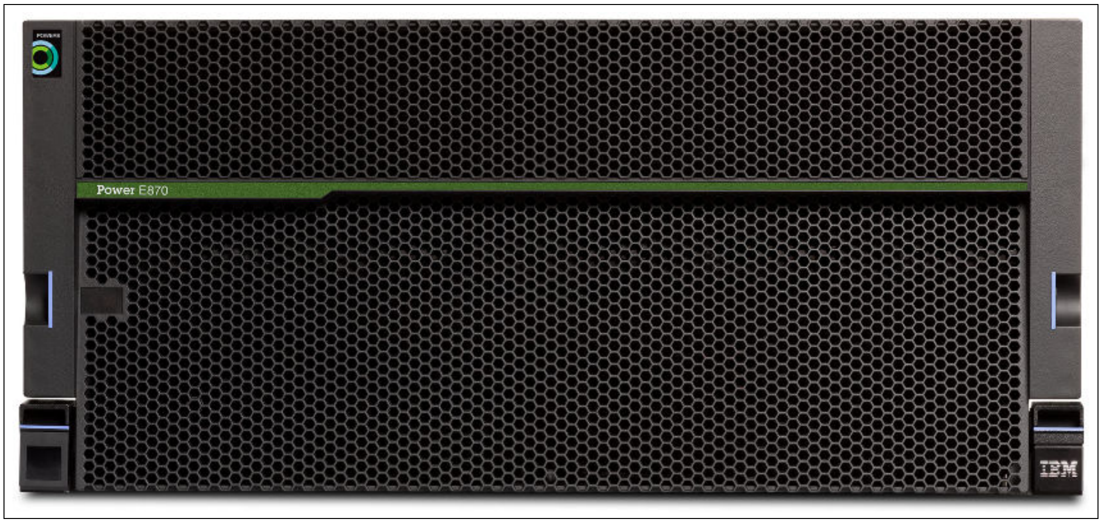


Figure 1-4 Front view of the system node

Figure 1-5 shows the rear view of the system node with notable features highlighted.

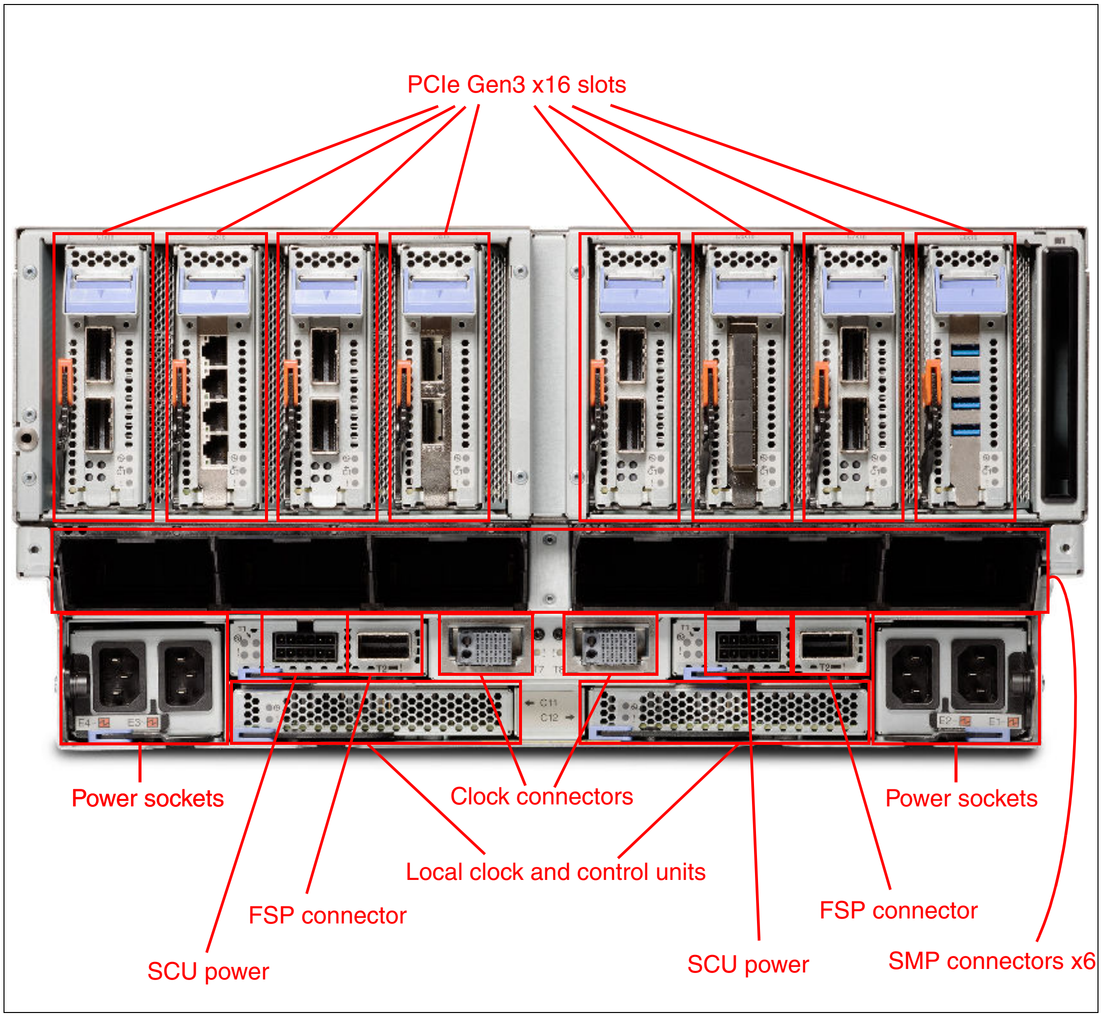


Figure 1-5 Rear view of the system node

Unlike the inter-node flex connectors on the Power 770 and Power 780 servers, the SMP cables on the Power E870 and E880 servers are fully flexible. These cables do not impose restrictions on allowed rack specification. Figure 1-6 shows a diagram of how SMP cables can be routed within a rack on a Power E880 with all four system nodes installed.

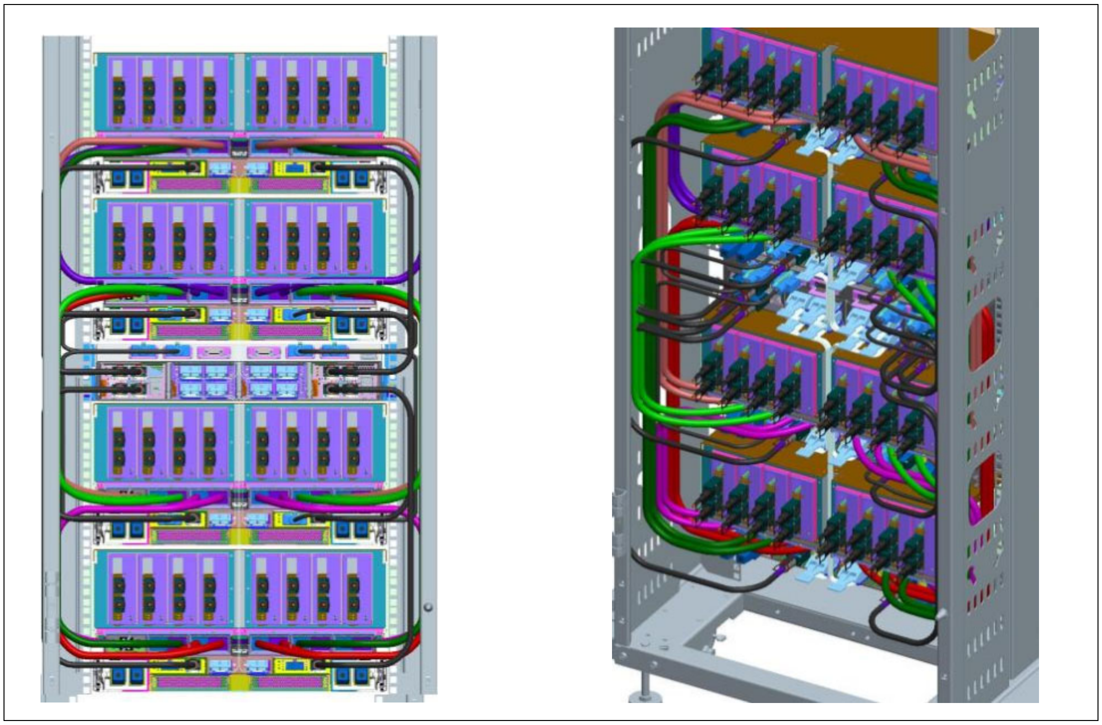


Figure 1-6 An example of SMP cable routing in a four-drawer system

Figure 1-7 shows SMP cable routing in a two-drawer Power E870 or E880.

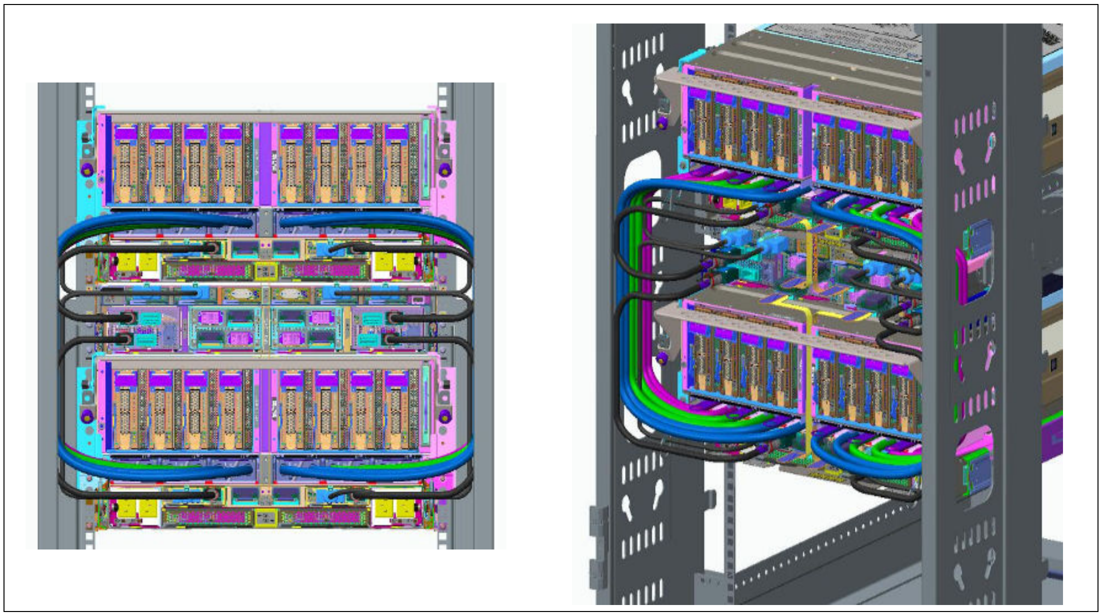


Figure 1-7 SMP cable routing in a two-drawer Power E870 or E880

1.1.5 I/O drawers

I/O drawers to provide additional PCIe adapter slots and SAS disk slots are available for the Power E870 and E880. More details can be found in 1.6, “I/O drawers” on page 22.

1.2 Operating environment

Table 1-1 details the operating environment for the Power E870 and E880 servers.

Table 1-1 Operating environment for Power E870 and Power E880

Power E870 and Power E880 operating environment				
System	**Power E870**	**Power E880**	**Power E870**	**Power E880**
	Operating		**Non-operating**	
Temperature	5 to 35 degrees C (41 to 95 F)		5 to 45 degrees C (41 to 114F)	
Relative humidity	20 - 80%		8 - 80%	
Maximum dew point	29 degrees C (84 F)		Less than 27 degrees C (81 F)	
Operating voltage	200 to 240 V AC 180 - 400 V DC		N/A	
Operating frequency	47 - 63 Hz AC		N/A	
Maximum power consumption	4150 Watts per system node 510 Watts per system control unit		N/A	
Maximum power source loading	4.20 kVA per system node 0.52 kVA per system control unit		N/A	
Maximum thermal output	14,164 BTU/hour per system node 1,740 BTU/hour per system control unit		N/A	
Maximum altitude	3,048 m (10,000 ft.)		N/A	

Power E870 and Power E880 operating environment				
System	**Power E870**	**Power E880**	**Power E870**	**Power E880**
Noise level	Declared A-weighted sound power level, LWad (B) [1, 2, 3]		Declared A-weighted sound pressure level, LpAm (dB) [1, 2, 3]	
	Operating	Idle	Operating	Idle
Model 9119-MHE or 9119-MME in a single drawer configuration.	7.7	7.7	60	60
Model 9119-MHE or 9119-MME in a single drawer configuration. The system is in dynamic power save (favor performance) mode with a heavy workload.	9.5[5]	9.5[5]	79[5]	79[5]
Model 9119-MHE or 9119-MME in a single drawer configuration. The system has acoustical doors (FC 6248 or 6249) installed.	7.15[5]	7.15[5]	56[5]	56[5]
Model 9119-MHE or 9119-MME in a single drawer configuration. The system is in dynamic power save (favor performance) mode with a heavy workload. The system has acoustical doors (FC 6248 or 6249) installed.	8.85[5]	8.85[5]	73[5]	73[5]
Model 9119-MHE or 9119-MME in a 2-drawer configuration.	8.0[4]	8.0[4]	63[4]	63[4]

Power E870 and Power E880 operating environment				
System	**Power E870**	**Power E880**	**Power E870**	**Power E880**
Noise level	Declared A-weighted sound power level, LWad (B) [1, 2, 3]		Declared A-weighted sound pressure level, LpAm (dB) [1, 2, 3]	
	Operating	Idle	Operating	Idle
Model 9119-MHE or 9119-MME in a 2-drawer configuration. The system is in dynamic power save (favor performance) mode with a heavy workload.	9.8[4,5]	9.8[4,5]	82[4,5]	82[4,5]
Model 9119-MHE or 9119-MME in a 2-drawer configuration. The system has acoustical doors (FC 6248 or 6249) installed.	7.45[4,5]	7.45[4,5]	59[4,5]	59[4,5]
Model 9119-MHE or 9119-MME in a 2-drawer configuration. The system is in dynamic power save (favor performance) mode with a heavy workload. The system has acoustical doors (FC 6248 or 6249) installed.	9.15[4,5]	9.15[4,5]	76[4,5]	76[4,5]

Power E870 and Power E880 operating environment				
System	**Power E870**	**Power E880**	**Power E870**	**Power E880**
Noise level	Declared A-weighted sound power level, LWad (B) [1, 2, 3]		Declared A-weighted sound pressure level, LpAm (dB) [1, 2, 3]	
	Operating	Idle	Operating	Idle

Notes:

1. Declared level LWad is the upper-limit A-weighted sound power level. Declared level LpAm is the mean A-weighted emission sound pressure level that is measured at the 1-meter bystander positions.
2. All measurements made in conformance with ISO 7779 and declared in conformance with ISO 9296.
3. 10 dB (decibel) equals 1 B (bel).
4. Estimated.
5. Notice: Government regulations (such as those prescribed by OSHA or European Community Directives) might govern noise level exposure in the workplace and might apply to you and your server installation. This IBM system is available in racks FC 7014-T00 and 7014-T42 with an optional acoustical door feature that can help reduce the noise that is emitted from this system. The actual sound pressure levels in your installation depend upon various factors, including the number of racks in the installation; the size, materials, and configuration of the room where you designate the racks to be installed; the noise levels from other equipment; the room ambient temperature, and employees' location in relation to the equipment. Further, compliance with such government regulations also depends upon various extra factors, including the duration of employees' exposure and whether employees wear hearing protection. IBM recommends that you consult with qualified experts in this field to determine whether you are in compliance with the applicable regulations.
6. A 1.0 B increase in the LWad for a product equals a sound that is approximately twice as loud or twice as noisy.

Environmental assessment: The IBM Systems Energy Estimator tool can provide more accurate information about power consumption and thermal output of systems based on a specific configuration, including adapter cards and I/O expansion drawers. The Energy Estimator tool can be accessed online:

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

1.3 Physical package

Table 1-2 on page 12 lists the physical dimensions of the system control unit and of individual system nodes. Both servers are available only in a rack-mounted form factor.

The Power E870 is a modular system that can be constructed from a single system control unit and one or two system nodes.

The Power E880 is a modular system that can be constructed from a single system control unit and one, two, three, or four system nodes.

The system control unit requires 2U and each system node requires 5U. Thus, a single-enclosure system requires 7U, a two-enclosure system requires 12U, a three-enclosure system requires 17U, and a four-enclosure system requires 22U.

Table 1-2 Physical dimensions of the Power E870 and Power E880

Dimension	System control unit	Power E870 or Power E880 system node	PCIe Gen3 I/O expansion drawer
Width	434 mm (17.1 in)	445 mm (17.5 in)	482 mm (19 in)
Depth	813 mm (32.0in)	902 mm (35.5 in)	802 mm (31.6in)
Height	86 mm (3.4 in) 2 EIA units	219 mm (8.6 in) 5 EIA units	173 mm 6.8 in) 4 EIA units
Weight	23.6 kg (52.0 lb)	75.7 kg (167 lb)	54.4 kg (120 lb)

Figure 1-8 shows a picture of a Power E870 system control unit and system node from the front.

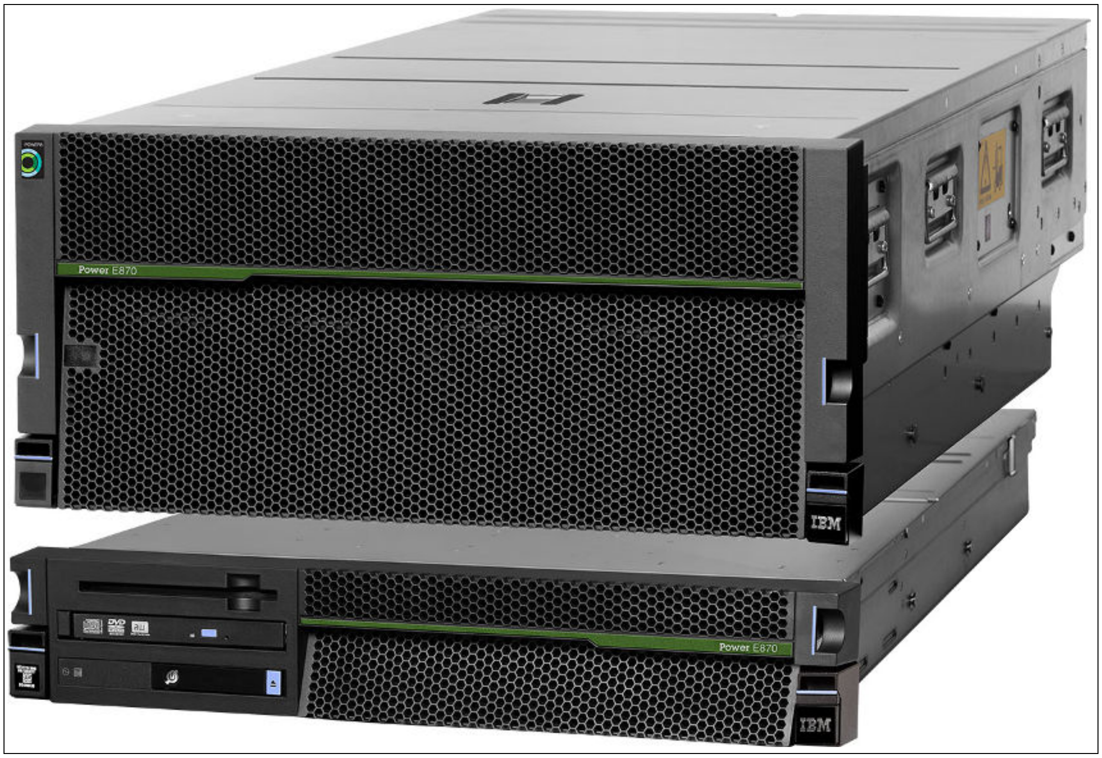


Figure 1-8 Power 870 system control unit and system node

1.4 System features

Both Power E870 and Power E880 system nodes contain four processor modules with 512 KB L2 cache and 8 MB L3 cache per core.

1.4.1 Power E870 system features

The following features are available on the Power E870:

- One or two 5U 19-inch rack mount system node drawers
- One 2U 19-inch rack-mount system control unit drawer

- 5U for a system with one system node drawer
- 12U for a system with two system node drawer
- One processor feature per system node. All system nodes must have the same feature:
  - 4.024 GHz, (4 X 0/8-core) 32-core POWER8 processor (#EPBA)
  - 4.190 GHz, (4 X 0/10-core) 40-core POWER8 processor (#EPBC)
- Static or mobile processor activation features available on a per core basis
- POWER8 DDR3 or DDR4 memory CDIMMs (32 CDIMM slots per system node, 16 sites populated per system node minimum - see 2.3.2, "Memory placement rules" on page 55for #EM8Y restrictions):
  - 0/64 GB (4 X 16 GB), 1600 MHz (#EM8J) DDR3
  - 0/128 GB (4 X 32 GB), 1600 MHz (#EM8K) DDR3
  - 0/256 GB (4 X 64 GB), 1600 MHz (#EM8L) DDR3
  - 0/512 GB (4 x 128 GB), 1600MHz (#EM8M) DDR3
  - 0/1024 GB (4 x 256 GB), 1600 MHz (#EM8Y) DDR4
- IBM Active Memory™ Expansion - optimized onto the processor chip (#EM82)
- Elastic CoD Processor Day - no-charge, temporary usage of inactive, Elastic CoD resources on initial orders (#EPJ3 or #EPJ5)
- 90 Days Elastic CoD Temporary Processor Enablement (#EP9T)
- Eight PCIe Gen3 x16 I/O expansion slots per system node drawer (maximum 16 with 2-drawer system)
- One slim-line, SATA media bay per system control unit enclosure (DVD drive selected as default with the option to de-select)
- Redundant hot-swap AC or DC power supplies in each system node drawer
- Two HMC 1 Gbps ports per Flexible Service Processor (FSP) in the system control unit
- Dynamic logical partition (LPAR) support
- Processor and memory Capacity Upgrade on Demand (CUoD)
- Single root I/O virtualization (SR-IOV)
- PowerVM virtualization (standard and optional):
  - IBM Micro-Partitioning®
  - Dynamic logical partitioning
  - Shared processor pools
  - Shared storage pools
  - Live Partition Mobility
  - Active Memory Sharing
  - Active Memory Deduplication
  - NPIV support
  - IBM PowerVP™ Performance Monitor
- Optional PowerHA for AIX and IBM i
- Optional PCIe Gen3 I/O Expansion Drawer with PCIe Gen3 slots:
  - Zero, one, two, three or four PCIe Gen3 Drawers per system node drawer (#EMX0)
  - Each Gen3 I/O drawer holds two 6-slot PCIe3 Fan-out Modules (#EMXF)

  - Each PCIe3 fan-out module attaches to the system node via two CXP Active Optical Cables (#ECC6, #ECC8, or #ECC9) to one PCIe3 Optical CXP Adapter (#EJ07)

1.4.2 Power E880 system features

The following features are available on the Power E880:

- One, two, three, or four 5U 19-inch rack-mount system node drawers
- One 2U 19-inch rack-mount system control unit drawer
- 7U for a system with one system node drawer plus one system control unit
- 22U for a system with four system node drawers
- One processor feature per system node:
  - 4.35 GHz, (4 X 0/8-core) 32-core POWER8 processor (#EPBB)
  - 4.19 GHz (4 X 0/10-core) 40-core POWER8 processor (#EPBS)
  - 4.02 GHz, (4 X 0/12-core) 48-core POWER8 processor (#EPBD)
- Static or mobile processor activation features available on a per core basis
- POWER8 DDR3 or DDR4 memory CDIMMs (32 CDIMM slots per system node, 16 slots populated per system node minimum - see 2.3.2, "Memory placement rules" on page 55for #EM8Y restrictions):
  - 0/64 GB (4 X 16 GB), 1600 MHz (#EM8J) DDR3
  - 0/128 GB (4 X 32 GB), 1600 MHz (#EM8K) DDR3
  - 0/256 GB (4 X 64 GB), 1600 MHz (#EM8L) DDR3
  - 0/512 GB (4 X 128 GB), 1600 MHz (#EM8M) DDR3
  - 0/1024 GB (4 x 256 GB), 1600 MHz (#EM8Y) DDR4
- Active Memory Expansion - optimized onto the processor chip (#EM82)
- 90 Days Elastic CoD Temporary Processor Enablement (#EP9T)
- Eight PCIe Gen3 x16 I/O expansion slots per system node drawer (maximum 16 with 2-drawer system)
- One slim-line, SATA media bay per system control unit enclosure (DVD drive defaulted on order, option to de-select)
- Redundant hot-swap AC or DC power supplies in each system node drawer
- Two HMC 1 Gbps ports per Flexible Service Processor (FSP) in the system control unit
- Dynamic logical partition (LPAR) support
- Processor and memory CUoD
- Single root I/O virtualization (SR-IOV)
- PowerVM virtualization:
  - Micro-Partitioning
  - Dynamic logical partitioning
  - Shared processor pools
  - Shared storage pools
  - Live Partition Mobility
  - Active Memory Sharing

  - Active Memory Deduplication
  - NPIV support
  - PowerVP Performance Monitor
- Optional PowerHA for AIX and IBM i
- Optional PCIe Gen3 I/O Expansion Drawer with PCIe Gen3 slots:
  - Zero, one, two, three, or four PCIe Gen3 Drawers per system node drawer (#EMX0)
  - Each Gen3 I/O drawer holds two 6-slot PCIe3 Fan-out Modules (#EMXF)
  - Each PCIe3 fan-out module attaches to the system node via two CXP Active Optical Cables (#ECC6, #ECC8, or #ECC9) to one PCIe3 Optical CXP Adapter (#EJ07)

1.4.3 Minimum features

Each Power E870 or Power E880 initial order must include a minimum of the following items:

- 1 x system node with a choice of the following processors:
  - 4.024 GHz, 32-core POWER8 processor module (#EPBA) for Power E870
  - 4.190 GHz, 40-core POWER8 processor module (#EPBC) for Power E870
  - 4.35 GHz, 32-core POWER8 processor module (#EPBB) for Power E880
  - 4.19 GHz, 40-core POWER8 processor module (#EPBS) for Power_E880
  - 4.02 GHz, 48-core POWER8 processor module (#EPBD) for Power E880
- 8 x 1 core processor activation:
  - 1 core permanent Processor Activation for #EPBA (#EPBJ) for Power E870
  - 1 core permanent Processor Activation for #EPBC (#EPBL) for Power E870
  - 1 core permanent Processor Activation for #EPBB (#EPBK) for Power E880
  - 1 core permanent Processor Activation for #EPBS (#EPBU) for Power E880
  - 1 core permanent Processor Activation for #EPBD (#EPBM) for Power E880
- 4 x 64 GB (4 x 16 GB) CDIMMs, 1600 MHz, 4 Gb DDR3 DRAM (#EM8J)
- Total of 50% of installed memory must be activated with a combination of permanent or mobile activation using features:
  - 1 GB Memory Activation (#EMA5)
  - 100 GB Memory Activations (#EMA6)
  - 100 GB Mobile Activations (#EMA9)
- A maximum of 75% of memory activations can be mobile activations
- 1 x 5U system node drawer (#EBA0)
- 2 x Service Processor (#EU0A)
- 1 x Load Source Specify:
  - EXP24S SFF Gen2 (#5887 or #EL1S) Load Source Specify (#0728)
  - SAN Load Source Specify (#0837)
- 1 x Rack-mount Drawer Bezel and Hardware:
  - IBM Rack-mount Drawer Bezel and Hardware (#EBA2) for Power E870
  - IBM Rack-mount Drawer Bezel and Hardware (#EBA3) for Power E880

  - OEM Rack-mount Drawer Bezel and Hardware (#EBA4)
- 1 x System Node to System Control Unit Cable Set for Drawer 1 (#ECCA)
- 1 x Power Chunnels for routing power cables from back of machine to front (#EBAA)
- 1 x Language Group Specify (#9300/#97xx)
- Optional 1 x Media:
  - SATA Slimline DVD-RAM with write cache (#EU13)
  - PCIe2 LP 4-Port USB 3.0 Adapter (#EC45)
  - PCIe2 4-Port USB 3.0 Adapter (#EC46)

When either AIX or Linux are the primary operating systems, the order must include a minimum of the following items also:

- 1 x Primary Operating System Indicator:
  - Primary Operating System Indicator - AIX (#2146)
  - Primary Operating System Indicator - Linux (#2147)

When IBM i is the primary operating system, the order must include a minimum of the following items:

- 1 x Specify Code:
  - Mirrored System Disk Level, Specify Code (#0040)
  - Device Parity Protection-All, Specify Code (#0041)
  - Mirrored System Bus Level, Specify Code (#0043)
  - Device Parity RAID 6 All, Specify Code (#0047)
  - Mirrored Level System Specify Code (#0308)
- 1 x System Console:
  - Sys Console On HMC (#5550)
  - System Console-Ethernet No IOP (#5557)
- 1 x Primary Operating System Indicator - IBM i (#2145)

Notes:

- Additional optional features can be added, as desired.
- IBM i systems require a DVD to be available to the system when required. This DVD can be located in the system control unit (DVD feature #EU13) or it can be located externally in an enclosure like the 7226-1U3. A USB PCIe adapter such as #EC45 or #EC46 is required for #EU13. A SAS PCIe adapter such as #EJ11 is required to attach a SATA DVD in the 7226-1U3. A virtual media repository may be used to substitute for a DVD device.
- Feature-coded racks are allowed for I/O expansion only.
- A machine type/model rack, if desired, should be ordered as the primary rack.
- A minimum number of eight processor activations must be ordered per system.
- A minimum of four memory features per drawer are required.
- At least 50% of available memory must be activated through a combination of feature #EMA5 and #EMA6, and #EMA9. 25% of the available memory must be permanently activated using #EMA5 and #EMA6. The remaining 75% of the active memory can be enabled using either permanent or mobile options.
- Memory sizes can differ across CPU modules, but the eight CDIMM slots connected to the same processor module must be filled with identical CDIMMs.
- If SAN Load Source Specify (#0837) is ordered #0040, #0041, #0043, #0047, #0308 are not supported.
- The language group is auto-selected based on geographic rules.
- No feature codes are assigned for the following items:
  - Four AC power supplies are delivered as part of the system node. No features are assigned to power supplies. Four line cords are auto-selected according to geographic rules.
  - There must be one system control unit on each system. The system control unit is considered the system with the serial number.
- An HMC is required for management of every Power E870 or Power E880; however, a shared HMC is acceptable. HMCs supported on POWER8 hardware are 7042-C08 and 7042-CR5 through 7042-CR9. For more details see 1.12, “Management consoles” on page 29

1.4.4 Power supply features

This section describes how the system nodes and system control units are powered.

System node power

Four AC or DC power supplies are required for each system node enclosure. This arrangement provides 2+2 redundant power with dual power sources for enhanced system availability. A failed power supply can be hot swapped but must remain in the system until the replacement power supply is available for exchange.

Four AC or DC power cords are used for each system node (one per power supply) and are ordered using the AC Power Chunnel feature (#EBAA). Each #EBAA provides all four AC power cords, therefore a single #EBAA should be ordered per system node. The chunnel carries power from the rear of the system node to the hot swap power supplies located in the front of the system node where they are more accessible for service.

System control unit power

The system control unit is powered from the system nodes. UPIC cables provide redundant power to the system control unit. Two UPIC cables attach to system node drawer 1 and two UPIC cables attach to system node drawer 2. They are ordered with #ECCA and #ECCB. The UPIC cords provides N+1 redundant power to the system control unit.

1.4.5 Processor card features

Each system node will deliver a set of four identical processors. All processors in the system must be identical. Cable features are required to connect system nodes to the system control unit and to other system nodes.

- For a single system node configuration, #ECCA is required. #ECCA provides cables to connect the system node with the system control unit.
- For a dual system node configuration, #ECCB is required. #ECCB provides cables to connect the system nodes with the system control unit and cables to connect the two system nodes.
- For three or four system node systems there are no additional cables required, as redundant power is provided through system nodes 1 and 2.

Each system must have a minimum of eight active cores:

The Power E870 has two types of processors, offering the following features:

- 4.024 GHz, (4 X 0/8-core) 32-core POWER8 processor (#EPBA)
- 4.190 GHz, (4 X 0/10-core) 40-core POWER8 processor (#EPBC)

The Power E880 has three types of processors, offering the following features:

- 4.35 GHz, (4 X 0/8-core) 32-core POWER8 processor (#EPBB)
- 4.19 GHz (4 X 0/10-core) 40-core POWER8 processor (#EPBS)
- 4.02 GHz, (4 X 0/12-core) 48-core POWER8 processor (#EPBD)

Several types of capacity on demand (CoD) processor options are available on the Power E870 and Power 880 servers to help meet changing resource requirements in an on-demand environment by using resources installed on the system but not activated. CoD allows you to purchase additional permanent processor or memory capacity and dynamically activate it when needed. The #EPJ3 provides no-charge elastic processor days for Power E880. The #EMJ8 feature provides no-charge elastic memory days for Power E880.

1.4.6 Summary of processor features

Table 1-3 summarizes the processor feature codes for the Power E870.

Table 1-3 Summary of processor features for the Power E870

Feature Code	Description
EPBA	4.02 GHz, 32-core POWER8 processor
EPBC	4.19 GHz, 40-core POWER8 processor
EB2R	Single 5250 Enterprise Enablement, IBM i
EB30	Full 5250 Enterprise Enablement, IBM i
EP2S	1-Core Mobile Activation

Feature Code	Description
EP2U	1-Core Mobile Activation from POWER® 7
EP9T	90 Days Elastic CoD Processor Core Enablement
EPBJ	1 core Processor Activation for #EPBA
EPBL	1 core Processor Activation for #EPBC
EPBN	1 core Processor Activation for #EPBA, Mobile Enabled
EPBQ	1 core Processor Activation for #EPBC, Mobile Enabled
EPJ6	1 Proc-Day Elastic CoD Billing for #EPBA, AIX/Linux
EPJ7	1 Proc-Day Elastic CoD Billing for #EPBA, IBM i
EPJ8	100 Elastic CoD Proc-Days of Billing for Processor #EPBA. AIX/Linux
EPJ9	100 Elastic CoD Proc-Days of Billing for Processor #EPBA. IBM i
EPJA	Proc CoD Utility Billing, 100 Proc-mins. for #EPBA, AIX/Linux
EPJB	Proc CoD Utility Billing, 100 Proc-mins. for #EPBA, IBM i
EPJJ	1 Proc-Day Elastic CoD Billing for #EPBC, AIX/Linux
EPJK	1 Proc-Day Elastic CoD Billing for #EPBC, IBM i
EPJL	100 Elastic CoD Proc-Days of Billing for Processor #EPBC. AIX/Linux
EPJM	100 Elastic CoD Proc-Days of Billing for Processor #EPBC. IBM i
EPJN	Proc CoD Utility Billing, 100 Proc-mins. for #EPBC, AIX/Linux
EPJP	Proc CoD Utility Billing, 100 Proc-mins. for #EPBC, IBM i
ELJG	Power Integrated Facility for Linux, adds #ELJ8 for 4 activations[a]

[a] This feature activates 4 processors (#ELJ8), as well as 32GB memory (#ELJH) and 4 licenses for PowerVM for Linux (ELJJ)

Table 1-4 summarizes the processor feature codes for the Power E880.

Table 1-4 Summary of processor features for the Power E880

Feature Code	Description
EPBB	4.35 GHz, 32-core POWER8 processor
EPBS	4.19 GHz, 40-core POWER8 processor
EPBD	4.02 GHz, 48-core POWER8 processor
EB2R	Single 5250 Enterprise Enablement
EB30	Full 5250 Enterprise Enablement
EP2T	1-Core Mobile Activation
EP2V	1-Core Mobile Activation from POWER 7
EP9T	90 Days Elastic CoD Processor Core Enablement
EPBK	1 core Processor Activation for #EPBB
EPBU	1 core Processor Activation for #EPBS
EPBM	1 core Processor Activation for #EPBD

Feature Code	Description
EPBP	1 core Processor Activation for #EPBB, Mobile Enabled
EPBV	1 core Processor Activation for #EPBS, Mobile Enabled
EPBR	1 core Processor Activation for #EPBD, Mobile Enabled
EPJ3	48 Proc-Days of Elastic CoD Processor Resources for #EPBB
EPJ5	48 Proc-Days of Elastic CoD Processor Resources for #EPBD
EPJC	1 Proc-Day Elastic CoD Billing for #EPBB, AIX/Linux
EPJQ	1 Proc-Day Elastic CoD Billing for #EPBD, AIX/Linux
EPJD	1 Proc-Day Elastic CoD Billing for #EPBB, IBM i
EPQR	1 Proc-Day Elastic CoD Billing for #EPBD, IBM i
EPJE	100 Elastic CoD Proc-Days of Billing for Processor #EPBB. AIX/Linux
EPJS	100 Elastic CoD Proc-Days of Billing for Processor #EPBD. AIX/Linux
EPJF	100 Elastic CoD Proc-Days of Billing for Processor #EPBB. IBM i
EPJT	100 Elastic CoD Proc-Days of Billing for Processor #EPBD. IBM i
EPJG	Proc CoD Utility Billing, 100 Proc-mins. for #EPBB, AIX/Linux
EPJU	Proc CoD Utility Billing, 100 Proc-mins. for #EPBD, AIX/Linux
EPJH	Proc CoD Utility Billing, 100 Proc-mins. for #EPBB, IBM i
EPJV	Proc CoD Utility Billing, 100 Proc-mins. for #EPBD, IBM i
ELJG	Power Integrated Facility for Linux, adds #ELJ8 for 4 activations[a]

[a] This feature activates 4 processors (#ELJ8), as well as 32GB memory (#ELJH) and 4 licenses for PowerVM for Linux (ELJJ)

1.4.7 Memory features

1600 MHz memory CDIMMs are available as 64 GB (#EM8J), 128 GB (#EM8K), 256 GB (#EM8L), 512 GB (#EM8M), and 1024 (#EM8Y) memory features. Each memory feature provides four CDIMMs. CDIMMs are custom DIMMs which enhance memory performance and memory reliability. Each system node has 32 CDIMM slots which support a maximum of eight memory features. Memory activations of 50% of the installed capacity are required. The feature #EMJ8 provides no-charge elastic memory days for the Power E880.

Table 1-5 lists memory features that are available for the Power E870 and E880.

Table 1-5 Summary of memory features

Feature Code	Description
EM8J	64 GB (4X16 GB) CDIMMs, 1600 MHz, 4 GBIT DDR3 DRAM
EM8K	128 GB (4X32 GB) CDIMMs, 1600 MHz, 4 GBIT DDR3 DRAM
EM8L	256 GB (4X64 GB) CDIMMs, 1600 MHz, 4 GBIT DDR3 DRAM
EM8M	512 GB (4X128 GB) CDIMMs, 1600 MHz, 4 GBIT DDR3 DRAM
EM8Y	1024 GB (4X 256 GB) CDIMMs, 1600 MHz, 4 GBIT DDR4 DRAM

Feature Code	Description
EMB6	Bundle of eight #EM8M, 512 GB 1600 MHz Memory features
EM82	Active Memory Expansion Enablement
EM9T	90 Days Elastic CoD Memory Enablement
EMA5	1 GB Memory Activation
EMA6	Quantity of 100 1 GB Memory Activations (#EMA5)
EMA7	100 GB Mobile Memory Activations
EMA9	100 GB Mobile Enabled Memory Activations
EMB7	Bundle of forty, #EMA6 1600 MHz memory features
EMB8	Bundle of 512 #EMA5 1600 MHz memory features
EMJ4	1 GB-Day billing for Elastic CoD memory
EMJ5	100 GB-Day billing for Elastic CoD memory
EMJ6	999 GB-Day billing for Elastic CoD memory
EMJ8	384 GB-Days of Elastic CoD Memory Resources (E880 only)
ELJG	Power Integrated Facility for Linux, adds #ELJH for 32 GB activations[a]

[a] This feature activates 4 processors (#ELJ8), as well as 32GB memory (#ELJH) and 4 licenses for PowerVM for Linux (ELJJ)

1.4.8 System node PCIe slots

Each Power E870 or Power E880 system node provides eight half-length, half-high x16 PCIe Gen3 slots. These PCIe slots can be used for either low profile PCIe adapters or for attaching a PCIe I/O drawer.

A new form factor blind swap cassette (BSC) is used to house the low profile adapters, which go into these slots.

PCIe Gen1, Gen2, and Gen3 adapter cards are supported in these Gen3 slots.

Table 1-6 provides details of the PCI slots in the Power E870 and Power E880 system.

Table 1-6 PCIe slot locations and descriptions for the Power E870 and Power E880

Slot	Location code	Description	PHB	Adapter size
Slot 1	P1-C1	PCIe3, x16	Processor Module 1, PHB1	Short (low-profile)
Slot 2	P1-C2	PCIe3, x16	Processor Module 1, PHB0	Short (low-profile)
Slot 3	P1-C3	PCIe3, x16	Processor Module 2, PHB1	Short (low-profile)
Slot 4	P1-C4	PCIe3, x16	Processor Module 2, PHB0	Short (low-profile)
Slot 5	P1-C5	PCIe3, x16	Processor Module 3, PHB1	Short (low-profile)
Slot 6	P1-C6	PCIe3, x16	Processor Module 3, PHB0	Short (low-profile)
Slot 7	P1-C7	PCIe3, x16	Processor Module 4, PHB1	Short (low-profile)
Slot 8	P1-C8	PCIe3, x16	Processor Module 4, PHB0	Short (low-profile)

- All slots support enhanced error handling (EEH).
- All PCIe slots are hot swappable and support concurrent maintenance.

1.5 Disk and media features

The Power E870 and Power E880 system control unit and system nodes do not support internal disks. Any required disk must reside within a SAN disk subsystem or an external disk drawer. The EXP24S SFF Gen2-bay Drawer (#5887) is the only supported disk drawer for Power E870 or Power E880.

Each system control unit enclosure has one slim-line bay which can support one DVD drive (#EU13). The #EU13 DVD is cabled to a USB PCIe adapter located in either a system node (#EC45) or in a PCIe Gen3 I/O drawer (#EC46). A USB to SATA converter is included in the configuration without a separate feature code.

IBM i systems require a DVD to be available to the system when required. This DVD can be located in the system control unit (DVD feature #EU13) or it can be located externally in an enclosure like the 7226-1U3. A USB PCIe adapter such as #EC45 or #EC46 is required for #EU13. A SAS PCIe adapter such as #EJ11 is required to attach a SATA DVD in the 7226-1U3. A virtual media repository may be used to substitute for a DVD device if using VIOS.

1.6 I/O drawers

If additional Gen3 PCIe slots beyond the system node slots are required, a system node x16 slot is used to attach a 6-slot expansion module in the I/O Drawer. A PCIe Gen3 I/O expansion drawer (#EMX0) holds two expansion modules which are attached to any two x16 PCIe slots in the same system node or in different system nodes.

Disk-only I/O drawers (#5887) are also supported, providing storage capacity.

1.6.1 PCIe Gen3 I/O expansion drawer

The 19-inch 4 EIA (4U) PCIe Gen3 I/O expansion drawer (#EMX0) and two PCIe FanOut Modules (#EMXF) provide twelve PCIe I/O full-length, full-height slots. One FanOut Module provides six PCIe slots labeled C1 through C6. C1 and C4 are x16 slots and C2, C3, C5, and C6 are x8 slots. PCIe Gen1, Gen2, and Gen3 full-high adapter cards are supported.

A blind swap cassette (BSC) is used to house the full-high adapters that go into these slots. The BSC is the same BSC as used with the previous generation server's 12X attached I/O drawers (#5802, #5803, #5877, #5873). The drawer is shipped with a full set of BSC, even if the BSC is empty.

Concurrent repair and add/removal of PCIe adapter cards is done through HMC guided menus or by operating system support utilities.

A PCIe x16 to Optical CXP converter adapter (#EJ07) and 2.0 M (#ECC6), 10.0 M (#ECC8), or 20.0 M (#ECC9) CXP 16X Active Optical cables (AOC) connect the system node to a PCIe FanOut module in the I/O expansion drawer. One feature #ECC6, #ECC8, or #ECC9 ships two AOC cables. Each PCIe Gen3 I/O expansion drawer has two power supplies.

Each system node supports zero, one, two, three, or four PCIe Gen3 I/O expansion drawers. A half drawer, consisting of just one PCIe fan-out module in the I/O drawer is also supported, allowing a lower-cost configuration if fewer PCIe slots are required. Thus a system node supports the following half drawer options: one half drawer, two half drawers, three half drawers, or four half drawers. Because there is a maximum of four #EMX0 drawers per node, a single system node cannot have more than four half drawers. A server with more system nodes can support more half drawers up to four per system node. A system can also mix half drawers and full PCIe Gen3 I/O expansion drawers. The maximum of four PCIe Gen3 drawers per system node applies whether a full or half PCIe drawer.

Drawers can be added to the server at a later time, but system downtime must be scheduled for adding a PCIe3 Optical Cable Adapter or a PCIe Gen3 I/O drawer (EMX0) or fan-out module (#EMXF).

Figure 1-9 shows a PCIe Gen3 I/O expansion drawer.

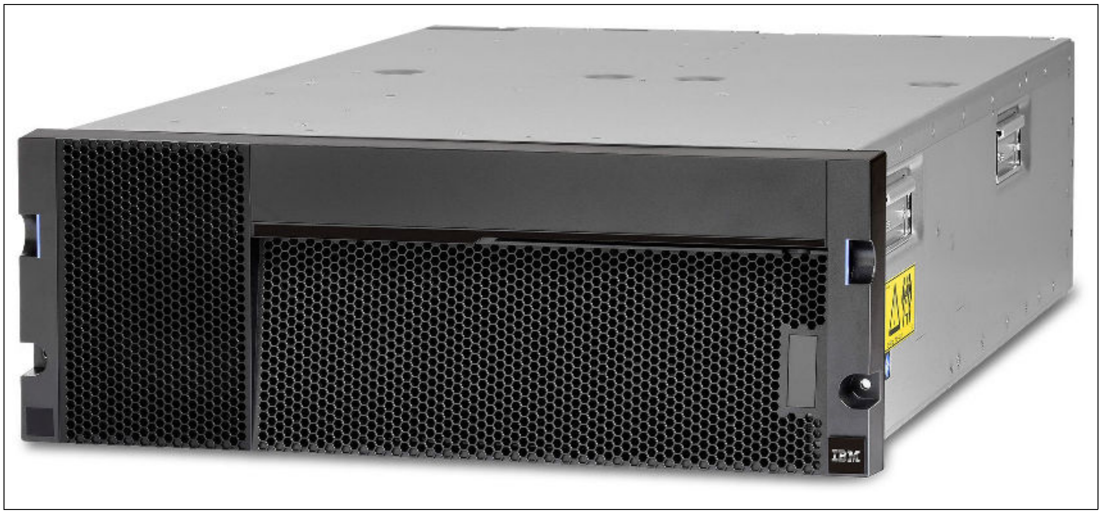


Figure 1-9   PCIe Gen3 I/O expansion drawer

For more details on connecting PCIe Gen3 I/O expansion drawers to the Power E870 and Power E880 servers, see 2.9.1, “PCIe Gen3 I/O expansion drawer” on page 85

1.6.2 I/O drawers and usable PCI slot

Figure 1-10 shows the rear view of the PCIe Gen3 I/O expansion drawer with the location codes for the PCIe adapter slots in the PCIe3 6-slot fanout module.

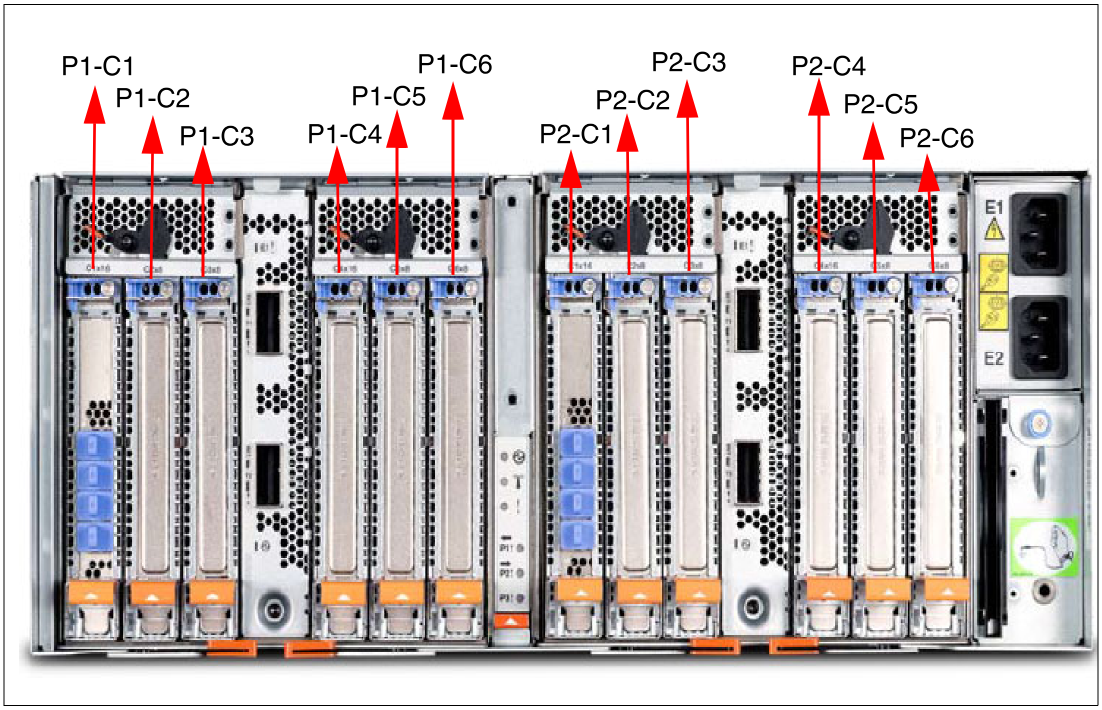


Figure 1-10 Rear view of a PCIe Gen3 I/O expansion drawer with PCIe slots location codes

Table 1-7 provides details of the PCI slots in the PCIe Gen3 I/O expansion drawer.

Table 1-7 PCIe slot locations and descriptions for the PCIe Gen3 I/O expansion drawer

Slot	Location code	Description
Slot 1	P1-C1	PCIe3, x16
Slot 2	P1-C2	PCIe3, x8
Slot 3	P1-C3	PCIe3, x8
Slot 4	P1-C4	PCIe3, x16
Slot 5	P1-C5	PCIe3, x8
Slot 6	P1-C6	PCIe3, x8
Slot 7	P2-C1	PCIe3, x16
Slot 8	P2-C2	PCIe3, x8
Slot 9	P2-C3	PCIe3, x8
Slot 10	P2-C4	PCIe3, x16
Slot 11	P2-C5	PCIe3, x8
Slot 12	P2-C6	PCIe3, x8

- All slots support full-length, regular-height adapter or short (low-profile) with a regular-height tailstock in single-wide, generation 3, blind-swap cassettes.

- Slots C1 and C4 in each PCIe3 6-slot fanout module are x16 PCIe3 buses and slots C2, C3, C5, and C6 are x8 PCIe buses.
- All slots support enhanced error handling (EEH).
- All PCIe slots are hot swappable and support concurrent maintenance.

Table 1-8 summarizes the maximum number of I/O drawers supported and the total number of PCI slots that are available when expansion consists of a single drawer type.

Table 1-8 Maximum number of I/O drawers supported and total number of PCI slots

System nodes	**Maximum #EMX0 drawers**	**Total number of slots**		
		PCIe3, x16	PCIe3, x8	Total PCIe3
1 system node	4	16	32	48
2 system nodes	8	32	64	96
3 system nodes	12	48	96	144
4 system nodes	16	64	128	192

1.7 EXP24S SFF Gen2-bay Drawer

The EXP24S SFF Gen2-bay Drawer (#5887) is an expansion drawer with 24 2.5-inch form-factor (SFF) SAS bays. The EXP24S supports up to 24 hot-swap SFF-2 SAS hard disk drives (HDDs) or solid state drives (SSDs). It uses 2 EIA of space in a 19-inch rack. The EXP24S includes redundant AC power supplies and uses two power cords.

With AIX, Linux, and VIOS, you can order the EXP24S with four sets of six bays, two sets of 12 bays, or one set of 24 bays (mode 4, 2, or 1). With IBM i, you can order the EXP24S as one set of 24 bays (mode 1). Mode setting is done by IBM Manufacturing. If you need to change the mode after installation, ask your IBM support rep. to refer to:

http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5121

The EXP24S SAS ports are attached to a SAS PCIe adapter or pair of adapters using SAS YO or X cables.

Figure 1-11 shows the EXP24S SFF Gen2-bay drawer.

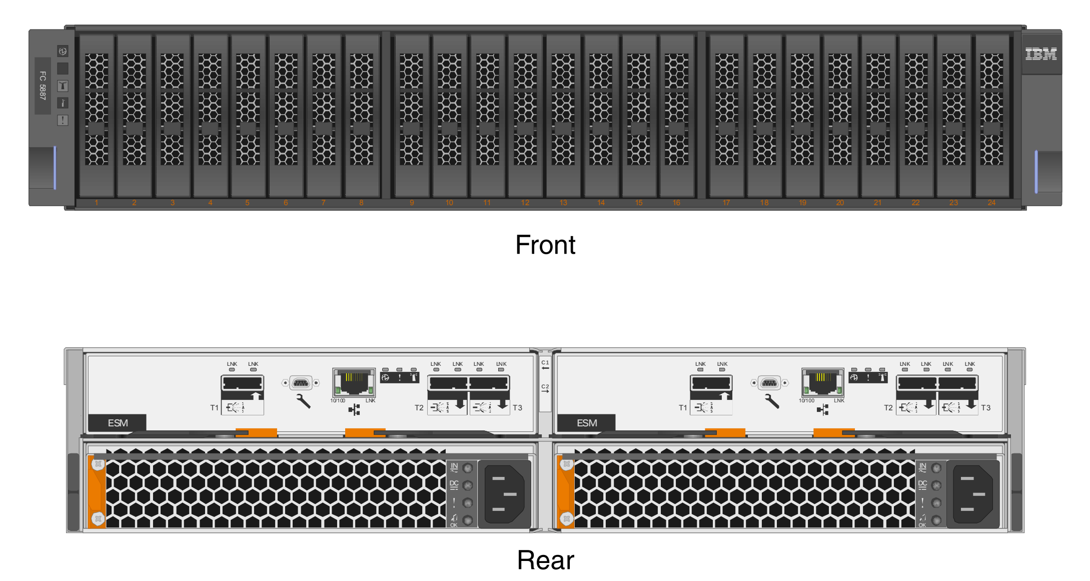


Figure 1-11 EXP24S SFF Gen2-bay drawer

For more information on connecting the EXP24S Gen2-bay drawer to the Power E870 and Power E880 servers, see 2.10.1, “EXP24S SFF Gen2-bay Drawer” on page 92.

1.8 Comparison between models

The Power E870 offers configuration options where the POWER8 processor can have one of two processor speeds. Each system node is populated with four single chip modules (SCMs). Each system node will contain one of the following processor configurations:

- Four 4.02 GHz 8-core SCMs
- Four 4.19 GHz 10-core SCMs

A Power E870 with either of the processor configurations can have as few as eight cores activated or up to 100% of the cores can be activated. Incrementing one core at a time is available through built-in capacity on demand (CoD) functions to the full capacity of the system. The Power E870 can be installed with one or two system nodes connected to the system control unit. Each system node can have up to 8 TB of memory installed.

The Power E880 also offers system nodes populated with four SCMs. The Power E880 system node has two processor configuration:

- Four 4.35 GHz 8-core SCMs
- Four 4.19 GHz 10-core SCMs
- Four 4.02 GHz 12-core SCMs

A Power E880 with either of the processor configurations can have as few as eight cores activated or up to 100% of the cores can be activated. Incrementing one core at a time is available through built-in capacity on demand (CoD) functions to the full capacity of the system. The Power E880 can be installed with one, two, three, or four system nodes connected to the system control unit. Each system node can have up to 8 TB of memory installed.

Table 1-9 shows a summary of processor and memory maximums for the Power E870 and Power E880.

Table 1-9 Summary of processor and memory maximums for the Power E870 and Power E880

System	Cores per SCM	Core speed	System node core maximum	System core maximum	System node memory maximum	System memory maximum
Power E870	8	4.02 GHz	32	64	8 TB	8 TB
Power E870	10	4.19 GHz	40	80	8 TB	16 TB
Power E880	8	4.35 GHz	32	128	8 TB	32 TB
Power E880	10	4.19 GHz	40	160	8 TB	32 TB
Power E880	12	4.02 GHz	48	192	8 TB	32 TB

1.9 Build to order

You can do a *build to order* (also called *a la carte*) configuration by using the IBM Configurator for e-business (e-config). With it, you specify each configuration feature that you want on the system.

This method is the only configuration method for the Power E870 and Power E880 servers.

1.10 IBM editions

IBM edition offerings are not available for the Power E870 and Power E880 servers.

1.11 Model upgrades

The following sections describe the various upgrades that are available.

1.11.1 Power E870

A model conversion from a Power 770 (9117-MMD) to a Power E870 (9119-MME) is available. One-step upgrades from previous models of the Power 770 (9117-MMB and 9117-MMC) are not available. To upgrade from a 9117-MMB or 9117-MMC, an upgrade to a 9117-MMD is required first.

The existing components being replaced during a model or feature conversion become the property of IBM and must be returned.

1.11.2 Power E880

A model conversion from a Power 780 (9179-MHD) to a Power E880 (9119-MHE) is available. One-step upgrades from previous models of the Power 780 (9179-MHB and 9179-MHC) are not available. To upgrade from a 9179-MHB or 9179-MHC, an upgrade to a 9179-MHD is required first.

A model conversion from a Power 770 (9117-MMD) to a Power E880 (9119-MHE) is also available. One-step upgrades from previous models of the Power 770 (9117-MMB and

9117-MMC) are not available. To upgrade from a 9117-MMB or 9117-MMC, an upgrade to a 9117-MMD is required first.

Upgrades to a Power E880 from a Power 795 (9119-FHB) are not available.

The existing components being replaced during a model or feature conversion become the property of IBM and must be returned.

Statement of direction:

For clients with Power E880 servers that require additional system capacity and growth, IBM plans to enhance the Power Systems' Enterprise system portfolio with greater flexibility by providing a serial number protected upgrade path from the Power E880 32-core system node server to the Power E880 48-core system node server.

1.11.3 Upgrade considerations

Feature conversions are set up for the following items:

- PCIe Crypto blind swap cassettes
- Power IFL processor activations
- Power IFL PowerVM for Linux
- Active Memory Expansion Enablement
- DDR3 memory DIMMS to CDIMMS
- Static and mobile memory activations
- 5250 enterprise enablement
- POWER7+™ processor cards to POWER8 processors
- Static and mobile processor activations
- System CEC enclosure and bezel to 5U system node drawer
- PowerVM standard and enterprise

The following features that are present on the current system can be moved to the new system if they are supported in the Power E870 and E880:

- Disks (within an EXP24S I/O drawer)
- SSDs (within an EXP24S I/O drawer)
- PCIe adapters with cables, line cords, keyboards, and displays
- Racks
- Doors
- EXP24S I/O drawers

For POWER7+ processor-based systems that have the Elastic CoD function enabled, you must reorder the elastic CoD enablement features when placing the upgrade MES order for the new Power E870 or E880 system to keep the elastic CoD function active. To initiate the model upgrade, the on/off enablement features should be removed from the configuration file before the MES order is started. Any temporary use of processors or memory owed to IBM on the existing system must be paid before installing the Power E870 or E880.

1.12 Management consoles

This section discusses the supported management interfaces for the servers.

The Hardware Management Console (HMC) is required for managing the IBM Power E870 and Power E880. It has a set of functions that are necessary to manage the system:

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

Multiple Power Systems servers can be managed by a single HMC. Each server can be connected to multiple HMC consoles to build extra resiliency into the management platform.

In 2015, the 7042-CR9 was announced.

The IBM Power E870 and Power E880 are not supported by the Integrated Virtualization Manager (IVM).

Several HMC models are supported to manage POWER8 based systems. The 7042-CR9 is are available for ordering at the time of writing, but you can also use one of the withdrawn models listed in Table 1-10.

Table 1-10 HMC models supporting POWER8 processor technology-based servers

Type-model	Availability	Description
7042-C08	Withdrawn	IBM 7042 Model C08 Deskside Hardware Management Console
7042-CR5	Withdrawn	IBM 7042 Model CR5 Rack-Mounted Hardware Management Console
7042-CR6	Withdrawn	IBM 7042 Model CR6 Rack mounted Hardware Management Console
7042-CR7	Withdrawn	IBM 7042 Model CR7 Rack mounted Hardware Management Console
7042-CR8	Withdrawn	IBM 7042 Model CR8 Rack mounted Hardware Management Console
7042-CR9	Available	IBM 7042 Model CR9 Rack mounted Hardware Management Console

HMC base Licensed Machine Code Version 8 Revision 8.2.0 or later is required to support the Power E870 (9119-MME) and Power E880 (9119-MHE).

System firmware level 8.3 or later, is a prerequisite for the 48-core system node, Power E880 systems with three or four system nodes, or for system nodes supporting more than two I/O expansion drawers or half drawers. This will require HMC firmware level 8.3.0 or higher.

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

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

Existing HMC models 7042 can be upgraded to Licensed Machine Code Version 8 to support environments that might include IBM POWER6®, IBM POWER6+™, IBM POWER7®, IBM POWER7+ and POWER8 processor-based servers.

If you want to support more than 254 partitions in total, the HMC will require a memory upgrade to a minimum of 4 GB.

For further information on managing the Power E870 and Power E880 servers from an HMC, see 2.11, “Hardware Management Console” on page 104.

1.13 System racks

The Power E870 and E880 and its I/O drawers are designed to be mounted in the following existing IBM racks: 7014-T00, 7014-T42, 7965-94Y, #0553, #0551, and #ER05.

However, for initial system orders, the racks must be ordered as machine type 7014-T42, #0553. The 36U (1.8-meter) rack (#0551) and the 42U (2.0-meter) rack (#ER05) are available to order only on Miscellaneous Equipment Specification (MES) upgrade orders.

Shipping without a rack: If you require the system to be shipped without an IBM rack, feature code #ER21 must be used to remove the IBM rack from the order. The server will then ship as separate packages for installation into an existing rack.

The Power E870 and Power E880 use a new type of connector between system drawers. Therefore, the systems do not require wider racks, and an OEM rack or cabinet which meet the requirements can be used.

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

1.13.1 IBM 7014 model T00 rack

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

- 36U (EIA units) of usable space.
- Optional removable side panels.
- Optional highly perforated front door.
- Optional side-to-side mounting hardware for joining multiple racks.
- Standard business black or optional white color in OEM format.
- Increased power distribution and weight capacity.
- Supports both AC and DC configurations.
- The rack height is increased to 1926 mm (75.8 in.) if a power distribution panel is fixed to the top of the rack.
- The #6068 feature provides a cost effective plain front door.

- Weights are as follows:
  - T00 base empty rack: 244 kg (535 lb.)
  - T00 full rack: 816 kg (1795 lb.)
  - Maximum Weight of Drawers is 572 kg (1260 lb.)
  - Maximum Weight of Drawers in a zone 4 earthquake environment is 490 kg (1080 lb.). This equates to 13.6 kg (30 lb.)/EIA.

1.13.2 IBM 7014 model T42 rack

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

- The T42 rack has 42U (EIA units) of usable space (6U of additional space).
- The model T42 supports AC power only.
- Weights are as follows:
  - T42 base empty rack: 261 kg (575 lb.)
  - T42 full rack: 930 kg (2045 lb.)
- The feature #ERG7 provides an attractive black full-height rack door. The door is steel, with a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack.
- The feature #6069 provides a cost-effective plain front door.
- The feature #6249 provides a special acoustic door
- A minimum of one of the following is required:
  - Feature #ER2B allows you to reserve 2U of space at the bottom of the rack.
  - Feature #ER2T allows you to reserve 2U of space at the top of the rack.

Special door: The Power 780 logo rack door (#6250) is not supported.

A rear rack extension of 8 inches or 20.3 cm (#ERG0) provides space to hold cables on the side of the rack and keep the center area clear for cooling and service access. Including this extension is very, very strongly recommended where large numbers of thicker I/O cables are present or may be added in the future. The definition of a large number depends on the type of I/O cables used. Probably around 64 short-length SAS cables per side of a rack or around 50 longer-length (thicker) SAS cables per side of a rack is a good rule of thumb. Generally, other I/O cables are thinner and easier to fit in the sides of the rack and the number of cables can be higher. SAS cables are most commonly found with multiple EXP24S SAS drawers (#5887) driven by multiple PCIe SAS adapters. For this reason, it can be a very good practice to keep multiple EXP24S drawers in the same rack as the PCIe Gen3 I/O drawer or in a separate rack close to the PCIe Gen3 I/O drawer, using shorter, thinner SAS cables. The feature ERG0 extension can be good to use even with a smaller numbers of cables as it enhances the ease of cable management with the extra space it provides.

Recommended Rack: The 7014-T42 System rack with additional rear extension (#ERG0) is recommended for all initial Power E870 and Power E880 system orders. These are the default options for new orders.

1.13.3 IBM 7953 model 94Y rack

The 2.0-meter (79.3 inch) model 94Y rack has the following features:

- 42U (EIA units)
- Weights:
  - Base empty rack: 187 kg (221 lb.)
  - Maximum load limit: 664 kg (1460 lb.)

The IBM 42U Slim Rack (7953-94Y) differs from the IBM 42U enterprise rack (7014-T42) in several aspects. Both provide 42U of vertical space, are 1100 mm deep, and have an interior rail-to-rail depth of 715 mm. However, the IBM 42U Slim Rack is 600 mm wide; the T42 is 645 mm wide with side covers. For clients with 2-foot floor tiles, the extra 45 mm (1.77-inch) width of the enterprise rack can sometimes cause challenges when cutting holes in the floor tiles for cabling.

The 42U Slim Rack has a lockable perforated front steel door, providing ventilation, physical security, and visibility of indicator lights in the installed equipment within. In the rear, either a lockable perforated rear steel door (#EC02) or a lockable rear door heat exchanger (#EC15) is used. Lockable optional side panels (#EC03) increase the rack's aesthetics, help control airflow through the rack, and provide physical security. Multiple 42U Slim Racks can be bolted together to create a rack suite (indicated with feature #EC04).

Weight calculation: Maximum weight limits must include everything that will be installed in the rack. This must include servers, I/O drawers, PDUs, switches, and anything else installed. In zone 4 earthquake environments, the rack should be configured starting with the heavier drawers at the bottom of the rack.

1.13.4 Feature code #0551 rack

The 1.8-meter rack (#0551) is a 36U (EIA units) rack. The rack that is delivered as #0551 is the same rack delivered when you order the 7014-T00 rack. The included features might differ. Several features that are delivered as part of the 7014-T00 must be ordered separately with the #0551. The #0551 is not available for initial orders of Power E870 and E880 servers.

1.13.5 Feature code #0553 rack

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

1.13.6 The AC power distribution unit and rack content

For rack models T00, T42, and the slim 94Y, 12-outlet PDUs are available. The PDUs available include these:

- PDUs Universal UTG0247 Connector (#7188)
- Intelligent PDU+ Universal UTG0247 Connector (#7109)

PDU mounting: Only horizontal PDUs are allowed in racks hosting Power E870 and Power E880 system. Vertically mounted PDUs limit access to the cable routing space on the side of the rack and cannot be used.

When mounting the horizontal PDUs, it is a good practice to place them almost at the top or almost at the bottom of the rack, leaving 2U or more of space at the very top or very bottom of the rack for cable management. Mounting a horizontal PDU in the middle of the rack is generally not optimal for cable management.

For the Power E870 or E880 installed in IBM 7014 or FC 055x racks, the following PDU rules apply:

- For PDU #7109 and #7188 when using 24 Amp power cord #6654, #6655, #6656, #6657, or #6658, each pair of PDUs can power one Power E870 or Power E880 system node and two I/O expansion drawers, or eight I/O expansion drawers. 24A PDU cables are used to supply 30A PDUs. In Figure 1-12 on page 33, you can see the rack configuration with two pairs of 30A PDUs suppling a two system node configuration and two I/O expansion drawers.

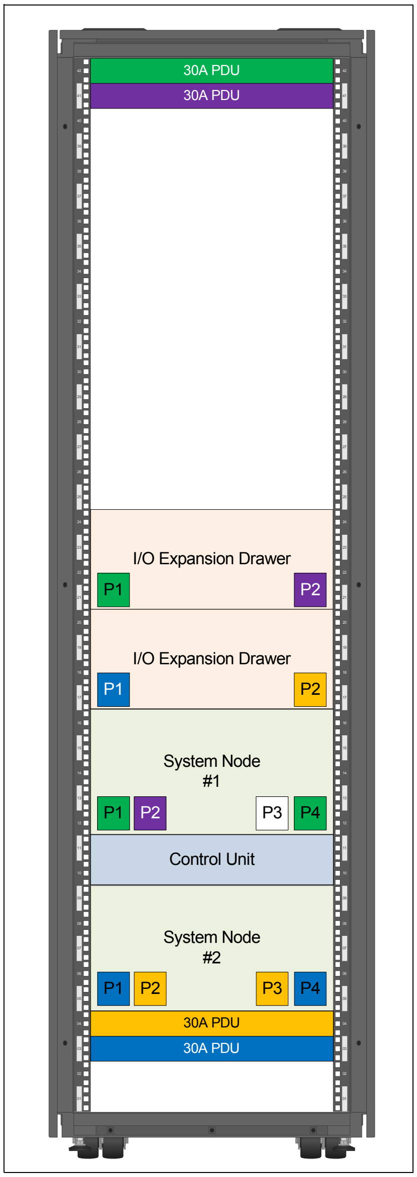


Figure 1-12 Two system node configuration and two I/O expansion drawers supplied by 30A PDUs

- For PDU #7109 and #7188 when using three phase power cords or 48 Amp power cords #6491 or #6492, each pair of PDUs can power up to two Power E870 or Power E880 system nodes and two I/O expansion drawers, or eight I/O expansion drawers. 48A PDU cables are used to supply 60A PDU. In Figure 1-13 on page 34, you can see the rack configuration with two pairs of 60A PDUs suppling a two system node configuration, four I/O expansion drawers, and four EXP24S disk drawers.

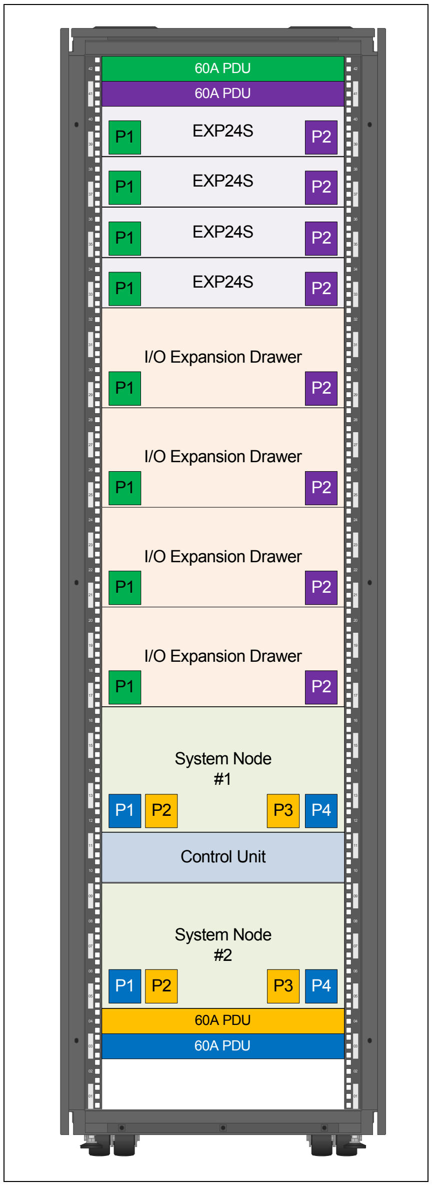


Figure 1-13 Two system nodes configuration with I/O expansions and disk drawers supplied by 60A PDUs

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

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_rpower.htm

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

For rack integrated systems, a minimum quantity of two PDUs (#7109 or #7188) are required.

The PDUs (#7109, #7188) support a wide range of country requirements and electrical power specifications. The PDU receives power through a UTG0247 power line connector. Each PDU requires one PDU-to-wall power cord. Various power cord features are available for countries and applications by selecting a PDU-to-wall power cord, which must be ordered separately.

Each power cord provides the unique design characteristics for the specific power requirements. To match new power requirements and save previous investments, these power cords can be requested with an initial order of the rack or with a later upgrade of the rack features.

The PDUs have 12 client-usable IEC 320-C13 outlets. There are six groups of two outlets fed by six circuit breakers. Each outlet is rated up to 10 Amps, but each group of two outlets is fed from one 20 A circuit breaker.

The Universal PDUs are compatible with previous models.

Power cord and PDU: Based on the power cord that is used, the PDU can supply a range of 4.8 - 21 kVA. The total kilovolt ampere (kVA) of all the drawers that are plugged into the PDU must not exceed the power cord limitation.

Each system node mounted in the rack requires four power cords. For maximum availability, be sure to connect power cords from the same system to two separate PDUs in the rack, and to connect each PDU to independent power sources.

2

Architecture and technical overview

This chapter describes the overall system architecture for the IBM Power System E870 (9119-MME) and IBM Power System E880 (9119-MHE) servers. The bandwidths that are provided throughout the section are theoretical maximums that are used for reference.

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

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

2.1 Logical diagrams

This section contains logical diagrams for the Power E870 and E880.

Figure 2-1 shows the logical system diagram for a single system node of a Power E870 or Power E880.

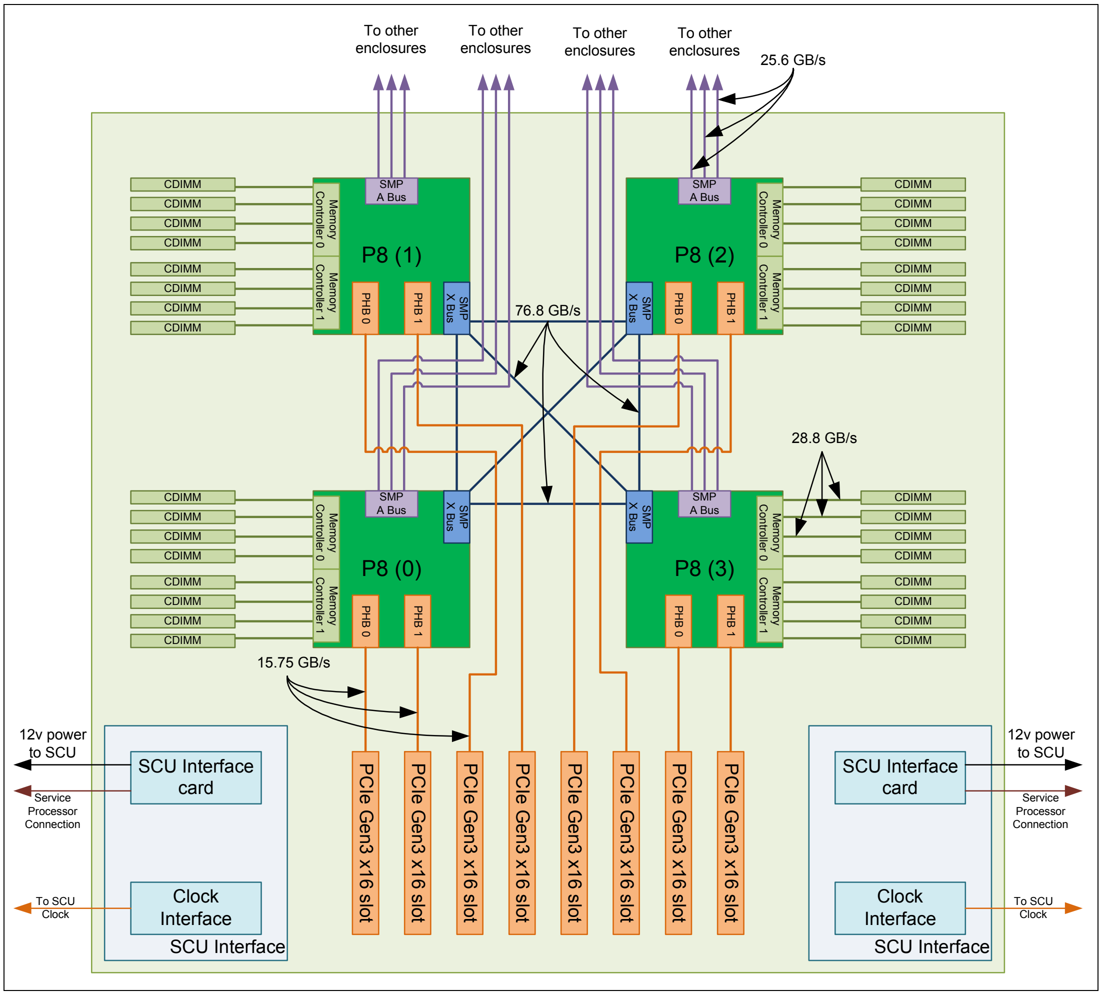


Figure 2-1 Logical system diagram for a system node of a Power E870 or a Power E880

Figure 2-2 shows the logical system diagram for the system control unit of a Power E870 or a Power E880.

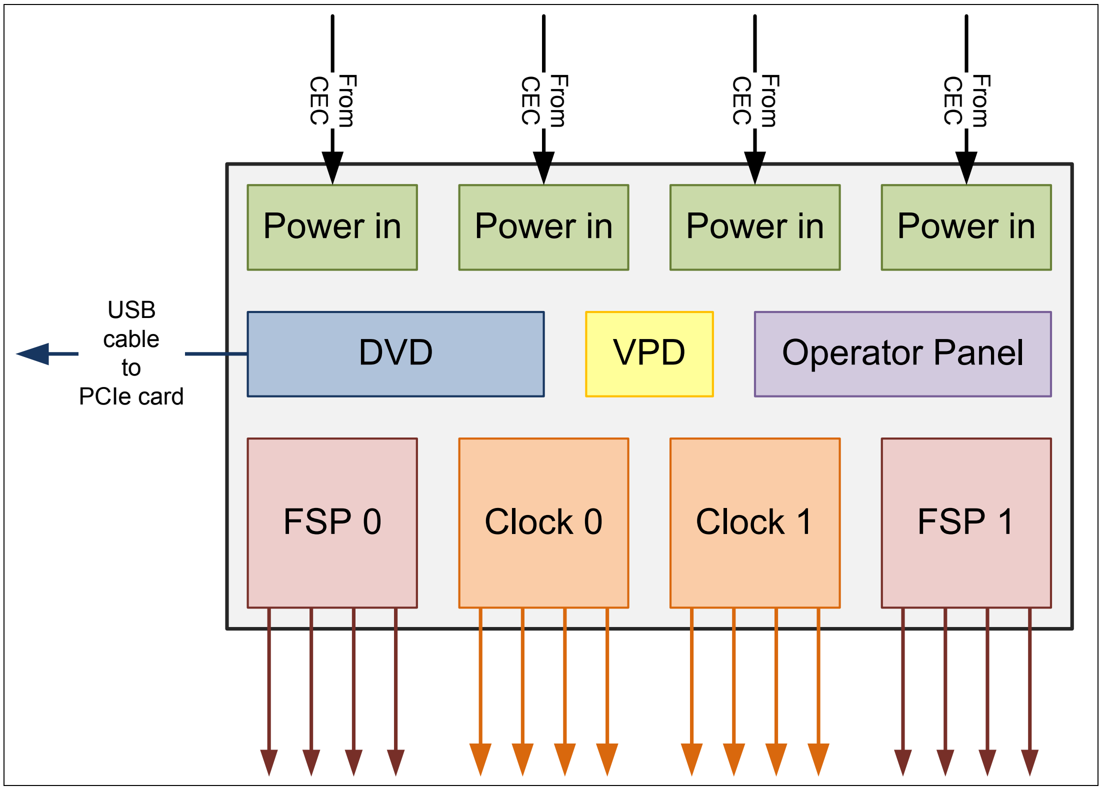


Figure 2-2 Logical system diagram for the system control unit

Flexible symmetric multiprocessing (SMP) cables are used to connect system nodes when a Power E870 or Power E880 is configured with more than one system node. Figure 2-3 shows the SMP connection topology for a two-drawer Power E870 or Power E880 system.

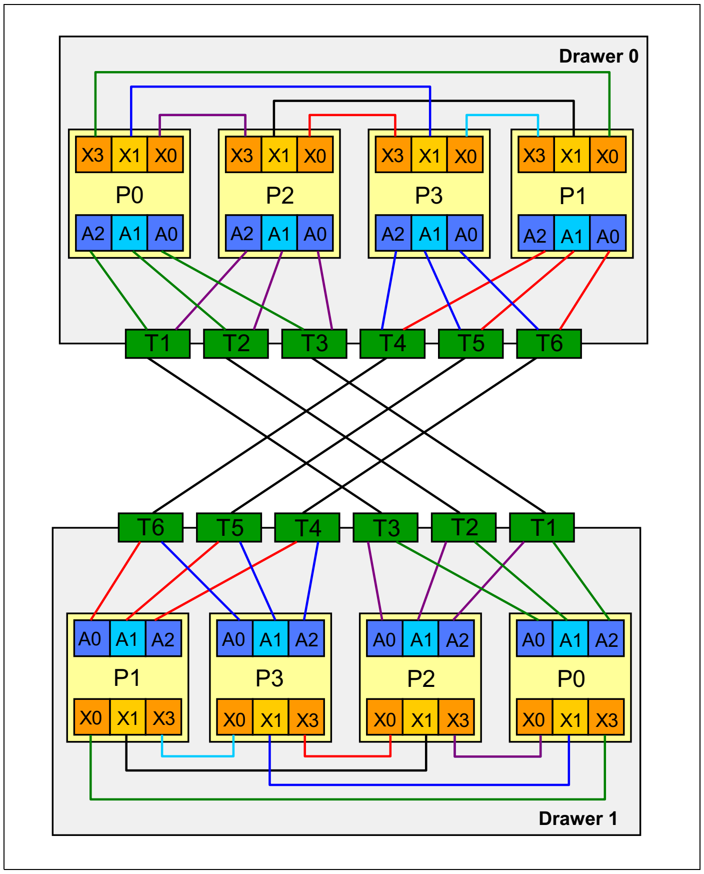


Figure 2-3 SMP connection topology for a two-drawer Power E870 or Power E880

Figure 2-4 shows the SMP connection topology for a three-drawer Power E880.

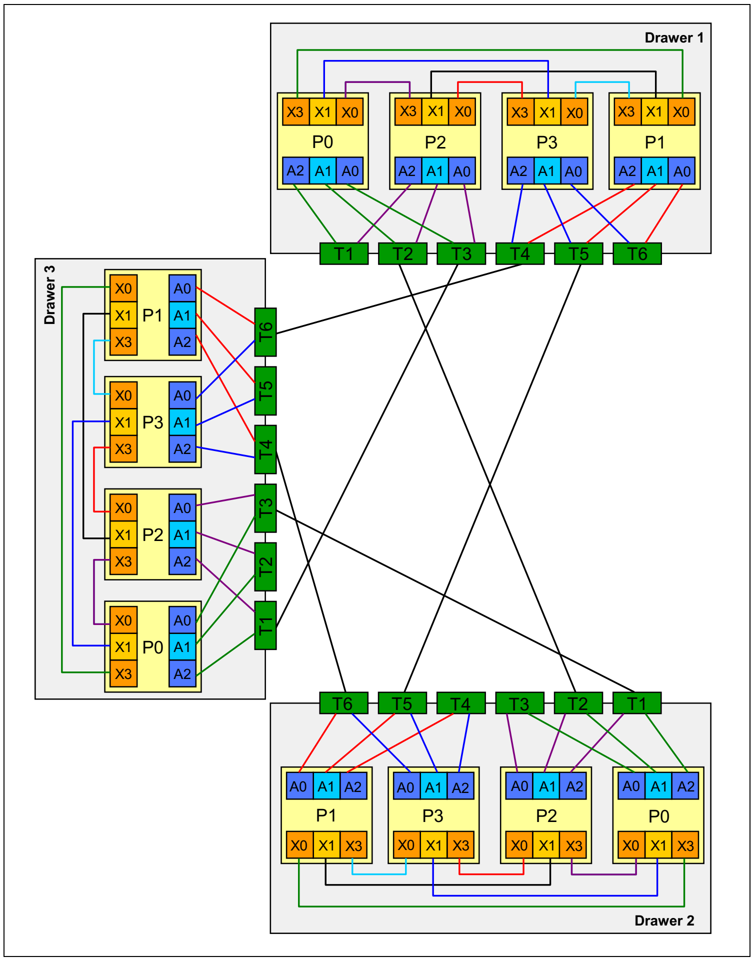


Figure 2-4 SMP connection topology for a three-drawer Power E880

Figure 2-5 shows the SMP connection topology for a four-drawer Power E880.

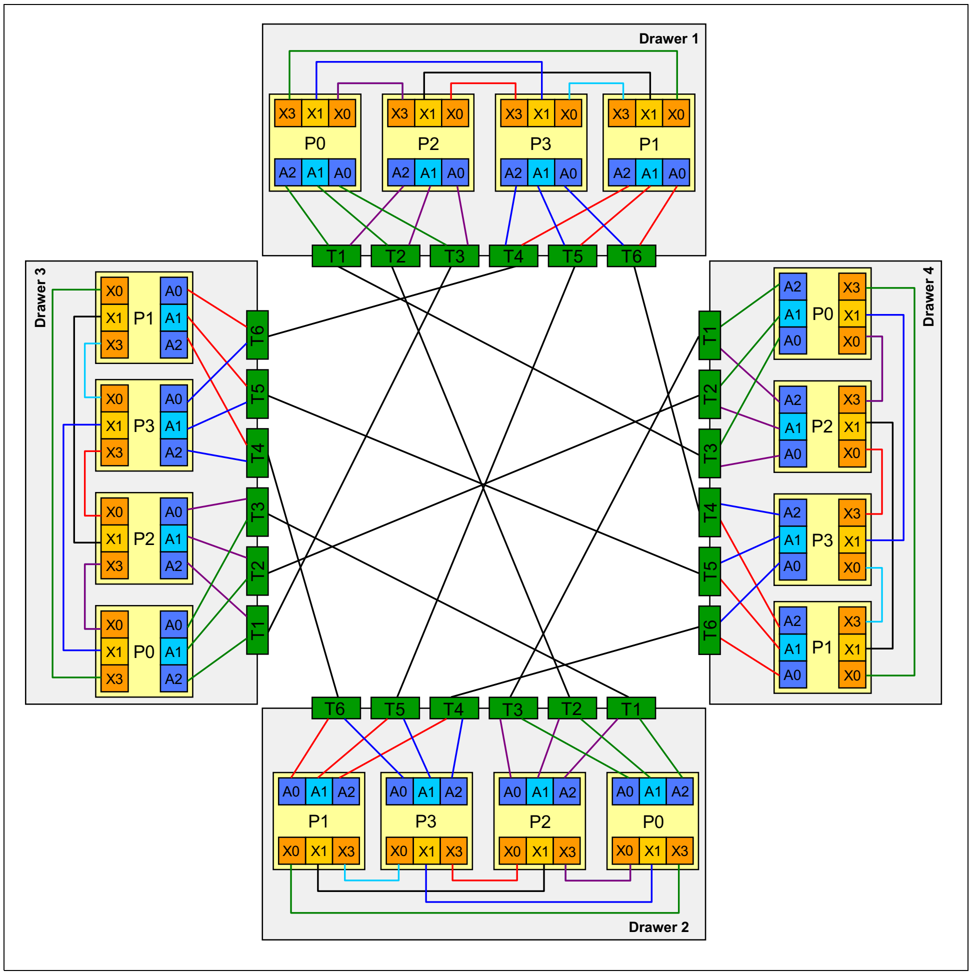


Figure 2-5 SMP connection topology for a four-drawer Power E880

2.2 The IBM POWER8 processor

This section introduces the latest processor in the IBM Power Systems product family, and describes its main characteristics and features.

2.2.1 POWER8 processor overview

The POWER8 processor is manufactured using the IBM 22 nm Silicon-On-Insulator (SOI) technology. Each chip is 649 mm^2 and contains 4.2 billion transistors. As shown in Figure 2-6, the chip contains 12 cores, two memory controllers, PCIe Gen3 I/O controllers, and an interconnection system that connects all components within the chip. On some systems only 6, 8, 10, or 12 cores per processor may be available to the server. Each core has 512 KB of L2 cache, and all cores share 96 MB of L3 embedded DRAM (eDRAM). The interconnect also extends through module and board technology to other POWER8 processors in addition to DDR3 memory and various I/O devices.

POWER8 systems use memory buffer chips to interface between the POWER8 processor and DDR3 or DDR4 memory. Each buffer chip also includes an L4 cache to reduce the latency of local memory accesses.

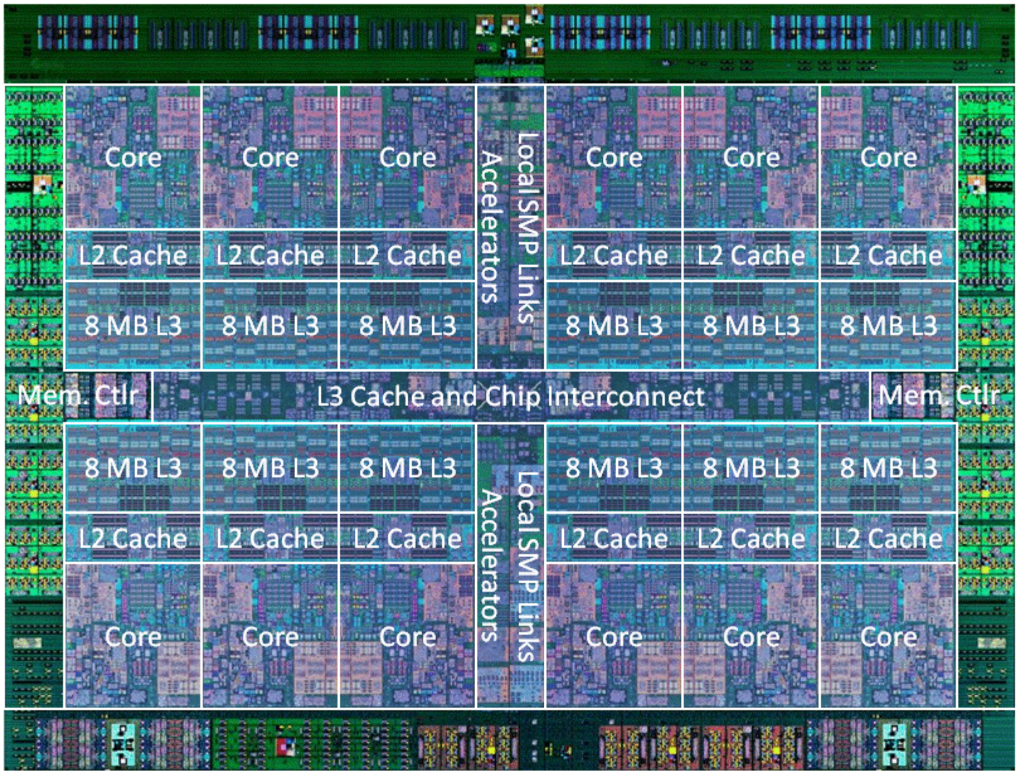


Figure 2-6 The POWER8 processor chip

The POWER8 processor is for system offerings from single-socket servers to multi-socket Enterprise servers. It incorporates a triple-scope broadcast coherence protocol over local and global SMP links to provide superior scaling attributes. Multiple-scope coherence protocols reduce the amount of SMP link bandwidth that is required by attempting operations on a limited scope (single chip or multi-chip group) when possible. If the operation cannot complete coherently, the operation is reissued using a larger scope to complete the operation.

The following additional features can augment the performance of the POWER8 processor:

- Support is provided for DDR3 and DDR4 memory through memory buffer chips that offload the memory support from the POWER8 memory controller.
- Each memory CDIMM has 16 MB of L4 cache within the memory buffer chip that reduces the memory latency for local access to memory behind the buffer chip; the operation of the

L4 cache is not apparent to applications running on the POWER8 processor. Up to 128 MB of L4 cache can be available for each POWER8 processor.

- Hardware transactional memory.
- On-chip accelerators, including on-chip encryption, compression, and random number generation accelerators.
- Coherent Accelerator Processor Interface, which allows accelerators plugged into a PCIe slot to access the processor bus using a low latency, high-speed protocol interface.
- Adaptive power management.

There are two versions of the POWER8 processor chip. Both chips use the same building blocks. The scale-out systems use a 6-core version of POWER8. The 6-core chip is installed in pairs in a Dual Chip Module (DCM) that plugs into a socket in the system board of the systems. Functionally, it works as a single chip module (SCM).

The Enterprise servers use a 12-core chip installed on a Single Chip module (SCM) that plugs into a socket in the system board of the systems. SCMs are used in the Power E870 and Power E880 systems.

Figure 2-7 shows a graphic representation of the 6-core processor on a DCM. DCMs are only available on the scale-out systems. It is shown here for informational purposes.

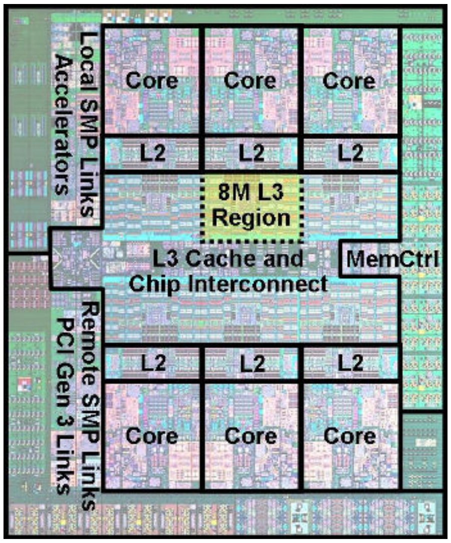


Figure 2-7 6-core POWER8 processor chip on a DCM

Figure 2-8 shows a graphic representation of the 12-core processor on a SCM. The SCM is used in Enterprise systems. It is shown here for informational purposes.

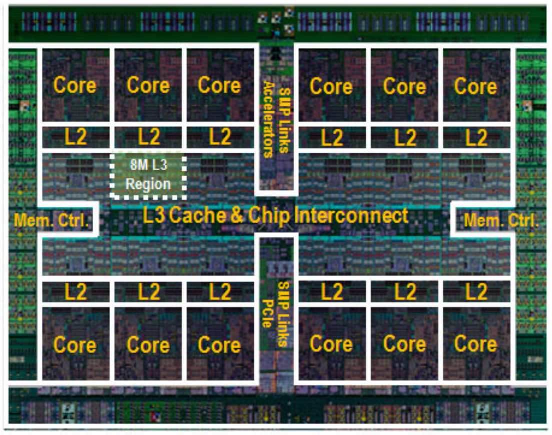


Figure 2-8 12-core POWER8 processor chip on an SCM

Table 2-1 summarizes the technology characteristics of the POWER8 processor.

Table 2-1 Summary of POWER8 processor technology

Technology	**POWER8 processor**
Die size	649 mm^2
Fabrication technology	▶ 22 nm lithography ▶ Copper interconnect ▶ SOI ▶ eDRAM
Maximum processor cores	6 or 12
Maximum execution threads core/chip	8/96
Maximum L2 cache core/chip	512 KB/6 MB
Maximum On-chip L3 cache core/chip	8 MB/96 MB
Maximum L4 cache per chip	128 MB
Maximum memory controllers	2
SMP design-point	16 sockets with IBM POWER8 processors
Compatibility	With prior generations of POWER processor

2.2.2 POWER8 processor core

The POWER8 processor core is a 64-bit implementation of the IBM Power Instruction Set Architecture (ISA) Version 2.07 and has the following features:

- Multi-threaded design, which is capable of up to eight-way simultaneous multithreading (SMT)
- 32 KB, eight-way set-associative L1 instruction cache
- 64 KB, eight-way set-associative L1 data cache

- Enhanced prefetch, with instruction speculation awareness and data prefetch depth awareness
- Enhanced branch prediction, using both local and global prediction tables with a selector table to choose the best predictor
- Improved out-of-order execution
- Two symmetric fixed-point execution units
- Two symmetric load/store units and two load units, all four of which can also run simple fixed-point instructions
- An integrated, multi-pipeline vector-scalar floating point unit for running both scalar and SIMD-type instructions, including the Vector Multimedia eXtension (VMX) instruction set and the improved Vector Scalar eXtension (VSX) instruction set, and capable of up to sixteen floating point operations per cycle (eight double precision or sixteen single precision)
- In-core Advanced Encryption Standard (AES) encryption capability
- Hardware data prefetching with 16 independent data streams and software control
- Hardware decimal floating point (DFP) capability

More information about Power ISA Version 2.07 can be found at the following website:

https://www.power.org/documentation/power-isa-v-2-07b/

Figure 2-9 shows a picture of the POWER8 core, with some of the functional units highlighted.

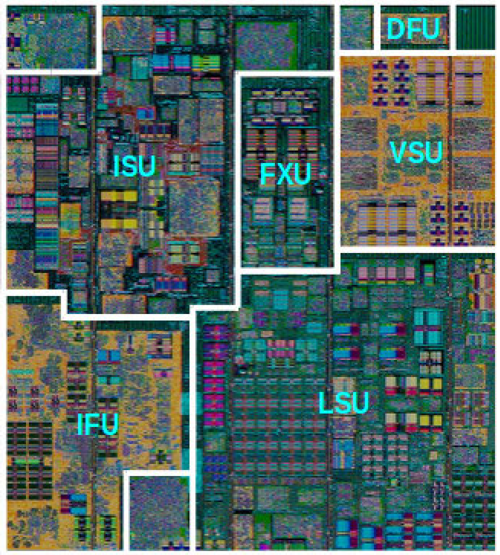


Figure 2-9 POWER8 processor core

2.2.3 Simultaneous multithreading

POWER8 processor advancements in multi-core and multi-thread scaling are remarkable. A significant performance opportunity comes from parallelizing workloads to enable the full potential of the microprocessor, and the large memory bandwidth. Application scaling is influenced by both multi-core and multi-thread technology.

Simultaneous Multithreading (SMT) allows a single physical processor core to simultaneously dispatch instructions from more than one hardware thread context. With SMT, each POWER8 core can present eight hardware threads. Because there are multiple hardware threads per physical processor core, additional instructions can run at the same time. SMT is primarily beneficial in commercial environments where the speed of an individual transaction is not as critical as the total number of transactions that are performed. SMT typically increases the throughput of workloads with large or frequently changing working sets, such as database servers and web servers.

Table 2-2 shows a comparison between the different POWER processors in terms of SMT capabilities that are supported by each processor architecture.

Table 2-2 SMT levels that are supported by POWER processors

Technology	Cores/system	Maximum SMT mode	Maximum hardware threads per system
IBM POWER4	32	Single Thread (ST)	32
IBM POWER5	64	SMT2	128
IBM POWER6	64	SMT2	128
IBM POWER7	256	SMT4	1024
IBM POWER8	192	SMT8	1536

The architecture of the POWER8 processor, with its larger caches, larger cache bandwidth, and faster memory, allows threads to have faster access to memory resources, which translates in to a more efficient usage of threads. Because of that, POWER8 allows more threads per core to run concurrently, increasing the total throughput of the processor and of the system.

2.2.4 Memory access

On the Power E870 and Power E880, each POWER8 processor has two memory controllers, each connected to four memory channels. Each memory channel operates at 1600 MHz and connects to a DIMM. Each DIMM on a POWER8 system has a memory buffer that is responsible for many functions that were previously on the memory controller, such as scheduling logic and energy management. The memory buffer also has 16 MB of level 4 (L4) cache.

On the Power E870, each memory channel can address up to 256 GB. Therefore a single system node can address 8 TB of memory. A two node system can address up to 16 TB of memory.

On the Power E880, each memory channel can address up to 256 GB. Therefore a single system node can address 8 TB of memory. A two node system can address up to 16 TB of memory and a four node system can address up to 32 TB of memory.

Figure 2-10 gives a simple overview of the POWER8 processor memory access structure in the Power E870 and Power E880.

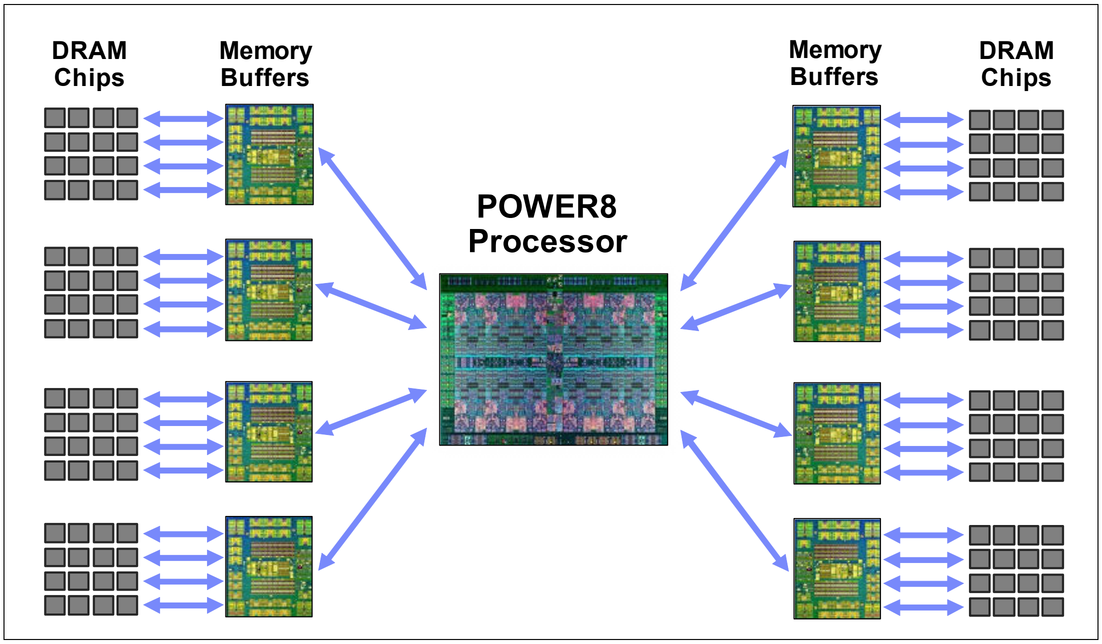


Figure 2-10 Overview of POWER8 memory access structure

2.2.5 On-chip L3 cache innovation and Intelligent Cache

Similar to POWER7 and POWER7+, the POWER8 processor uses a breakthrough in material engineering and microprocessor fabrication to implement the L3 cache in eDRAM and place it on the processor die. L3 cache is critical to a balanced design, as is the ability to provide good signaling between the L3 cache and other elements of the hierarchy, such as the L2 cache or SMP interconnect.

The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a fast 8 MB 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 that is associated with another core, depending on the reference patterns. Data can also be cloned and stored in more than one core's FLR-L3 cache, again depending on the reference patterns. This Intelligent Cache management enables the POWER8 processor to optimize the access to L3 cache lines and minimize overall cache latencies.

Figure 2-6 on page 43 and Figure 2-7 on page 44 show the on-chip L3 cache, and highlight one fast 8 MB L3 region closest to a processor core.

The benefits of using eDRAM on the POWER8 processor die is significant for several reasons:

- Latency improvement

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

- Bandwidth improvement

  A 2x bandwidth improvement occurs with on-chip interconnect. Frequency and bus sizes are increased to and from each core.

- No off-chip driver or receivers

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

- Small physical footprint

  The performance of eDRAM when implemented on-chip is similar to conventional SRAM but requires far less physical space. IBM on-chip eDRAM uses only a third of the components that conventional SRAM uses, 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.

2.2.6 Level 4 cache and memory buffer

POWER8 processor-based systems introduce an additional level of memory hierarchy. The Level 4 (L4) cache is implemented together with the memory buffer in the Custom DIMM (CDIMM). Each memory buffer contains 16 MB of L4 cache. Figure 2-11 shows a picture of the memory buffer, where you can see the 16 MB L4 cache, and processor links and memory interfaces.

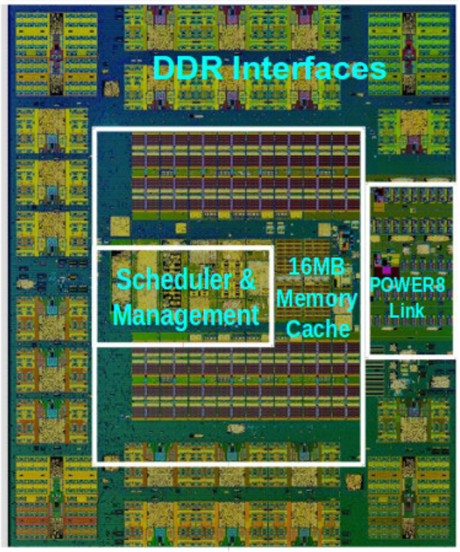


Figure 2-11 Memory buffer chip

Table 2-3 shows a comparison of the different levels of cache in the POWER7, POWER7+, and POWER8 processors.

Table 2-3 POWER8 cache hierarchy

Cache	POWER7	POWER7+	POWER8
L1 instruction cache: Capacity/associativity	32 KB, 4-way	32 KB, 4-way	32 KB, 8-way
L1 data cache: Capacity/associativity bandwidth	32 KB, 8-way Two 16 B reads or one 16 B writes per cycle	32 KB, 8-way Two 16 B reads or one 16 B writes per cycle	64 KB, 8-way Two 16 B reads or one 16 B writes per cycle

Cache	POWER7	POWER7+	POWER8
L2 cache: Capacity/associativity bandwidth	256 KB, 8-way Private 32 B reads and 16 B writes per cycle	256 KB, 8-way Private 32 B reads and 16 B writes per cycle	512 KB, 8-way Private 32 B reads and 16 B writes per cycle
L3 cache: Capacity/associativity bandwidth	On-Chip 4 MB/core, 8-way 16 B reads and 16 B writes per cycle	On-Chip 10 MB/core, 8-way 16 B reads and 16 B writes per cycle	On-Chip 8 MB/core, 8-way 32 B reads and 32 B writes per cycle
L4 cache: Capacity/associativity bandwidth	N/A	N/A	Off-Chip 16 MB/buffer chip, 16-way Up to 8 buffer chips per socket

For more information about the POWER8 memory subsystem, see 2.3, “Memory subsystem” on page 53.

2.2.7 Hardware transactional memory

Transactional memory is an alternative to lock-based synchronization. It attempts to simplify parallel programming by grouping read and write operations and running them like a single operation. Transactional memory is like database transactions where all shared memory accesses and their effects are either committed all together or discarded as a group. All threads can enter the critical region simultaneously. If there are conflicts in accessing the shared memory data, threads try accessing the shared memory data again or are stopped without updating the shared memory data. Therefore, transactional memory is also called a lock-free synchronization. Transactional memory can be a competitive alternative to lock-based synchronization.

Transactional Memory provides a programming model that makes parallel programming easier. A programmer delimits regions of code that access shared data and the hardware runs these regions atomically and in isolation, buffering the results of individual instructions, and retrying execution if isolation is violated. Generally, transactional memory allows programs to use a programming style that is close to coarse-grained locking to achieve performance that is close to fine-grained locking.

Most implementations of transactional memory are based on software. The POWER8 processor-based systems provide a hardware-based implementation of transactional memory that is more efficient than the software implementations and requires no interaction with the processor core, therefore allowing the system to operate at maximum performance.

2.2.8 Coherent Accelerator Processor Interface

The Coherent Accelerator Interface Architecture (CAIA) defines a coherent accelerator interface structure for attaching peripherals to Power Systems.

The Coherent Accelerator Processor Interface (CAPI) can attach accelerators that have coherent shared memory access to the processors in the server and share full virtual address translation with these processors, using a standard PCIe Gen3 bus.

Applications can have customized functions in Field Programmable Gate Arrays (FPGA) and be able to enqueue work requests directly in shared memory queues to the FPGA, and using the same effective addresses (pointers) it uses for any of its threads running on a host processor. From the practical perspective, CAPI allows a specialized hardware accelerator to be seen as an additional processor in the system, with access to the main system memory, and coherent communication with other processors in the system.

The benefits of using CAPI include the ability to access shared memory blocks directly from the accelerator, perform memory transfers directly between the accelerator and processor cache, and reduction in the code path length between the adapter and the processors. This is because the adapter is not operating as a traditional I/O device, and there is no device driver layer to perform processing. It also presents a simpler programming model.

Figure 2-12 on page 51 shows a high-level view of how an accelerator communicates with the POWER8 processor through CAPI. The POWER8 processor provides a Coherent Attached Processor Proxy (CAPP), which is responsible for extending the coherence in the processor communications to an external device. The coherency protocol is tunneled over standard PCIe Gen3, effectively making the accelerator part of the coherency domain.

The accelerator adapter implements the Power Service Layer (PSL), which provides address translation and system memory cache for the accelerator functions. The custom processors on the board, consisting of an FPGA or an Application Specific Integrated Circuit (ASIC) use this layer to access shared memory regions, and cache areas as though they were a processor in the system. This ability greatly enhances the performance of the data access for the device and simplifies the programming effort to use the device. Instead of treating the hardware accelerator as an I/O device, it is treated as a processor. This eliminates the requirement of a device driver to perform communication, and the need for Direct Memory Access that requires system calls to the operating system kernel. By removing these layers, the data transfer operation requires fewer clock cycles in the processor, greatly improving the I/O performance.

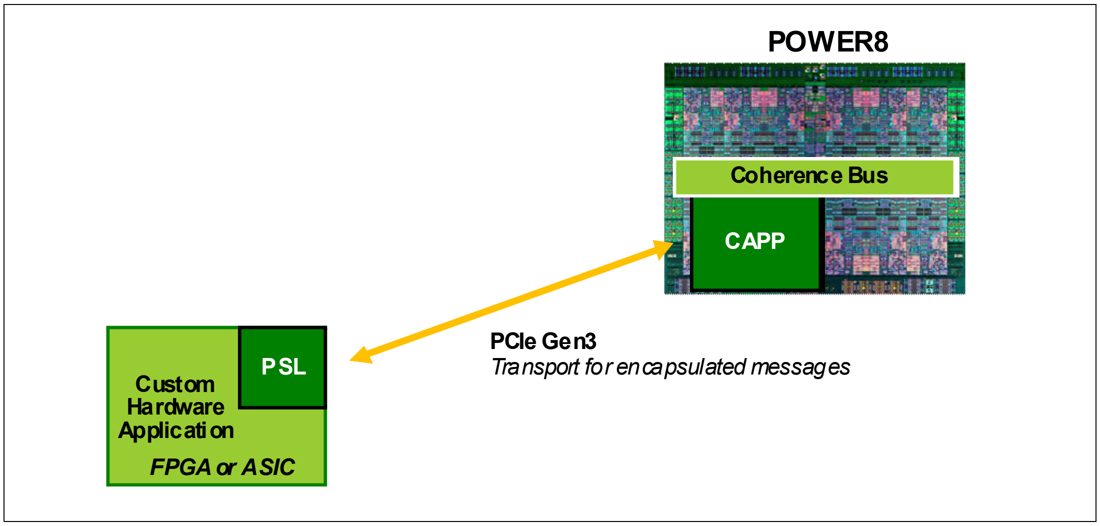

POWER8
Coherence Bus
CAPP
PSL
Custom Hardware Application
FPGA or ASIC
PCIe Gen3
Transport for encapsulated messages

Figure 2-12 CAPI accelerator that is attached to the POWER8 processor

The implementation of CAPI on the POWER8 processor allows hardware companies to develop solutions for specific application demands and use the performance of the POWER8 processor for general applications and custom acceleration of specific functions using a hardware accelerator, with a simplified programming model and efficient communication with the processor and memory resources.

2.2.9 Power management and system performance

The POWER8 processor has power saving and performance enhancing features that can be used to lower overall energy usage, while yielding higher performance when needed. The following modes can be enabled and modified to use these features.

Dynamic Power Saver: Favor Performance

This mode is intended to provide the best performance. If the processor is being used even moderately, the frequency is raised to the maximum frequency possible to provide the best performance. If the processors are lightly used, the frequency is lowered to the minimum frequency, which is potentially far below the nominal shipped frequency, to save energy. The top frequency that is achieved is based on system type and is affected by environmental conditions. Also, when running at the maximum frequency, more energy is being consumed, which means this mode can potentially cause an increase in overall energy consumption.

Dynamic Power Saver: Favor Power

This mode is intended to provide the best performance per watt consumed. The processor frequency is adjusted based on the processor usage to maintain the workload throughput without using more energy than required to do so. At high processor usage levels, the frequency is raised above nominal, as in the Favor Performance mode. Likewise, at low processor usage levels, the frequency is lowered to the minimum frequency. The frequency ranges are the same for the two Dynamic Power Saver modes, but the algorithm that determines which frequency to set is different.

Dynamic Power Saver: Tunable Parameters

The Dynamic Power Saver: Favor Performance and Dynamic Power Saver: Favor Power modes are tuned to provide both energy savings and performance increases. However, there might be situations where only top performance is of concern, or, conversely, where peak power consumption is an issue. The tunable parameters can be used to modify the setting of the processor frequency in these modes to meet these various objectives. Modifying these parameters should be done only by advanced users. If you must address any issues concerning the Tunable Parameters, IBM support personal should be directly involved in the parameter value selection.

Idle Power Saver

This mode is intended to save the maximum amount of energy when the system is nearly idle. When the processors are found to be nearly idle, the frequency of all processors is lowered to the minimum. Additionally, workloads are dispatched onto a smaller number of processor cores so that the other processor cores can be put into a low energy usage state. When processor usage increases, the process is reversed: The processor frequency is raised back up to nominal, and the workloads are spread out once again over all of the processor cores. There is no performance boosting aspect in this mode, but entering or exiting this mode might affect overall performance. The delay times and usage levels for entering and exiting this mode can be adjusted to allow for more or less aggressive energy savings.

The controls for all of these modes are available on the Advanced System Management Interface (ASMI) and are described in more detail in a white paper that is found at the following link:

http://public.dhe.ibm.com/common/ssi/ecm/po/en/pow03125usen/POW03125USEN.PDF

For more information, see 2.13, “Energy management” on page 113.

2.2.10 Comparison of the POWER7, POWER7+, and POWER8 processors

Table 2-4 shows comparable characteristics between the generations of POWER7, POWER7+, and POWER8 processors.

Table 2-4 Comparison of technologies for the POWER8 processor and the prior generations

Characteristics	POWER7	POWER7+	POWER8
Technology	45 nm	32 nm	22 nm
Die size	567 mm^2	567 mm^2	649 mm^2
Number of transistors	1.2 billion	2.1 billion	4.2 billion
Maximum cores	8	8	12
Maximum SMT threads per core	4 threads	4 threads	8 threads
Maximum frequency	4.25 GHz	4.4 GHz	4.35 GHz
L2 Cache	256 KB per core	256 KB per core	512 KB per core
L3 Cache	4 MB or 8 MB of FLR-L3 cache per core with each core having access to the full 32 MB of L3 cache, on-chip eDRAM	10 MB of FLR-L3 cache per core with each core having access to the full 80 MB of L3 cache, on-chip eDRAM	8 MB of FLR-L3 cache per core with each core having access to the full 96 MB of L3 cache, on-chip eDRAM
Memory support	DDR3	DDR3	DDR3 and DDR4
I/O bus	GX++	GX++	PCIe Gen3

2.3 Memory subsystem

The Power E870 can have up to two system nodes per system with each system node having 32 CDIMM slots capable of supporting 16 GB, 32 GB, 64 GB, 128 GB, and 256 GB CDIMMs running at speeds of 1600 MHz. This allows for a maximum system memory of 16 TB for the 256 CDIMM slots in a system comprised of two system nodes.

The Power E880 can have up to four system nodes per system with each system node having 32 CDIMM slots capable of supporting 16 GB, 32 GB, 64 GB, 128 GB, and 256 GB CDIMMs running at speeds of 1600 MHz. This allows for a maximum system memory of 32 TB for the 256 CDIMM slots of a system comprised of four system nodes.

The memory on the systems are Capacity on Demand capable, allowing for the purchase of additional physical memory capacity and dynamically activate it when needed. It is required that at least 50% of the installed memory capacity is active.

The Power E870 and E880 servers support an optional feature called Active Memory Expansion (#EM82). This allows the effective maximum memory capacity to be much larger than the true physical memory. This feature runs innovative compression and decompression of memory content by using a dedicated coprocessor present on each POWER8 processor to provide memory expansion up to 125%, depending on the workload type and its memory usage. As an example, a server with 256 GB of memory physically installed can effectively be expanded over 512 GB of memory. This approach can enhance virtualization and server consolidation by allowing a partition to do more work with the same physical amount of memory or allowing a server to run more partitions and do more work with the same physical amount of memory.

2.3.1 Custom DIMM

Custom DIMMs (CDIMMs) are innovative memory DIMMs that house industry-standard DRAM memory chips and a set of components that allow for higher bandwidth, lower latency communications and increased availability. These components include:

- Memory Scheduler
- Memory Management (RAS Decisions & Energy Management)
- Memory Buffer

By adopting this architecture for the memory DIMMs, several decisions and processes regarding memory optimizations are run internally into the CDIMM. This saves bandwidth and allows for faster processor-to-memory communications. This also allows for a more robust RAS. For more information, see Chapter 4, “Reliability, availability, and serviceability” on page 159.

The CDIMMs exists in two different form factors, a 152 SDRAM design named the Tall CDIMM and an 80 SDRAM design named the Short CDIMM. Each design will be comprised of multiple 4 GB SDRAM devices depending on its total capacity. The CDIMM slots for the Power E870 and Power E880 are tall CDIMMs slots. A filler is added to the short CDIMM allowing it to properly latch into the same physical location of a tall CDIMM and allows for proper airflow and ease of handling. Tall CDIMMs slots allow for larger DIMM sizes and the adoption of future technologies more seamlessly.

A detailed diagram of the CDIMMs available for the Power E870 and Power E880 can be seen in Figure 2-13.

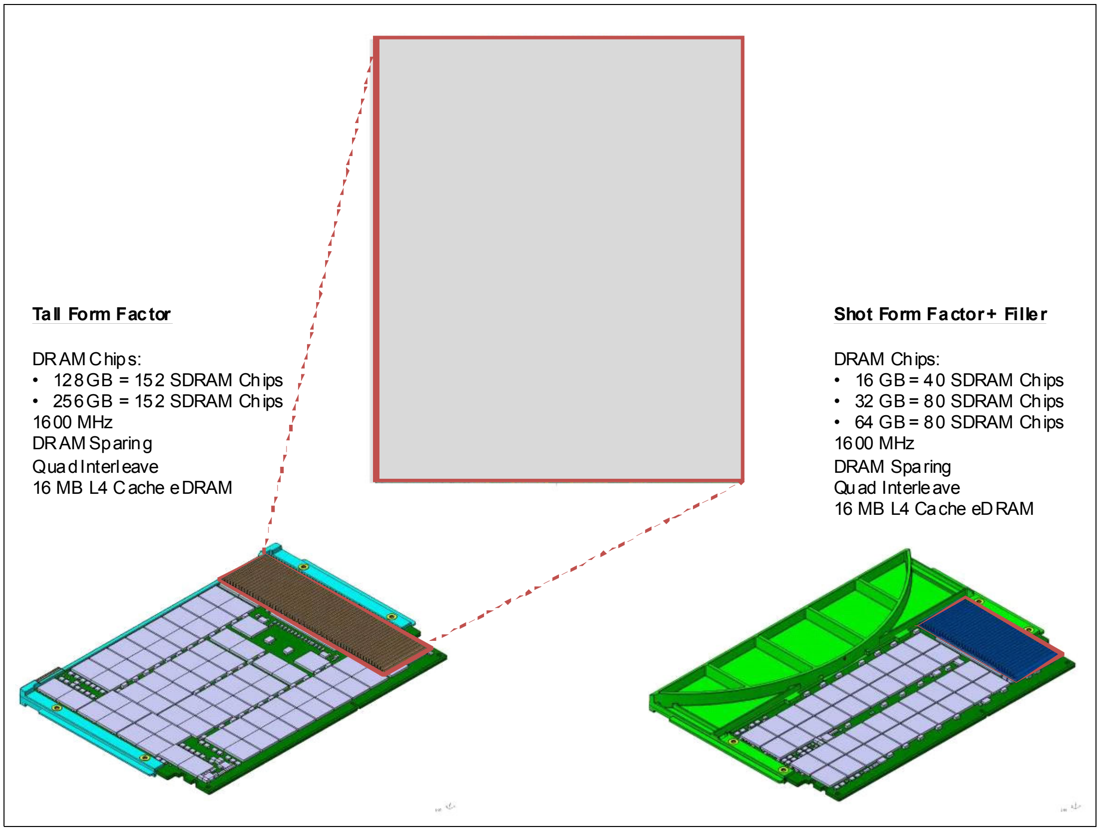


Figure 2-13 Short CDIMM and Tall CDIMM details

The Memory Buffer is a L4 cache and is built on eDRAM technology (same as the L3 cache), which has a lower latency than regular SRAM. Each CDIMM has 16 MB of L4 cache and a fully populated Power E870 server has 1 GB of L4 Cache while a fully populated Power E880 has 2 GB of L4 Cache. The L4 Cache performs several functions that have direct impact on performance and bring a series of benefits for the Power E870 and Power E880:

- Reduces energy consumption by reducing the number of memory requests.
- Increases memory write performance by acting as a cache and by grouping several random writes into larger transactions.
- Partial write operations that target the same cache block are gathered within the L4 cache before being written to memory, becoming a single write operation.
- Reduces latency on memory access. Memory access for cached blocks has up to 55% lower latency than non-cached blocks.

2.3.2 Memory placement rules

For the Power E870 and Power E880, each memory feature code provides four CDIMMs. Therefore a maximum of 8 memory feature codes per system node are allowed in order to fill all the 32 CDIMM slots.

All the memory CDIMMs are capable of capacity upgrade on demand and must have a minimum of 50% of their physical capacity activated. For example, the minimum installed memory for both servers is 256 GB, which requires a minimum of 128 GB active.

For the Power E870 and Power E880, the following memory options are orderable:

- 64 GB (4 X 16 GB) CDIMMs, 1600 MHz DDR3 DRAM (#EM8J)
- 128 GB (4 X 32 GB) CDIMMs, 1600 MHz DDR3 DRAM (#EM8K)
- 256 GB (4 X 64 GB) CDIMMs, 1600 MHz DDR3 DRAM (#EM8L)
- 512 GB (4 X 128 GB) CDIMMs, 1600 MHz DDR3 DRAM (#EM8M)
- 1024 GB (4 X 256 GB) CDIMMs, 1600 MHz DDR4 DRAM (#EM8Y)
- 64 GB (4 X 16 GB) CDIMMs, 1600 MHz DDR4 DRAM (#EM8U)
- 128 GB (4 X 32 GB) CDIMMs, 1600 MHz DDR4 DRAM (#EM8V)
- 256 GB (4 X 64 GB) CDIMMs, 1600 MHz DDR4 DRAM (#EM8W)
- 512 GB (4 X 128 GB) CDIMMs, 1600 MHz DDR4 DRAM (#EM8X)

Each processor has two memory controllers. These memory controllers must have at least a pair of CDIMMs attached to it. This set of mandatory four CDIMMs is called a memory quad. A logical diagram of a POWER8 processor with its two memory quads can be seen in Figure 2-14.

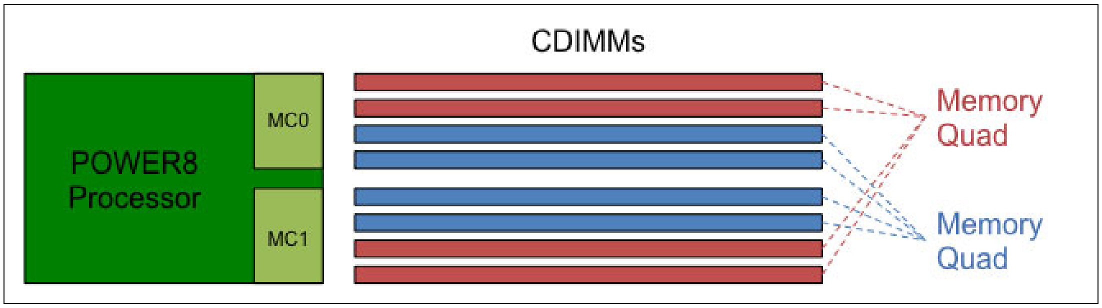


Figure 2-14 Logical diagram of a POWER8 processor and its two memory quads

The basic rules for memory placement follow:

- Each feature code equals a set of four physical CDIMMs; a memory quad.
- Each installed processor must have at least one memory quad populated, which equals to at least one feature code per installed processor.
- A given processor can only have four or eight CDIMMs attached to it.
- All the CDIMMs connected to the same POWER8 processor must be identical. However, it is permitted to mix different CDIMM sizes between different POWER8 processors on a system.
- At least 50% of the installed memory must be activated via memory activation features.

Note: The DDR4 #EM8Y CDIMM has the following usage guidelines:

- No mixing #EM8Y with another size CDIMM on the same system node
- At the time of writing, no mixing #EM8Y with another size CDIMM on the same server
- 100% of the memory slots must be filled.
- Firmware 840.12 is the required minimum level

Development may change these usage guidelines at any time.

The suggested approach is to install memory evenly across all processors in the system and across all system nodes in a system. Balancing memory across the installed processors allows memory access in a consistent manner and typically results in the best possible performance for your configuration. You should account for any plans for future memory upgrades when you decide which memory feature size to use at the time of the initial system order.

A physical diagram with the location codes of the memory CDIMMs of a system node, as well as their grouping as memory quads, can be seen in Figure 2-15.

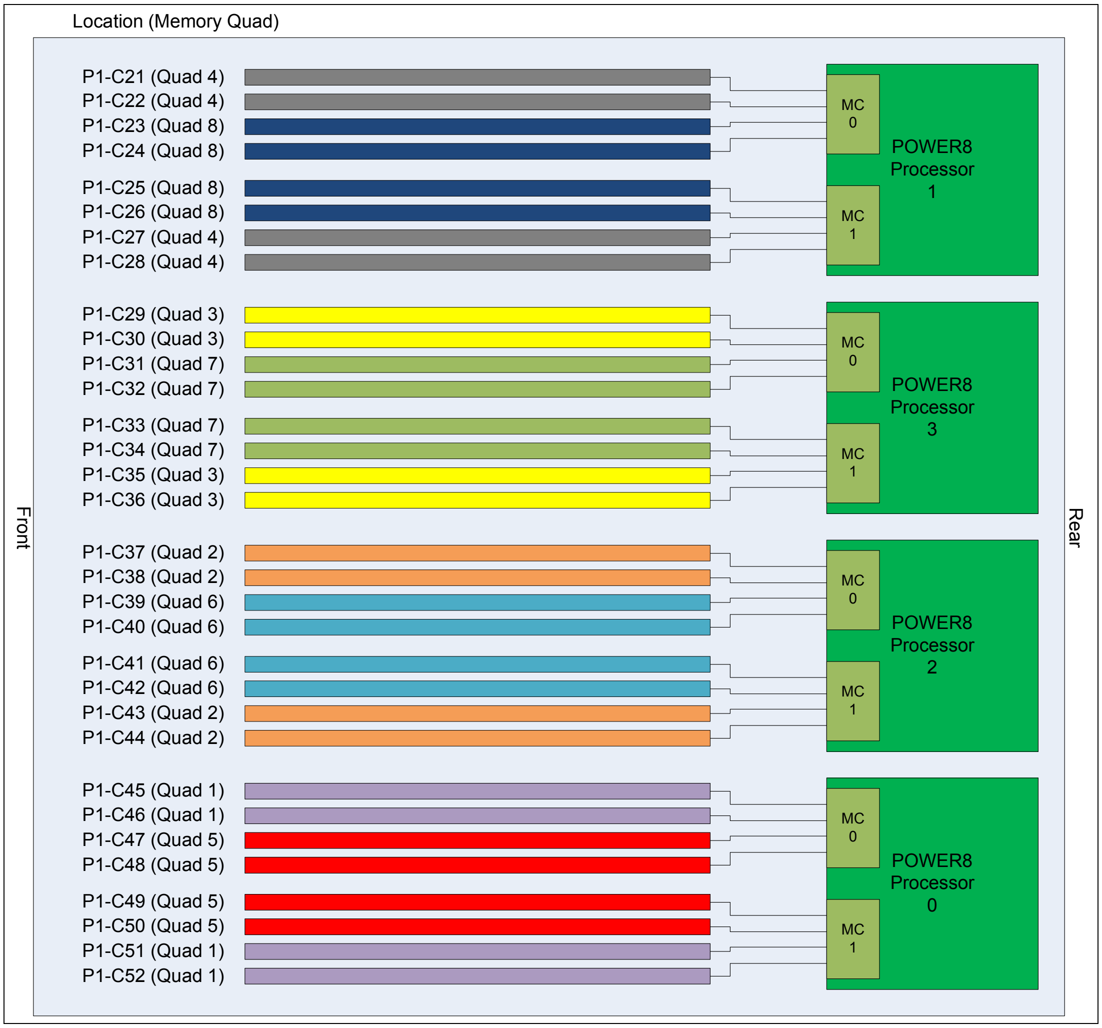


Figure 2-15 System node physical diagram with location codes for CDIMMs

Each system node has eight memory quads identified by the different colors in Figure 2-15. These are the location codes for the slots on each memory quad:

- Quad 1: P1-C45, P1-C46, P1-C51, and P1-C52
- Quad 2: P1-C37, P1-C38, P1-C43, and P1-C44
- Quad 3: P1-C29, P1-C30, P1-C35, and P1-C36
- Quad 4: P1-C21, P1-C22, P1-C27, and P1-C28
- Quad 5: P1-C47, P1-C48, P1-C49, and P1-C50
- Quad 6: P1-C39, P1-C40, P1-C41, and P1-C42
- Quad 7: P1-C31, P1-C32, P1-C33, and P1-C34
- Quad 8: P1-C23, P1-C24, P1-C25, and P1-C26

Table 2-5 shows the CDIMM plugging order for a Power E870 or Power E880 with a single system node.

Table 2-5 Optimal CDIMM memory quad placement for a single system node

System Node 1							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
1	5	2	6	3	7	4	8
Notes: Memory quads 1-4 must be populated. Memory quads on the same processor must be populated with CDIMMs of the same capacity.							

Table 2-6 shows the CDIMM plugging order for a Power E870 or Power E880 with two system nodes.

Table 2-6 Optimal CDIMM memory quad placement for two system nodes

System Node 1							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
1	9	2	11	3	13	4	15
System Node 2							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
5	10	6	12	7	14	8	16
Notes: Memory quads 1-8 must be populated. Memory quads on the same processor must be populated with CDIMMs of the same capacity.							

Table 2-7 shows the CDIMM plugging order for a Power E880 with a three system nodes.

Table 2-7 Optimal CDIMM memory quad placement for three system nodes

System Node 1							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
1	13	2	16	3	19	4	22
System Node 2							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
5	14	6	17	7	20	8	23
System Node 3							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
9	15	10	18	11	21	12	24
Notes: Memory quads 1-12 must be populated. Memory quads on the same processor must be populated with CDIMMs of the same capacity.							

Table 2-8 shows the CDIMM plugging order for a Power E880 with a four system nodes.

Table 2-8 Optimal CDIMM memory quad placement for four system nodes

System Node 1							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
1	17	2	21	3	25	4	29
System Node 2							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
5	18	6	22	7	26	8	30
System Node 3							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
9	19	10	23	11	27	12	31
System Node 4							
Processor 0		**Processor 2**		**Processor 3**		**Processor 1**	
Quad 1	Quad 5	Quad 2	Quad 6	Quad 3	Quad 7	Quad 4	Quad 8
13	20	14	24	15	28	16	32
Notes: Memory quads 1-16 must be populated. Memory quads on the same processor must be populated with CDIMMs of the same capacity.							

2.3.3 Memory activation

All the memory CDIMMs are capable of capacity upgrade on demand and must have a minimum of 50% of their physical capacity activated. For example, the minimum physical installed memory for both Power E870 and Power E880 is 256 GB, which requires a minimum of 128 GB activated.

There are two activation types that can be used to accomplish this:

- Static memory activations: Memory activations that are exclusive for a single server
- Mobile memory activations: Memory activations that can be moved from server to server in a power enterprise pool.

Both types of memory activations can co-reside in the same system, as long as at least 25% of the memory activations are static. This leads to a maximum of 75% of the memory activations as mobile.

In Figure 2-16, there is an example of the minimum required activations for a system with 1 TB of installed memory.

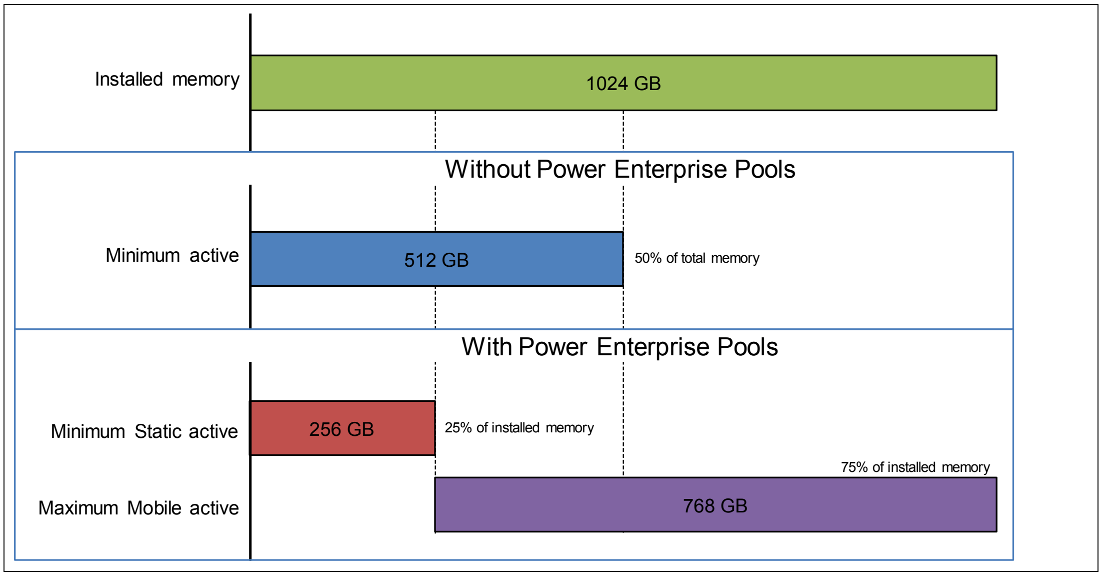


Figure 2-16 Example of the minimum required activations for a system with 1 TB of installed memory

The granularity for static memory activation is 1 GB, while for mobile memory activation, the granularity is 100 GB. In Table 2-9, there is a list of the feature codes that can be used to achieve the desired number of activations:

Table 2-9 Static and mobile memory activation feature codes

Feature code	Feature description	Amount of memory	Type of activation
EMA5	1 GB Memory activation	1 GB	Static
EMA6	100 GB Memory activation	100 GB	Static
EMA7	100 GB Mobile memory activation	100 GB	Mobile

Static memory activations can be converted to mobile memory activations after system installation. In order to enable mobile memory activations, the systems must be part of a power enterprise pool and have feature code #EB35 configured. For more information about power enterprise pools, see 2.4.2, “Power enterprise pools and mobile capacity on demand (Mobile CoD)” on page 68.

2.3.4 Memory throughput

The peak memory and I/O bandwidths per system node have increased over 300% compared to previous POWER7 servers, providing the next generation of data intensive applications with a platform capable of handling the needed amount of data.

DDR4 256 GB CDIMMs have similar performance as 128 GB CDIMMs. They both operate 1600 MHz and have the same memory bandwidth considerations as with the 512 GB memory feature.

Power E870

Table 2-10 on page 62 shows the maximum bandwidth estimates for a single core on the Power E870 system.

Table 2-10 Power E870 single core bandwidth estimates

Single core	Power E870	Power E870
	1 core @ 4.024 GHz	**1 core @ 4.190 GHz**
L1 (data) cache	193.15 GBps	201.12 GBps
L2 cache	193.15 GBps	201.12 GBps
L3 cache	257.54 GBps	268.16 GBps

The bandwidth figures for the caches are calculated as follows:

- L1 cache: In one clock cycle, two 16-byte load operations and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 4.024 GHz Core: (2 * 16 B + 1 * 16 B) * 4.024 GHz = 193.15 GBps
  - 4.190 GHz Core: (2 * 16 B + 1 * 16 B) * 4.190 GHz = 201.12 GBps
- L2 cache: In one clock cycle, one 32-byte load operation and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 4.024 GHz Core: (1 * 32 B + 1 * 16 B) * 4.024 GHz = 193.15 GBps
  - 4.190 GHz Core: (1 * 32 B + 1 * 16 B) * 4.190 GHz = 201.12 GBps
- L3 cache: In one clock cycle, one 32-byte load operation and one 32-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 4.024 GHz Core: (1 * 32 B + 1 * 32 B) * 4.024 GHz = 257.54 GBps
  - 4.190 GHz Core: (1 * 32 B + 1 * 32 B) * 4.190 GHz = 286.16 GBps

For each system node of a Power E870 populated with four processors and all its memory CDIMMs filled, the overall bandwidths are shown in Table 2-11.

Table 2-11 Power E870 system node bandwidth estimates

System node bandwidths	Power E870	Power E870
	32 cores @ 4.024 GHz	**40 cores @ 4.190 GHz**
L1 (data) cache	6,181 GBps	8,045 GBps
L2 cache	6,181 GBps	8,045 GBps
L3 cache	8,241 GBps	10,726 GBps
Total Memory	922 GBps	922 GBps
PCIe Interconnect	252.064 GBps	252.064 GBps
Intra-node buses (two system nodes)	922 GBps	922 GBps

PCIe Interconnect: Each POWER8 processor has 32 PCIe lanes running at 7.877Gbps full-duplex. The bandwidth formula is calculated as follows:

32 lanes * 4 processors * 7.877 Gbps * 2 = 252.064 GBps

Rounding: The bandwidths listed here may appear slightly differently in other materials due to rounding of some figures.

For the entire Power E870 system populated with two system nodes, the overall bandwidths are shown in Table 2-12.

Table 2-12 Power E870 total bandwidth estimates

Total bandwidths	Power E870	Power E870
	64 cores @ 4.024 GHz	**80 cores @ 4.190 GHz**
L1 (data) cache	12,362 GBps	16,090 GBps
L2 cache	12,362 GBps	16,090 GBps
L3 cache	16,484 GBps	21,453 GBps
Total Memory	1,844 GBps	1,844 GBps
PCIe Interconnect	504.128 GBps	504.128 GBps
Inter-node buses (two system nodes)	307 GBps	307 GBps
Intra-node buses (two system nodes)	1,844 GBps	1,844 GBps

Power E880

Table 2-13 shows the maximum bandwidth estimates for a single core on the Power E880 system.

Table 2-13 Power E880 single core bandwidth estimates

Single core	Power E880	Power E880
	1 core @ 4.024 GHz	**1 core @ 4.350 GHz**
L1 (data) cache	193.15 GBps	208.80 GBps
L2 cache	193.15 GBps	208.80 GBps
L3 cache	257.54 GBps	278.40 GBps

The bandwidth figures for the caches are calculated as follows:

- L1 cache: In one clock cycle, two 16-byte load operations and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 4.024 GHz Core: (2 * 16 B + 1 * 16 B) * 4.024 GHz = 193.15 GBps
  - 4.350 GHz Core: (2 * 16 B + 1 * 16 B) * 4.350 GHz = 208.80 GBps
- L2 cache: In one clock cycle, one 32-byte load operation and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 4.024 GHz Core: (1 * 32 B + 1 * 16 B) * 4.024 GHz = 193.15 GBps
  - 4.350 GHz Core: (1 * 32 B + 1 * 16 B) * 4.350 GHz = 208.80 GBps
- L3 cache: In one clock cycle, one 32-byte load operation and one 32-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 4.024 GHz Core: (1 * 32 B + 1 * 32 B) * 4.024 GHz = 257.54 GBps
  - 4.350 GHz Core: (1 * 32 B + 1 * 32 B) * 4.350 GHz = 278.40 GBps

For each system node of a Power E880 populated with four processors and all its memory CDIMMs filled, the overall bandwidths are shown in Table 2-14.

Table 2-14 Power E880 system node bandwidth estimates

System node bandwidths	Power E880	Power E880
	48 cores @ 4.024 GHz	**32 cores @ 4.350 GHz**
L1 (data) cache	9,271 GBps	6,682 GBps
L2 cache	9,271 GBps	6,682 GBps
L3 cache	12,362 GBps	8,909 GBps
Total Memory	922 GBps	922 GBps
PCIe Interconnect	252.064 GBps	252.064 GBps
Intra-node buses (two system nodes)	922 GBps	922 GBps

PCIe Interconnect: Each POWER8 processor has 32 PCIe lanes running at 7.877 Gbps full-duplex. The bandwidth formula is calculated as follows:

32 lanes * 4 processors * 7.877Gbps * 2 = 252.064 GBps

Rounding: The bandwidths listed here may appear slightly differently in other materials due to rounding of some figures.

For the entire Power E880 system populated with four system nodes, the overall bandwidths are shown in Table 2-15.

Table 2-15 Power E880 total bandwidth estimates

Total bandwidths	Power E880	Power E880
	192 cores @ 4.024 GHz	**128 cores @ 4.350 GHz**
L1 (data) cache	37,084 GBps	26,726 GBps
L2 cache	37,084 GBps	26,726 GBps
L3 cache	49,448 GBps	35,635 GBps
Total Memory	3,688 GBps	3,688 GBps
PCIe Interconnect	1008.256 GBps	1008.256 GBps
Inter-node buses (four system nodes)	307 GBps	307 GBps
Intra-node buses (four system nodes)	3,688 GBps	3,688 GBps

2.3.5 Active Memory Mirroring

The Power E870 and Power E880 systems have the ability to provide mirroring of the hypervisor code across multiple memory CDIMMs. If a CDIMM that contains the hypervisor code develops an uncorrectable error, its mirrored partner will enable the system to continue to operate uninterrupted.

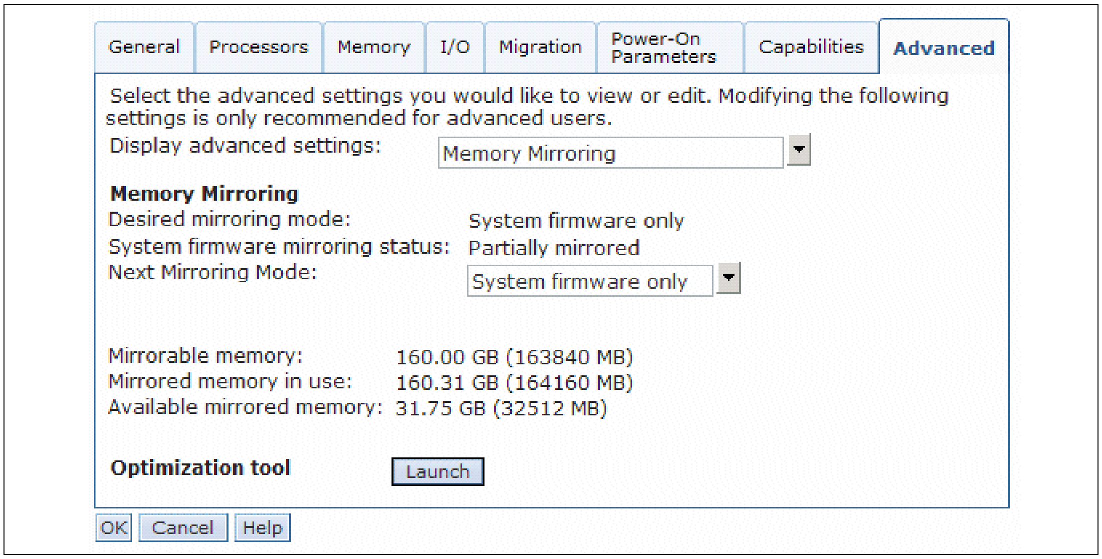
Active Memory Mirroring (AMM) is included with all Power E870 and Power E880 systems at no additional charge. It can be enabled, disabled, or re-enabled depending on the user's requirements.

The hypervisor code logical memory blocks will be mirrored on distinct CDIMMs to allow for more usable memory. There is no specific CDIMM that hosts the hypervisor memory blocks so the mirroring is done at the logical memory block level, not at the CDIMM level. To enable the AMM feature it is mandatory that the server has enough free memory to accommodate the mirrored memory blocks.

Besides the hypervisor code itself, other components that are vital to the server operation are also mirrored:

- Hardware page tables (HPTs), responsible for tracking the state of the memory pages assigned to partitions
- Translation control entities (TCEs), responsible for providing I/O buffers for the partition's communications
- Memory used by the hypervisor to maintain partition configuration, I/O states, virtual I/O information, and partition state

It is possible to check whether the Active Memory Mirroring option is enabled and change its current status through HMC, under the Advanced Tab on the CEC Properties panel (Figure 2-17).



Select the advanced settings you would like to view or edit. Modifying the following settings is only recommended for advanced users.

Display advanced settings: Memory Mirroring

Memory Mirroring

Desired mirroring mode: System firmware only

System firmware mirroring status: Partially mirrored

Next Mirroring Mode: System firmware only

Mirrorable memory: 160.00 GB (163840 MB)

Mirrored memory in use: 160.31 GB (164160 MB)

Available mirrored memory: 31.75 GB (32512 MB)

Optimization tool

Figure 2-17 CEC Properties panel on an HMC

After a failure on one of the CDIMMs containing hypervisor data occurs, all the server operations remain active and flexible service processor (FSP) will isolate the failing CDIMMs. Systems stay in the partially mirrored state until the failing CDIMM is replaced.

There are components that are not mirrored because they are not vital to the regular server operations and require a larger amount of memory to accommodate its data:

- Advanced Memory Sharing Pool
- Memory used to hold the contents of platform dumps

Partition data: Active Memory Mirroring will *not* mirror partition data. It was designed to mirror only the hypervisor code and its components, allowing this data to be protected against a DIMM failure

With AMM, uncorrectable errors in data that are owned by a partition or application are handled by the existing Special Uncorrectable Error handling methods in the hardware, firmware, and operating system.

2.3.6 Memory Error Correction and Recovery

The memory has error detection and correction circuitry is designed such that the failure of any one specific memory module within an ECC word can be corrected without any other fault.

In addition, a spare DRAM per rank on each memory port provides for dynamic DRAM device replacement during runtime operation. Also, dynamic lane sparing on the DMI link allows for repair of a faulty data lane.

Other memory protection features include retry capabilities for certain faults detected at both the memory controller and the memory buffer.

Memory is also periodically scrubbed to allow for soft errors to be corrected and for solid single-cell errors reported to the hypervisor, which supports operating system deallocation of a page associated with a hard single-cell fault.

For more details on Memory RAS, see 4.3.10, “Memory protection” on page 168.

2.3.7 Special Uncorrectable Error handling

Special Uncorrectable Error (SUE) handling prevents 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. If the error is irrelevant, it does not force a checkstop. If the data is used, termination can be limited to the program/kernel or hypervisor owning the data, or freeze of the I/O adapters controlled by an I/O hub controller if data is to be transferred to an I/O device.

2.4 Capacity on Demand

Several types of Capacity on Demand (CoD) offerings are optionally available on the Power 870 and Power E880 servers to help meet changing resource requirements in an on-demand environment, by using resources that are installed on the system but that are not activated.

2.4.1 Capacity Upgrade on Demand (CUoD)

Power E870 and Power E880 systems include a number of active processor cores and memory units. They can also include inactive processor cores and memory units. Active processor cores or memory units are processor cores or memory units that are already available for use on your server when it comes from the manufacturer. Inactive processor cores or memory units are processor cores or memory units that are included with your server, but not available for use until you activate them. Inactive processor cores and memory units can be permanently activated by purchasing an activation feature called Capacity Upgrade on Demand (CUoD) and entering the provided activation code on your server.

With the CUoD offering, you can purchase additional static processor or memory capacity and dynamically activate them when needed, without requiring you to restart your server or interrupt your business. All the static processor or memory activations are restricted to a single server.

Capacity Upgrade on Demand can have several applications to allow for a more flexible environment. One of its benefits is to allow for a given company to reduce the initial investment on a system. Traditional projects using other technologies require that the a system is acquired with all the resources available to support the whole lifecycle of the project. This might incur in costs that would only be necessary on later stages of the project, usually with impacts on software licensing costs and software maintenance.

By using Capacity Upgrade on Demand a company could start with a system with enough installed resources to support the whole project lifecycle but only with enough active resources necessary for the initial project phases. More resources could be added along with the project, adjusting the hardware platform with the project needs. This would allow for a company to reduce the initial investment in hardware and only acquire software licenses that are needed on each project phase, reducing the Total Cost of Ownership and Total Cost of Acquisition of the solution. Figure 2-18 shows a comparison between two scenarios: a fully activated system versus a system with CUoD resources being activated along with the project timeline.

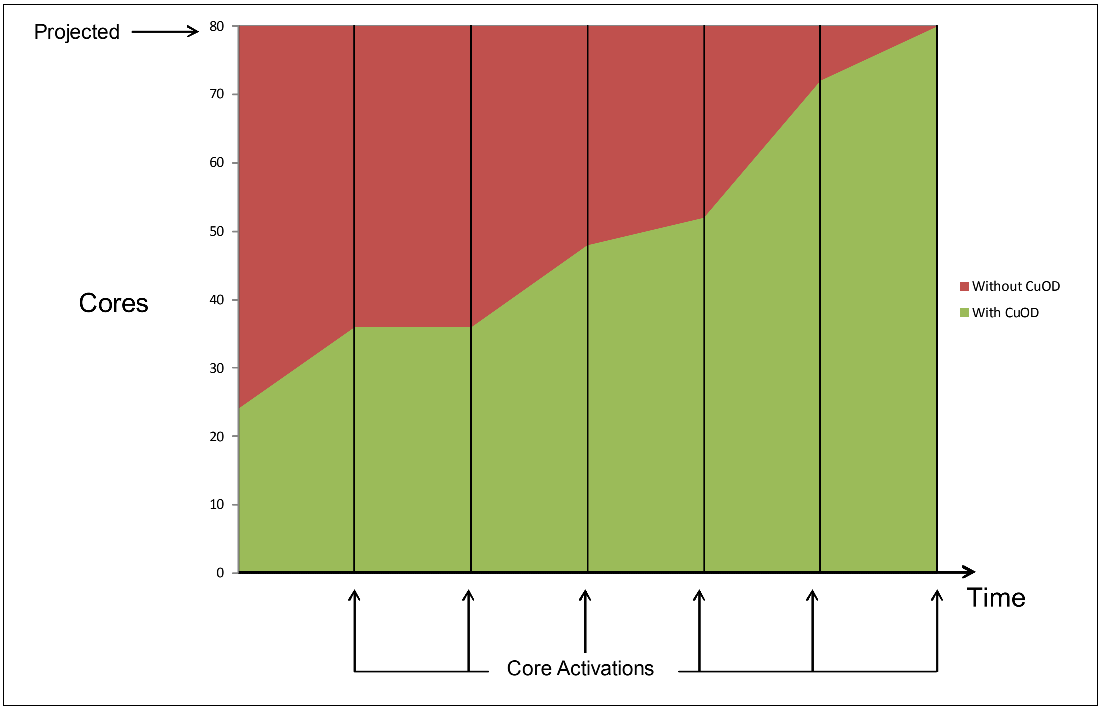


Figure 2-18 Active cores scenarios comparison during a project lifecycle

Table 2-16 lists the static processor activation features that are available for the Power E870 and Power E880.

Table 2-16 Power Systems CUoD static processor activation features

System	Processor feature	Processor core static activation feature
Power E870	#EPBA (4.02 GHz Processor Card)	EPBJ
Power E870	#EPBC (4.19 GHz Processor Card)	EPBL
Power E880	#EPBB (4.35 GHz Processor Card)	EPBK
Power E880	#EPBD (4.02 GHz Processor Card)	EPBM

Table 2-17 lists the static memory activation features that are available for the Power E870 and Power E880.

Table 2-17 Power Systems CUoD static memory activation features

System	Description	Feature code
Power E870, Power E880	Activation of 1 GB DDR3 POWER8 memory	EMA5
Power E870, Power E880	Activation of 100 GB DDR3 POWER8 memory	EMA6

2.4.2 Power enterprise pools and mobile capacity on demand (Mobile CoD)

While static activations are valid for a single system, some customers could benefit from moving processor and memory activations among different servers due to workload rebalance or disaster recovery.

IBM power enterprise pools, is a technology for dynamically sharing processor and memory activations among a group (or pool) of IBM Power Systems servers. Using Mobile Capacity on Demand (CoD) activation codes, the systems administrator can perform tasks without contacting IBM.

Two types of power enterprise pools are available:

- Power 770 (9117-MMD) and Power E870 (9119-MME) class systems
- Power 780 (9117-MHD), Power 795 (9119-FHB), and Power E880 (9119-MHE) class systems

Each pool type can support systems with different clock speeds or processor generations.

The basic rules for the Mobile Capacity on Demand follows:

- The Power 770 and Power 780 systems require a minimum of four static processor activations.
- The Power 870 and Power 880 require a minimum of eight static processor activations.
- The Power 795 requires a minimum of 24 static processor activations or 25% of the installed processor capacity whichever is bigger.
- For all systems, 25% of the active memory capacity must have static activations.

All the systems in a pool must be managed by the same HMC or by the same pair of redundant HMCs. If redundant HMCs are used, the HMCs must be connected to a network so that they can communicate with each other. The HMCs must have at least 2 GB of memory.

An HMC can manage multiple power enterprise pools and can also manage systems that are not part of a power enterprise pool. Systems can belong to only one power enterprise pool at a time. Powering down an HMC does not limit the assigned resources of participating systems in a pool but does limit the ability to perform pool change operations.

After a power enterprise pool is created, the HMC can be used to perform the following functions:

- Mobile CoD processor and memory resources can be assigned to systems with inactive resources. Mobile CoD resources remain on the system to which they are assigned until they are removed from the system.
- New systems can be added to the pool and existing systems can be removed from the pool.
- New resources can be added to the pool or existing resources can be removed from the pool.
- Pool information can be viewed, including pool resource assignments, compliance, and history logs.

In order for the Mobile activation features to be configured, it is necessary that a power enterprise pool is registered with IBM as well as the systems that are going to be included as members of the pool. Also, it is necessary that the systems have the feature code #EB35 for mobile enablement configured, and the required contracts must be in place.

Table 2-18 lists the mobile processor activation features that are available for the Power E870 and Power E880.

Table 2-18 Mobile processor activation features

System	Description	CUoD mobile processor core activation feature
Power E870	1-Core Mobile activation	#EP2S
Power E880	1-Core Mobile activation	#EP2T

Table 2-19 lists the mobile memory activation features that are available for the Power E870 and Power E880.

Table 2-19 Mobile memory activation features

System	Description	Feature code
Power E870, Power E880	100 GB Mobile memory activation	#EMA7

For more information about power enterprise pools, see the Redpaper publication, *Power Enterprise Pools on IBM Power Systems*, REDP5101:

http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp5101.html?Open

2.4.3 Elastic Capacity on Demand (Elastic CoD)

> **Note:** Some web sites or documents still refer to Elastic Capacity on Demand as On/Off Capacity on Demand.

With the Elastic CoD offering, you can temporarily activate and deactivate processor cores and memory units to help meet the demands of business peaks such as seasonal activity, period-end, or special promotions. Elastic CoD was previously called On/Off CoD. When you order an Elastic CoD feature, you receive an enablement code that allows a system operator to make requests for additional processor and memory capacity in increments of one processor day or 1 GB memory day. The system monitors the amount and duration of the activations. Both prepaid and post-pay options are available.

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

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

New to both Power E870 and Power E880 are 90-day temporary Elastic CoD processor and memory enablement features. These features enable a system to temporarily activate all inactive processor and memory CoD resources for a maximum of 90 days before ordering another temporary elastic enablement. A feature code is required.

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

If a Power E870 or Power E880 server uses the IBM i operating system in addition to any other supported operating system on the same server, the client must inform IBM which operating system caused the temporary Elastic CoD processor usage so that the correct feature can be used for billing.

The features that are used to order enablement codes and support billing charges on the Power E870 and Power E880 are described in 1.4, “System features” on page 12 and 1.4.7, “Memory features” on page 20.

The Elastic CoD process consists of three steps, enablement, activation, and billing.

- Enablement

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

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

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

- Activation requests

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

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

- Billing

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

For more information about registration, enablement, and usage of Elastic CoD, visit the following location:

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

2.4.4 Utility Capacity on Demand (Utility CoD)

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

With Utility CoD, you can place a quantity of inactive processors into the server’s shared processor pool, which then becomes available to the pool's resource manager. When the server recognizes that the combined processor utilization within the shared processor pool exceeds 100% of the level of base (purchased and active) processors that are assigned across uncapped partitions, then a Utility CoD processor minute is charged and this level of performance is available for the next minute of use.

If additional workload requires a higher level of performance, the system automatically allows the additional Utility CoD processors to be used, and the system automatically and continuously monitors and charges for the performance needed above the base (permanent) level.

Registration and usage reporting for utility CoD is made using a public website and payment is based on reported usage. Utility CoD requires PowerVM Standard Edition or PowerVM Enterprise Edition to be active.

If a Power E870 or Power E880 server uses the IBM i operating system in addition to any other supported operating system on the same server, the client must inform IBM which operating system caused the temporary Utility CoD processor usage so that the correct feature can be used for billing.

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

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

2.4.5 Trial Capacity on Demand (Trial CoD)

A *standard request* for Trial CoD requires you to complete a form including contact information and vital product data (VPD) from your Power E870 or Power E880 system with inactive CoD resources.

A standard request activates two processors or 64 GB of memory (or 8 processor cores and 64 GB of memory) for 30 days. Subsequent standard requests can be made after each purchase of a permanent processor activation. An HMC is required to manage Trial CoD activations.

An *exception request* for Trial CoD requires you to complete a form including contact information and VPD from your Power E870 or Power E880 system with inactive CoD resources. An exception request will activate all inactive processors or all inactive memory (or all inactive processor and memory) for 30 days. An exception request can be made only one time over the life of the machine. An HMC is required to manage Trial CoD activations.

To request either a Standard or an Exception Trial, visit the following location:

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

2.4.6 Software licensing and CoD

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

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

2.5 System bus

This section provides additional information related to the internal buses.

2.5.1 PCI Express Gen3

The internal I/O subsystem on the Power E870 and Power E880 is connected to the PCIe Controllers on a POWER8 processor in the system. Each POWER8 processor module has two buses that have 16 PCIe lanes each (for a total of 32 PCIe lanes) running at 7.877 Gbps full-duplex and provides 31.508 GBps of I/O connectivity to the PCIe slots. A diagram with the connections can be seen in Figure 2-19.

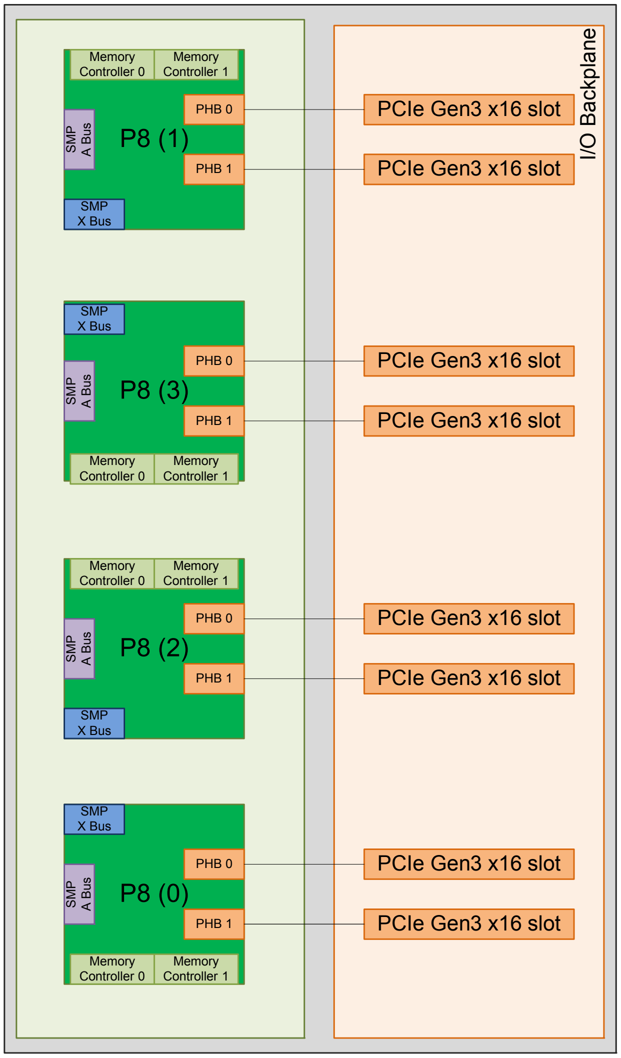


Figure 2-19 System nodes PCIe slots directly attached to PCIe controllers on POWER8 chips

Besides the slots directly attached to the processors PCI Gen3 controllers, the systems also allow for additional PCIe adapters on external PCIe Expansion Drawers and disks on external drawers connected through PCIe SAS adapters.

Figure 2-20 on page 74 shows a diagram with the I/O connectivity options available for the Power E870 and Power E880. The system nodes allow for eight PCIe Gen3 x16 slots. Additional slots can be added by attaching PCIe Expansion Drawers and SAS disks can be attached to EXP24S SFF Gen2 Drawers. The EXP24S can be either attached to SAS adapters on the system nodes or on the PCIe Expansion Drawer.

For a list of adapters and their supported slots, see 2.7, “PCI adapters” on page 77.

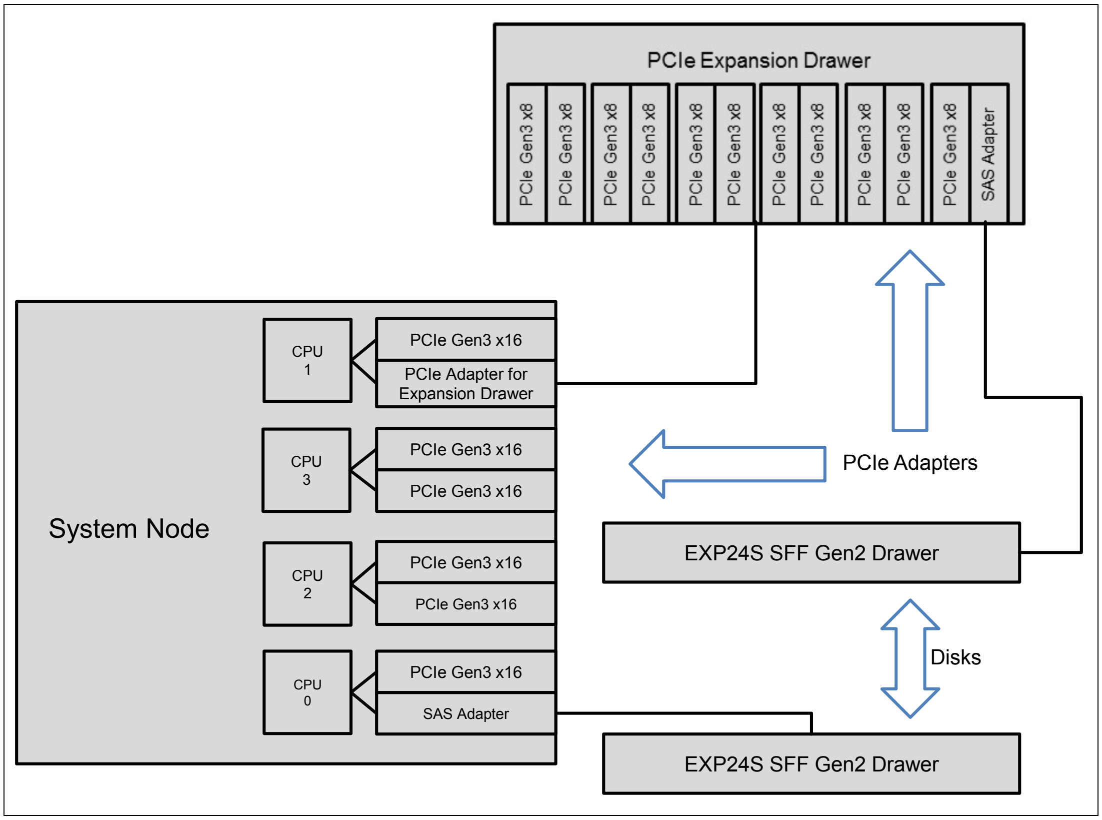


Figure 2-20 I/O connectivity options available for Power E870 and Power E880

Disk support: There is no support for disks directly installed on the system nodes and PCIe Expansion Drawers. If directly attached SAS disk are required, they need to be installed in a SAS disk drawer and connected to a supported SAS controller in one of the PCIe slots.

For more information about PCIe Expansion Drawers, see 2.9.1, “PCIe Gen3 I/O expansion drawer” on page 85.

2.5.2 Service Processor Bus

The redundant service processor bus connectors are located on the rear of the control unit and the system nodes. All of the service processor (SP) communication between the control unit and the system nodes flows though these cables.

Unlike the previous generations where a given pair of enclosures would host the service processors, on Power E870 and Power E880 as a standard, redundant service processor cards are installed on the control unit as well as redundant clock cards.

The cables used to provide communications between the control units and system nodes depend on the amount of system nodes installed. When a system node is added, a new set of cables is also added.

The cables necessary for each system node have been grouped under a single feature code, allowing for an easier configuration. Each cable set includes a pair of FSP cables, a pair of clock cables, and when applicable SMP cables and UPIC cables. Table 2-20 shows a list of the feature codes available.

Table 2-20 Features for cable sets

Feature code	Description
ECCA	System node to system control unit cable set for drawer 1
ECCB	System node to system control unit cable set for drawer 2
ECCC	System node to system control unit cable set for drawer 3
ECCD	System node to system control unit cable set for drawer 4

Cable sets feature codes are incremental and depend on the number of installed drawers as follows:

- 1 system node: #ECCA
- 2 system nodes: #ECCA and #ECCB
- 3 system nodes: #ECCA, #ECCB, and #ECCC
- 4 system nodes: #ECCA, #ECCB, #ECCC, and #ECCD

The system connection topology is shown in 2.1, “Logical diagrams” on page 38.

2.6 Internal I/O subsystem

The internal I/O subsystem resides on the I/O planar, which supports eight PCIe Gen3 x16 slots. All PCIe slots are hot-pluggable and enabled with enhanced error handling (EEH). In the unlikely event of a problem, EEH-enabled adapters respond to a special data packet that is generated from the affected PCIe 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 more information about RAS on the I/O buses, see 4.3.11, “I/O subsystem availability and Enhanced Error Handling” on page 169.

Table 2-21 lists the slot configuration of Power E870 and Power E880 system nodes.

Table 2-21 Slot configuration and capabilities

Slot	Location code	Slot type	CAPI capable[a]	SRIOV capable
Slot 1	P1-C1	PCIe Gen3 x16	No	Yes
Slot 2	P1-C2	PCIe Gen3 x16	Yes	Yes
Slot 3	P1-C3	PCIe Gen3 x16	No	Yes
Slot 4	P1-C4	PCIe Gen3 x16	Yes	Yes
Slot 5	P1-C5	PCIe Gen3 x16	No	Yes
Slot 6	P1-C6	PCIe Gen3 x16	Yes	Yes
Slot 7	P1-C7	PCIe Gen3 x16	No	Yes
Slot 8	P1-C8	PCIe Gen3 x16	Yes	Yes

[a] At the time of writing, there are no supported CAPI adapters for the Power E870 and E880 system unit.

The physical location of the slots can be seen in Figure 2-21.

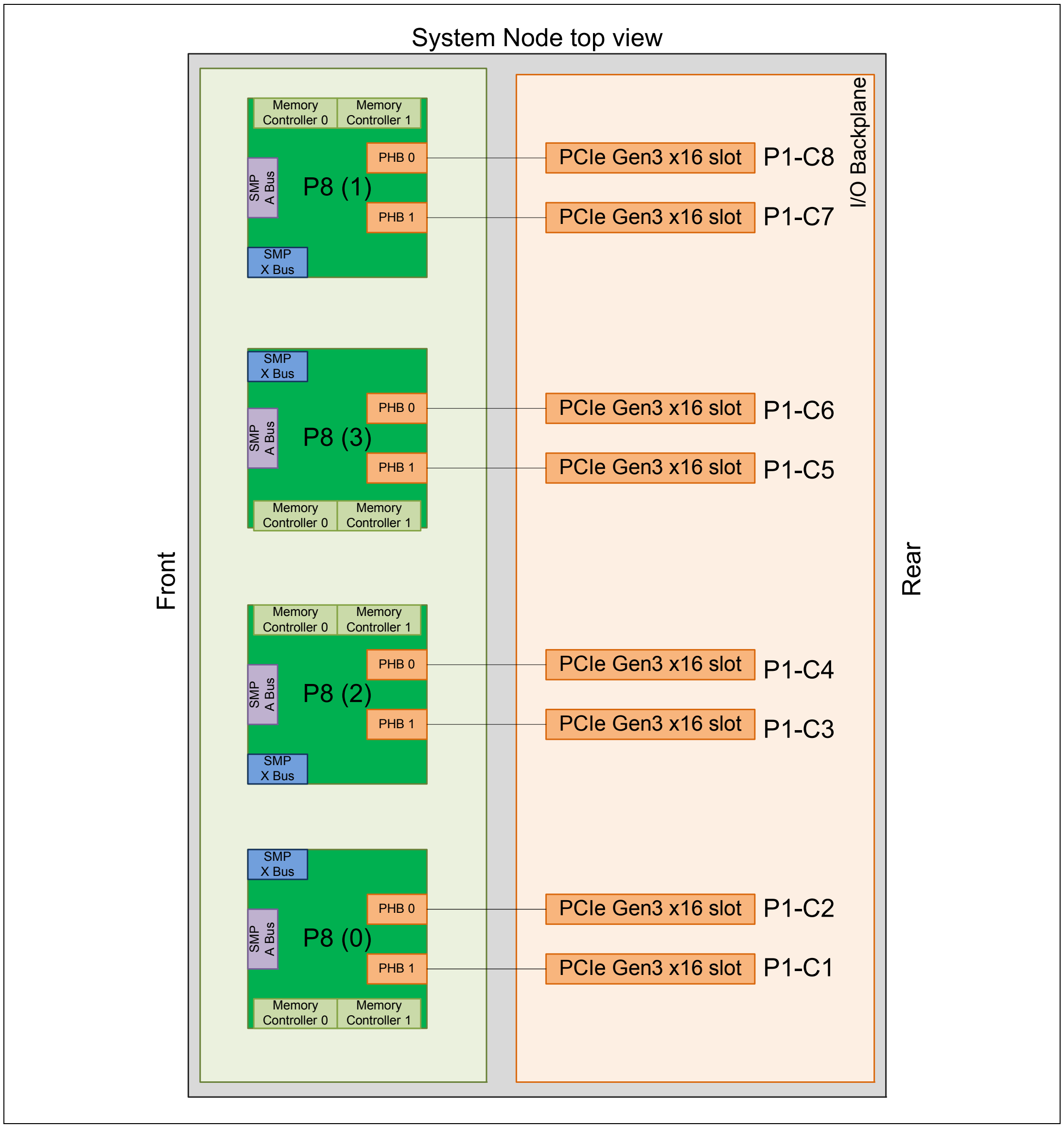


Figure 2-21 System node top view and PCIe slot location codes

2.6.1 Blind-swap cassettes

The Power E870 and Power E880 use a next generation blind-swap cassette to manage the installation and removal of PCIe adapters. This mechanism requires an interposer card that allows the PCIe adapters to plug in vertically to the system, allows more airflow through the cassette, and allows for faster hot swap procedures. Cassettes can be installed and removed without removing the system nodes or PCIe expansion drawers from the rack.

2.6.2 System ports

The system nodes do not have integrated ports. All networking and storage for the virtual machines must be provided via PCIe adapters installed in standard PCIe slots.

The system control unit has one USB Port dedicated to the DVD drive and 4 Ethernet ports used for HMC communications. There is no serial port so an HMC is mandatory for system management. The FSP's virtual console will be on the HMC.

The location of the USB and HMC Ethernet ports can be seen in Figure 2-22.

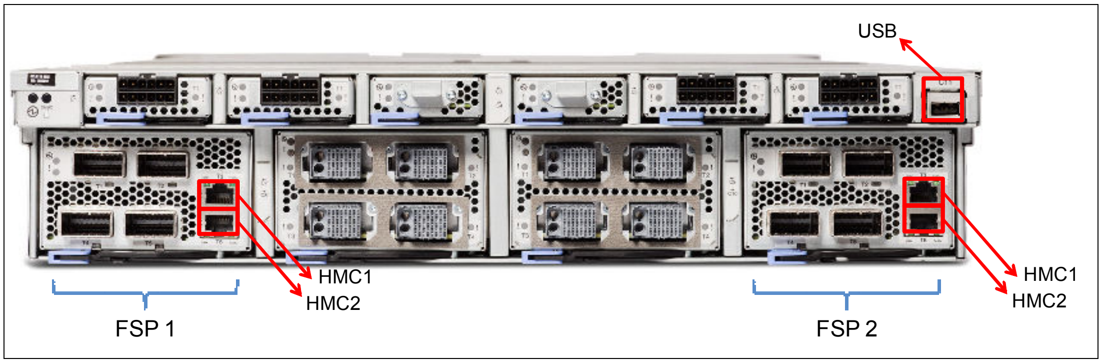


Figure 2-22 Physical location of the USB and HMC ports on the system control unit

The connection and usage of the DVD can be seen in detail in 2.8.1, "DVD" on page 84.

2.7 PCI adapters

This section covers the types and functions of the PCI cards supported by Power E870 and Power E880 systems.

2.7.1 PCI Express (PCIe)

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 that uses two pairs of wires, one pair for transmit and one pair for receive, and can transmit only one bit per cycle. These two pairs of wires are called a *lane*. A PCIe link can consist of multiple lanes. In such configurations, the connection is labelled as x1, x2, x8, x12, x16, or x32, where the number is effectively the number of lanes.

The PCIe interfaces supported on this server are PCIe Gen3, capable of 16 GBps simplex (32 GBps duplex) on a single x16 interface. PCIe Gen3 slots also support previous generations (Gen2 and Gen1) adapters, which operate at lower speeds, according to the following rules:

- Place x1, x4, x8, and x16 speed adapters in same connector size slots first, before mixing adapter speed with connector slot size.
- Adapters with smaller speeds are allowed in larger sized PCIe connectors but larger speed adapters are not compatible in smaller connector sizes (i.e. a x16 adapter cannot go in an x8 PCIe slot connector).

IBM POWER8 processor-based servers can support two different form factors of PCIe adapters:

- PCIe low profile (LP) cards, which are used with system node PCIe slots.

- PCIe full height and full high cards are used in the PCIe Gen3 I/O expansion drawer (#EMX0)

Low-profile PCIe adapter cards are supported only in low-profile PCIe slots, and full-height and full-high cards are supported only in full-high slots.

Before adding or rearranging adapters, use the System Planning Tool to validate the new adapter configuration. For more information, see the System Planning Tool website:

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

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

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

The following sections describe the supported adapters and provide tables of orderable feature numbers. The tables indicate operating system support (AIX, IBM i, and Linux) for each of the adapters.

2.7.2 LAN adapters

To connect the Power E870 and Power E880 servers to a local area network (LAN), you can use the LAN adapters that are supported in the PCIe slots of the system. Table 2-22 lists the available LAN adapters. Information about FCoE adapters can be found in Table 2-26 on page 82.

Table 2-22 Available LAN adapters

Feature Code	CCIN	Description	Placement	OS support
5260	576F	PCIe2 LP 4-port 1 GbE Adapter	CEC	AIX, IBM i, Linux
5274	5768	PCIe LP 2-Port 1 GbE SX Adapter	CEC	AIX, IBM i, Linux
5744	2B44	PCIe2 4-Port 10 GbE&1 GbE SR&RJ45 Adapter	I/O drawer	Linux
5767	5767	2-Port 10/100/1000 Base-TX Ethernet PCI Express Adapter	I/O drawer	AIX, IBM i, Linux
5768	5768	2-Port Gigabit Ethernet-SX PCI Express Adapter	I/O drawer	AIX, IBM i, Linux
5769	5769	10 Gigabit Ethernet-SR PCI Express Adapter	I/O drawer	AIX, IBM i, Linux
5772	576E	10 Gigabit Ethernet-LR PCI Express Adapter	I/O drawer	AIX, IBM i, Linux
5899	576F	PCIe2 4-port 1 GbE Adapter	I/O drawer	AIX, IBM i, Linux
EC29	EC29	PCIe2 LP 2-Port 10 GbE RoCE SR Adapter	CEC	AIX, IBM i, Linux
EC2M	57BE	PCIe3 LP 2-port 10 GbE NIC&RoCE SR Adapter	CEC	AIX, IBM i, Linux

Feature Code	CCIN	Description	Placement	OS support
EC2N		PCIe3 2-port 10 GbE NIC&RoCE SR Adapter	I/O drawer	AIX, IBM i, Linux
EC37	57BC	PCIe3 LP 2-port 10 GbE NIC&RoCE SFP+ Copper Adapter	CEC	AIX, IBM i, Linux
EC38		PCIe3 2-port 10 GbE NIC&RoCE SFP+ Copper Adapter	I/O drawer	AIX, IBM i, Linux
EC3A	57BD	PCIe3 LP 2-Port 40 GbE NIC RoCE QSFP+ Adapter	CEC	AIX, IBM i, Linux
EC3B	57B6	PCIe3 2-Port 40 GbE NIC RoCE QSFP+ Adapter	I/O drawer	AIX, IBM i, Linux
EC3L		PCIe3 LP 2-port 100GbE (NIC& RoCE) QSFP28 Adapter x16	I/O drawer	AIX, IBM i, Linux
EN0M	2CC0	PCIe2 4-port(10Gb FCoE & 1 GbE) LR&RJ45 Adapter	I/O drawer	AIX, IBM i, Linux
EN0N	2CC0	PCIe2 LP 4-port(10Gb FCoE & 1 GbE) LR&RJ45 Adapter	CEC	AIX, IBM i, Linux
EN0S	2CC3	PCIe2 4-Port (10Gb+1 GbE) SR+RJ45 Adapter	I/O drawer	AIX, IBM i, Linux
EN0T	2CC3	PCIe2 LP 4-Port (10Gb+1 GbE) SR+RJ45 Adapter	CEC	AIX, IBM i, Linux
EN0U	2CC3	PCIe2 4-port (10Gb+1 GbE) Copper SFP+RJ45 Adapter	I/O drawer	AIX, IBM i, Linux
EN0V	2CC3	PCIe2 LP 4-port (10Gb+1 GbE) Copper SFP+RJ45 Adapter	CEC	AIX, IBM i, Linux
EN0W	2CC4	PCIe2 2-port 10/1 GbE BaseT RJ45 Adapter	I/O drawer	AIX, IBM i, Linux
EN0X	2CC4	PCIe2 LP 2-port 10/1 GbE BaseT RJ45 Adapter	CEC	AIX, IBM i, Linux
EN15	2CE3	PCIe3 4-port 10 GbE SR Adapter	I/O drawer	AIX, IBM i, Linux
EN16		PCIe3 LPX 4-port 10 GbE SR Adapter	CEC	AIX, IBM i, Linux
EN17	2CE4	PCIe3 4-port 10 GbE SFP+ Copper Adapter		AIX, IBM i, Linux
EN18		PCIe3 LPX 4-port 10 GbE SFP+ Copper Adapter		AIX, IBM i, Linux

2.7.3 Graphics accelerator adapters

Table 2-23 lists the available graphics accelerator adapters. An adapter can be configured to operate in either 8-bit or 24-bit color modes. The adapter supports both analog and digital monitors.

Table 2-23 Available graphics accelerator adapters

Feature Code	CCIN	Description	Placement	OS support
5269	5269	PCIe LP POWER GXT145 Graphics Accelerator	CEC	AIX, Linux
EC41		PCIe2 LP 3D Graphics Adapter x1	CEC	AIX, Linux

2.7.4 SAS adapters

Table 2-24 lists the SAS adapters that are available for Power E870 and Power E880 systems.

Table 2-24 Available SAS adapters

Feature Code	CCIN	Description	Placement	OS support
5901	57B3	PCIe Dual-x4 SAS Adapter	I/O drawer	AIX, Linux
5913	57B5	PCIe2 1.8 GB Cache RAID SAS Adapter Tri-port 6Gb	I/O drawer	AIX, Linux
ESA3	57BB	PCIe2 1.8 GB Cache RAID SAS Adapter Tri-port 6Gb CR	I/O drawer	AIX, Linux
EJ0J	57B4	PCIe3 RAID SAS Adapter Quad-port 6Gb x8	I/O drawer	AIX, Linux
EJ0L	57CE	PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6Gb x8	I/O drawer	AIX, Linux
EJ0M		PCIe3 LP RAID SAS ADAPTER	CEC	AIX, Linux
EJ10	57B4	PCIe3 SAS Tape/DVD Adapter Quad-port 6Gb x8	I/O drawer	AIX, Linux
EJ11	57B4	PCIe3 LP SAS Tape/DVD Adapter Quad-port 6Gb x8	CEC	AIX, Linux
EJ14	57B1	PCIe3 12 GB Cache RAID PLUS SAS Adapter Quad-port 6Gb x8	I/O drawer	AIX, IBM i, Linux

2.7.5 Fibre Channel adapter

The systems support direct or SAN connection to devices that use Fibre Channel adapters. Table 2-25 summarizes the available Fibre Channel adapters, which all have LC connectors.

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

Table 2-25 Available Fibre Channel adapters

Feature Code	CCIN	Description	Placement	OS support
5273	577D	PCIe LP 8Gb 2-Port Fibre Channel Adapter	CEC	AIX, IBM i, Linux

Feature Code	CCIN	Description	Placement	OS support
5276	5774	PCIe LP 4 Gb 2-Port Fibre Channel Adapter	CEC	AIX, IBM i, Linux
5729	5729	PCIe2 8Gb 4-port Fibre Channel Adapter	I/O drawer	AIX, Linux
5735	577D	Gigabit PCI Express Dual Port Fibre Channel Adapter	I/O drawer	AIX, IBM i, Linux
5774	5774	4 Gigabit PCI Express Dual Port Fibre Channel Adapter	I/O drawer	AIX, IBM i, Linux
EN0A	577F	PCIe2 16Gb 2-port Fibre Channel Adapter	I/O drawer	AIX, Linux
EN0B	577F	PCIe2 LP 16Gb 2-port Fibre Channel Adapter	CEC	AIX, IBM i, Linux
EN0F	578D	PCIe2 LP 8Gb 2-Port Fibre Channel Adapter	CEC	AIX, Linux
EN0G		PCIe2 8Gb 2-Port Fibre Channel Adapter	I/O drawer	AIX, Linux
EN0Y	EN0Y	PCIe2 LP 8Gb 4-port Fibre Channel Adapter	CEC	AIX, IBM i, Linux
EN12		PCIe2 8Gb 4-port Fibre Channel Adapter	I/O drawer	AIX, Linux

2.7.6 Fibre Channel over Ethernet

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

Figure 2-23 compares existing Fibre Channel and network connections and FCoE connections.

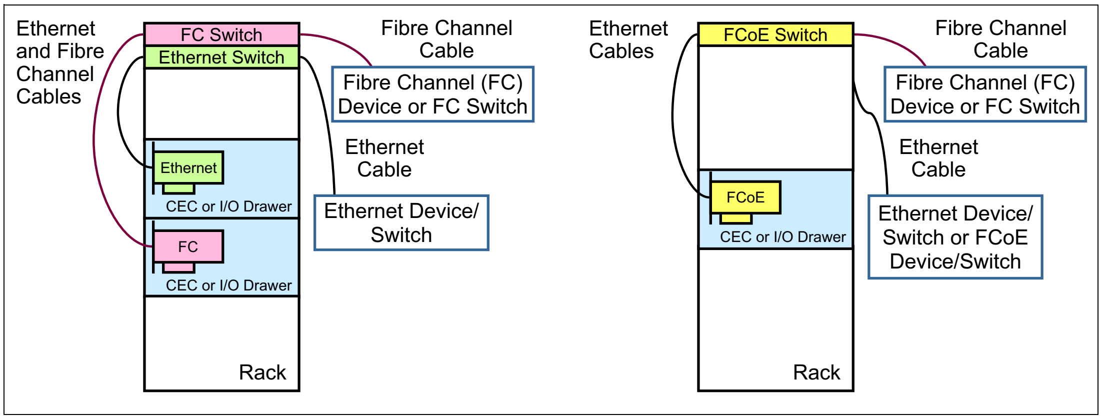


Figure 2-23 Comparison between existing Fibre Channel and network connections and FCoE connections

Table 2-26 lists the available FCoE adapters. They are high-performance Converged Network Adapters (CNAs) using SR optics. Each port can simultaneously provide network interface card (NIC) traffic and Fibre Channel functions.

Table 2-26 Available FCoE adapters

Feature Code	CCIN	Description	Placement	OS support
EN0H	2B93	PCIe2 4-port (10Gb FCoE & 1 GbE) SR&RJ45	I/O drawer	AIX, IBM i, Linux
EN0J	2B93	PCIe2 LP 4-port (10Gb FCoE & 1 GbE) SR&RJ45	CEC	AIX, IBM i, Linux
EN0K	2CC1	PCIe2 4-port (10Gb FCoE & 1 GbE) SFP+Copper&RJ45	I/O drawer	AIX, IBM i, Linux
EN0L	2CC1	PCIe2 LP 4-port(10Gb FCoE & 1 GbE) SFP+Copper&RJ45	CEC	AIX, IBM i, Linux
EN0M		PCIe3 4-port(10Gb FCoE & 1GbE) LR&RJ45 Adapter	CEC	AIX, Linux
EN0N		PCIe3 LP 4-port(10Gb FCoE & 1GbE) LR&RJ45 Adapter	CEC	AIX, Linux

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

Note: Adapters EN0J and EN0L support SR-IOV when minimum firmware and software levels are met. See 3.4, “Single root I/O virtualization (SR-IOV)” on page 126 for more information.

2.7.7 USB adapters

Each system control unit enclosure can have one slim-line bay that can support one DVD drive (#EU13). The DVD drive is cabled to a USB PCIe adapter located in either the system node or in a PCIe Gen3 I/O drawer.

Table 2-27 lists the available USB adapters.

Table 2-27 Available asynchronous and USB adapters

Feature Code	CCIN	Description	Placement	OS support
EC45		PCIe2 LP 4-Port USB 3.0 Adapter	CEC	AIX, IBM i, Linux
EC46		PCIe2 4-Port USB 3.0 Adapter	I/O drawer	AIX, IBM i, Linux

2.7.8 InfiniBand host channel adapter

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

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

http://www.infinibandta.org/

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

The InfiniBand physical connection consists of multiple byte lanes. Each individual byte lane is a four-wire, 2.5, 5.0, or 10.0 Gbps bidirectional connection. Combinations of link width and byte-lane speed allow for overall link speeds from 2.5 Gbps to 120 Gbps. The architecture defines a layered hardware protocol, and also 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.

Table 2-28 lists the available InfiniBand adapters.

Table 2-28 Available InfiniBand adapters

Feature Code	CCIN	Description	Placement	OS support
EC3E		PCIe3 LP 2-port 100Gb EDR IB Adapter x16	CEC	Linux
EC3T		PCIe3 LP 1-port 100Gb EDR IB Adapter x16	CEC	Linux

2.7.9 Cryptographic Coprocessor

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

The IBM PCIe Cryptographic Coprocessor adapter highlights the following features:

- Integrated Dual processors that operate in parallel for higher reliability
- Supports IBM Common Cryptographic Architecture or PKCS#11 standard
- Ability to configure adapter as coprocessor or accelerator
- Support for smart card applications using Europay, MasterCard, and Visa
- Cryptographic key generation and random number generation
- PIN processing: generation, verification, translation
- Encrypt and Decrypt using AES and DES keys

See the following site for the most recent firmware and software updates:

http://www.ibm.com/security/cryptocards/

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

Table 2-29 Available cryptographic adapters

Feature Code	CCIN	Description	Placement	OS support
EJ33		PCIe3 Crypto Coprocessor BSC-Gen3 4767	I/O drawer	AIX, IBM i, Linux

2.7.10 CAPI adapters

The available CAPI adapters for JAVA and WebSphere acelleration are shown in Table 2-30.

Table 2-30 Available CAPI adapters

Feature code	CCIN	Description	One Processor	Two Processors	OS support
EJ18		PCIe3 CAPI FlashSystem Accelerator Adapter	1	1	AIX

2.8 Internal storage

The system nodes for Power E870 and Power E880 do not allow for conventional physical storage. All storage must be provided externally via I/O expansion drawers or SAN. At the time of writing, the only external I/O expansion drawer that can be used in Power E870 and Power E880 is EXP24S, attached via SAS ports to a SAS PCIe adapter installed either in a system node drawer or in a PCIe expansion slot located in an I/O expansion drawer.

The system control unit has a DVD drive which is connected to an external USB port on the rear of the unit. In order to use the DVD drive, at least on PCIe USB adapter must be installed and connected via an USB cable to the DVD drive.

The following NVMe Flash Adapters are available for internal storage

- PCIe3 LP 1.6 TB NVMe Flash Adapter (#EC54)
- PCIe3 LP 3.2 TB NVMe Flash Adapter (#EC56)

2.8.1 DVD

There is one DVD media bay per system, located on the front of the system control unit. This media bay allows for a DVD (#EU13) to be installed on the system control unit and it enables the USB port on the rear of the control unit.

The USB port must be connected via a USB cable to a USB PCIe adapter, installed in one of the available PCIe slots on the system nodes or I/O expansion drawers. A diagram of the DVD connection can be seen in Figure 2-24 on page 85.

The basic rules for DVD drives on these system are as follows:

- Include a DVD drive with #EC13.
- Include the PCI USB adapter #EC45 or #EC46.
- Include a USB cable male-male with the proper length. As a suggestion, #EBK4 is a 1.6 m cable that allows enough length for the adapters on the first two system nodes to be connected to the USB DVD port.

This architecture allows for a more flexible infrastructure where the DVD drive is completely independent of another components and can be freely moved between partition.

It is also important to note that several daily functions where a DVD drive was needed can now be executed by using other methods such as Virtual Media Repository on the Virtual I/O Server or the Remote Virtual I/O Server install on HMC.

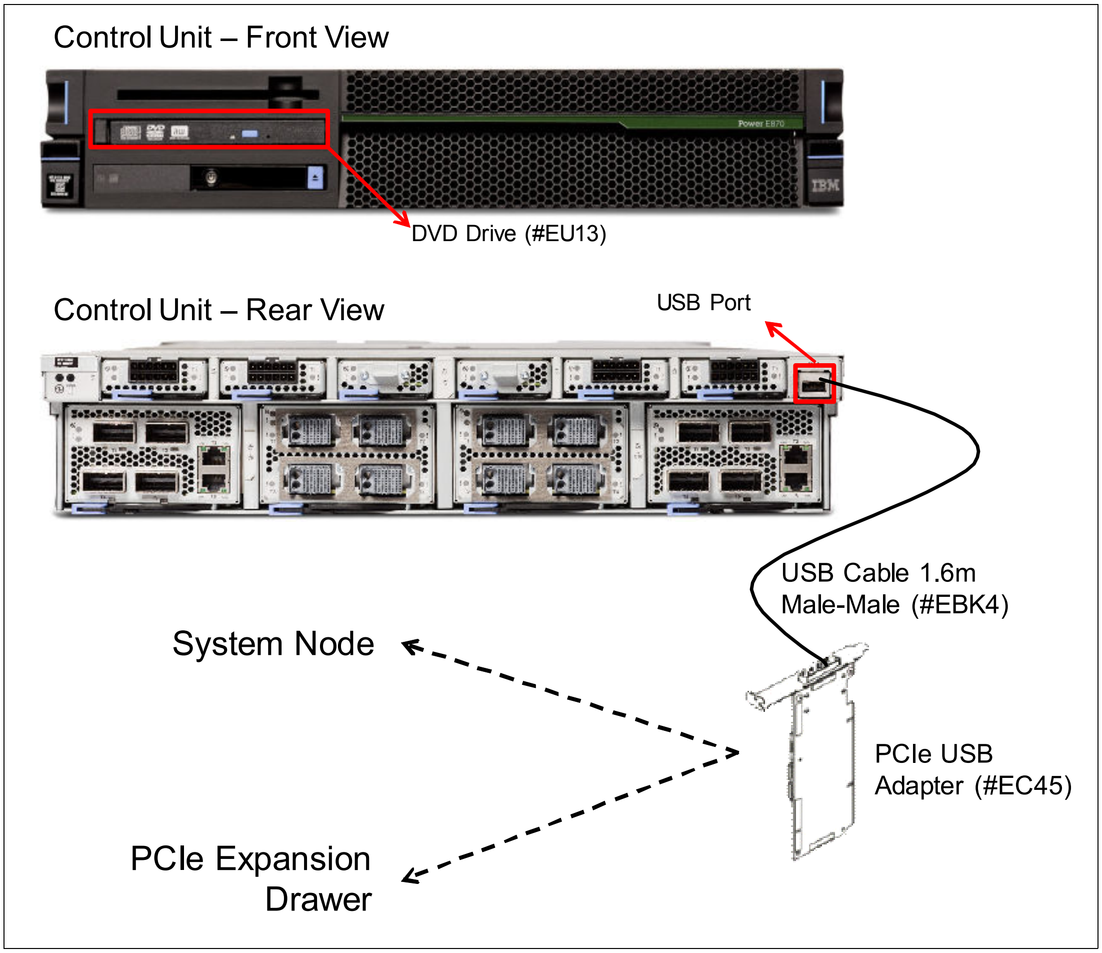


Figure 2-24 DVD drive physical location on control unit and suggested cabling

2.9 External I/O subsystems

This section describes the PCIe Gen3 I/O expansion drawer that can be attached to the Power E870 and Power E880.

2.9.1 PCIe Gen3 I/O expansion drawer

The PCIe Gen3 I/O expansion drawer is a 4U high, PCI Gen3-based and rack mountable I/O drawer. It offers two PCIe Fan Out Modules (#EMXF) each of them providing six PCIe slots.

The physical dimensions of the drawer are 444.5 mm (17.5 in.) wide by 177.8 mm (7.0 in.) high by 736.6 mm (29.0 in.) deep for use in a 19-inch rack.

A PCIe x16 to Optical CXP converter adapter (#EJ07) and 2.0 m (#ECC6), 10.0 m (#ECC8), or 20.0 m (#ECC9) CXP 16X Active Optical cables (AOC) connect the system node to a PCIe Fan Out module in the I/O expansion drawer. One feature #ECC6, one #ECC8, or one #ECC9 ships two AOC cables.

Concurrent repair and add/removal of PCIe adapter cards is done by HMC guided menus or by operating system support utilities.

A blind swap cassette (BSC) is used to house the full high adapters which go into these slots. The BSC is the same BSC as used with the previous generation server's #5802/5803/5877/5873 12X attached I/O drawers.

Figure 2-25 shows the back view of the PCIe Gen3 I/O expansion drawer.

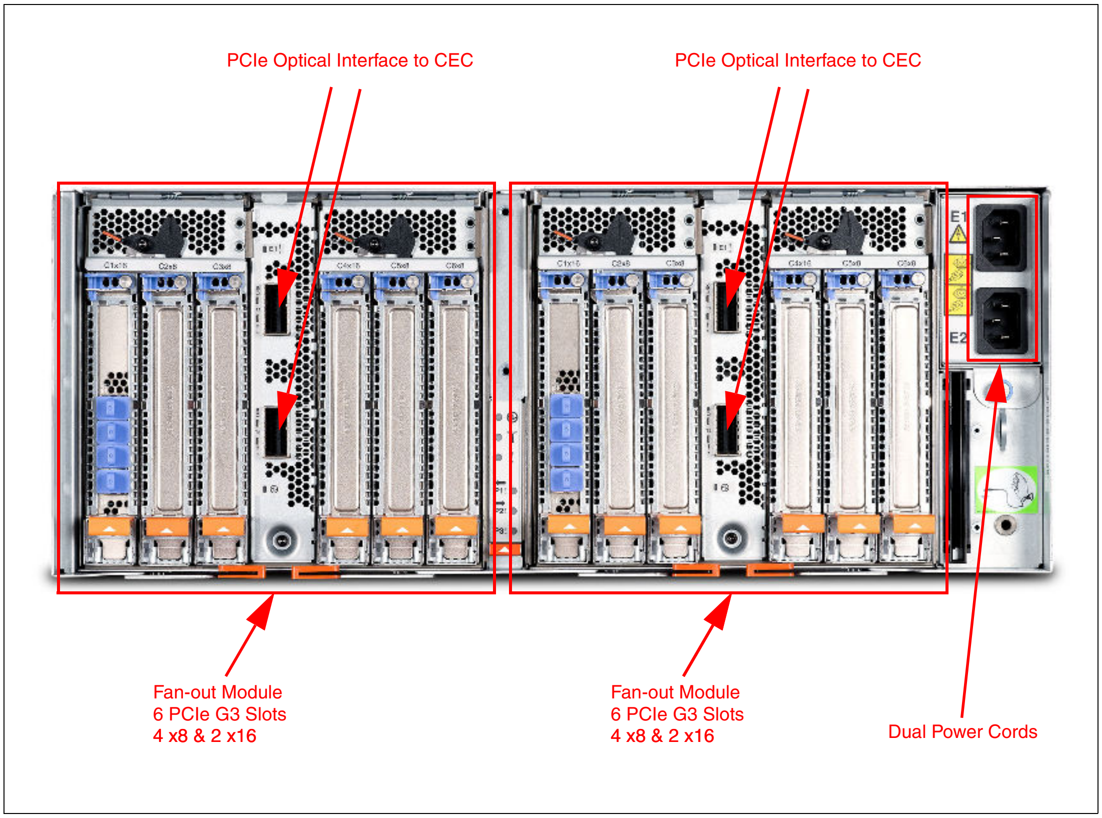


Figure 2-25 Rear view of the PCIe Gen3 I/O expansion drawer

2.9.2 PCIe Gen3 I/O expansion drawer optical cabling

I/O drawers are connected to the adapters in the system node with data transfer cables:

- 2.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC6)
- 10.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC8)
- 20.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC9)

Cable lengths: Use the 2.0 m cables for intra-rack installations. Use the 10.0 m or 20.0 m cables for inter-rack installations.

A minimum of one PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ07) is required to connect to the PCIe3 6-slot Fan Out module in the I/O expansion drawer. The top port of the fan out module must be cabled to the top port of the #EJ07 port. Likewise, the bottom two ports must be cabled together.

1. Connect an active optical cable to connector T1 on the PCIe3 optical cable adapter in your server.

2. Connect the other end of the optical cable to connector T1 on one of the PCIe3 6-slot Fan Out modules in your expansion drawer.
3. Connect another cable to connector T2 on the PCIe3 optical cable adapter in your server.
4. Connect the other end of the cable to connector T2 on the PCIe3 6-slot Fan Out module in your expansion drawer.
5. Repeat the four steps above for the other PCIe3 6-slot Fan Out module in the expansion drawer, if required.

> **Drawer connections:** Each Fan Out module in a PCIe3 Expansion Drawer can only be connected to a single PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ07). However the two Fan Out modules in a single I/O expansion drawer can be connected to different system nodes in the same server.

Figure 2-26 shows connector locations for the PCIe Gen3 I/O expansion drawer.

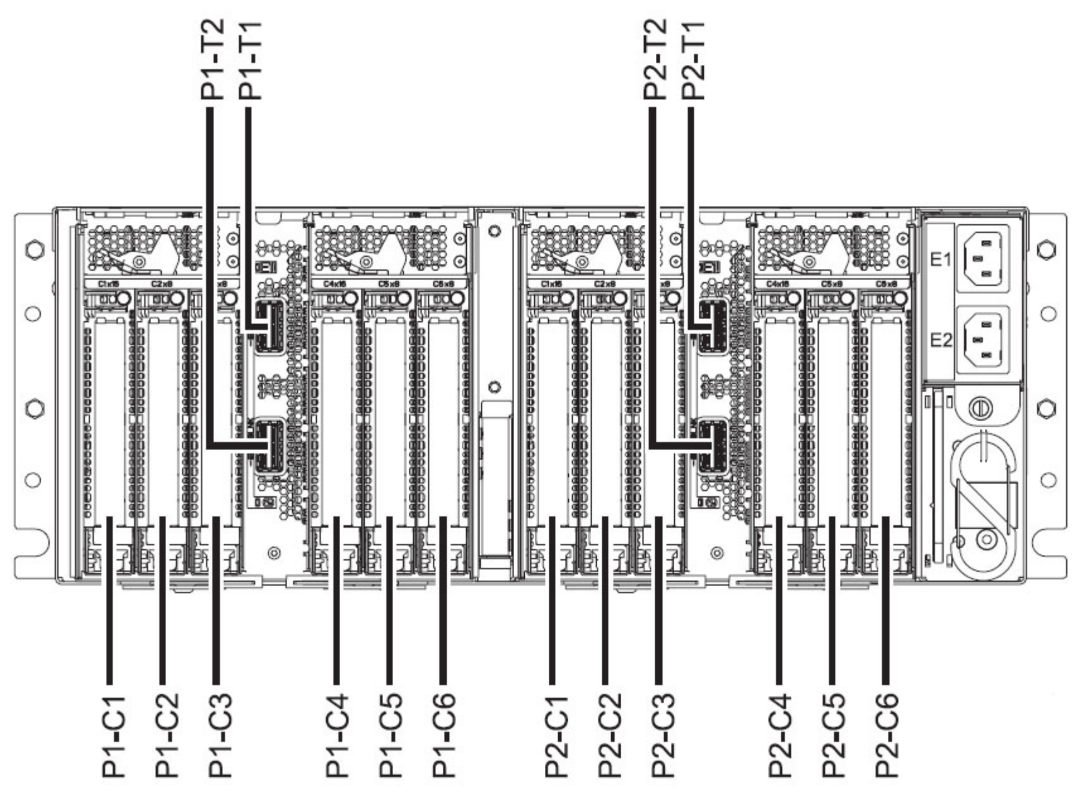


Figure 2-26 Connector locations for the PCIe Gen3 I/O expansion drawer

Figure 2-27 shows typical optical cable connections.

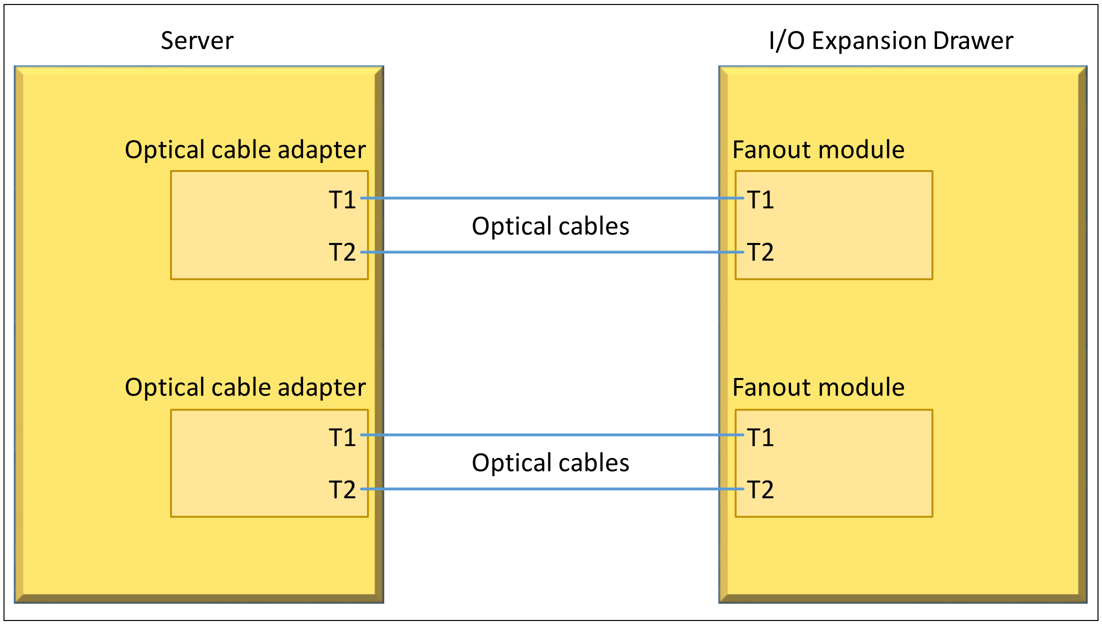


Figure 2-27 Typical optical cable connection

General rules for the PCI Gen3 I/O expansion drawer configuration

The PCIe3 optical cable adapter can be in any of the PCIe adapter slots in the Power E870 and Power E880 system node. However, we advise that you use the PCIe adapter slot priority information while selecting slots for installing PCIe3 Optical Cable Adapter (#EJ07).

Table 2-31 shows PCIe adapter slot priorities in the Power E870 and Power E880 system.

Table 2-31 PCIe adapter slot priorities

Feature code	Description	Slot priorities
EJ07	PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer	1, 7, 3, 5, 2, 8, 4, 6

The following figures show several examples of supported configurations. For simplification we have not shown every possible combination of the I/O expansion drawer to server attachments.

Figure 2-28 on page 89 shows an example of a single system node and two PCI Gen3 I/O expansion drawers.

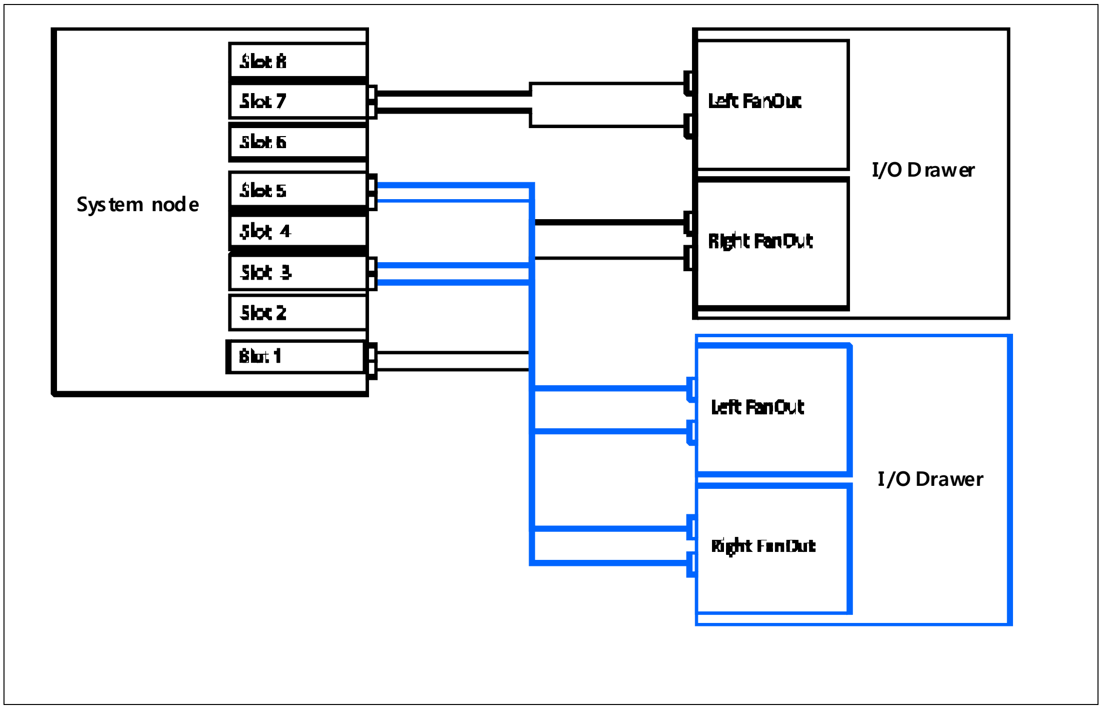


Figure 2-28 Example of a single system node and two I/O drawers

Figure 2-29 shows an example of two system nodes and two PCI Gen3 I/O expansion drawers.

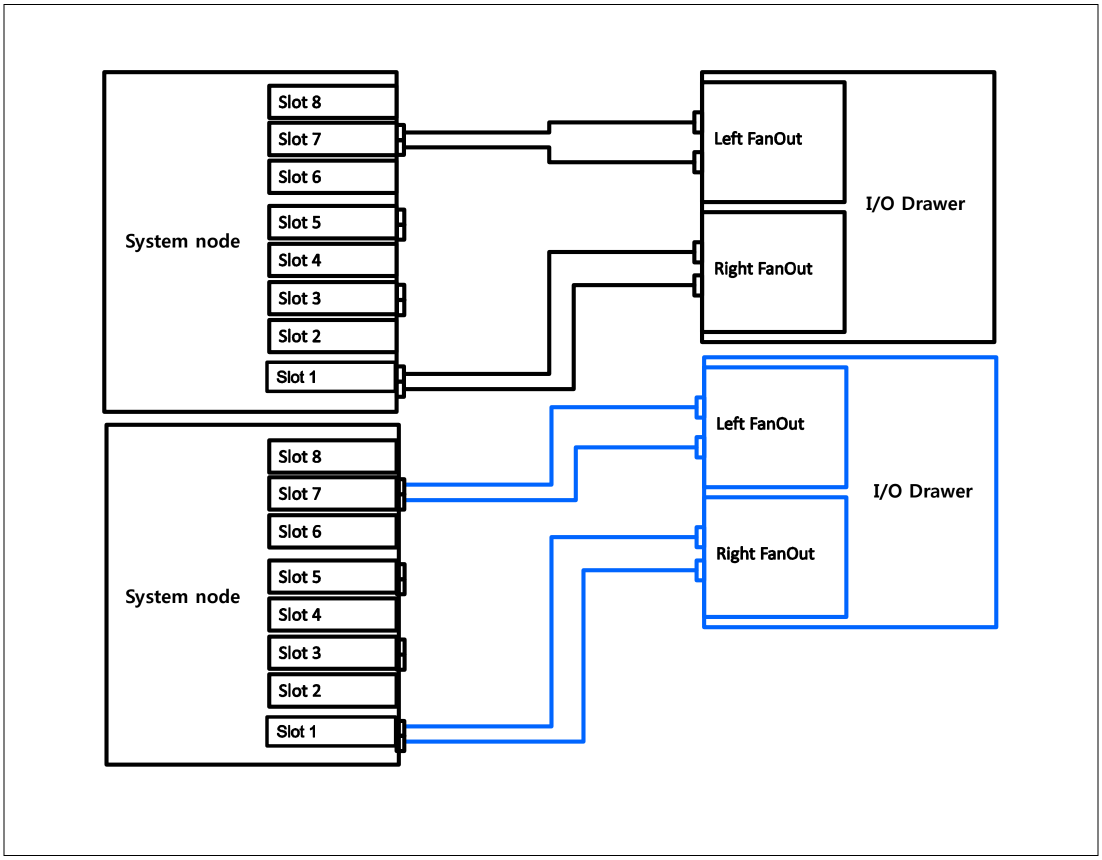


Figure 2-29 Example of two system nodes and two I/O drawers

Figure 2-30 shows an example of two system nodes and four PCI Gen3 I/O expansion drawers.

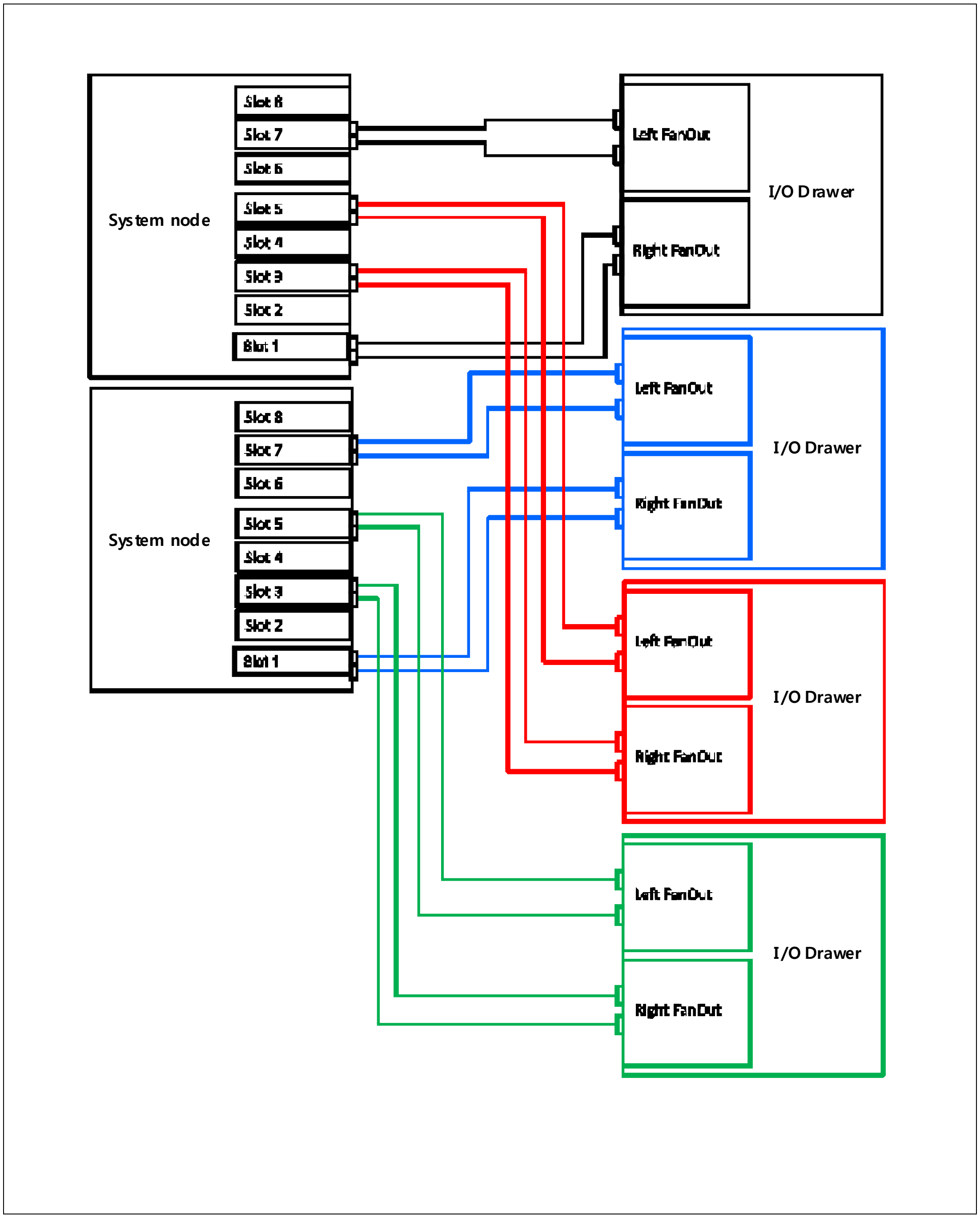


Figure 2-30 Example of two system nodes and four I/O drawers

2.9.3 PCIe Gen3 I/O expansion drawer SPCN cabling

There is no system power control network (SPCN) used to control and monitor the status of power and cooling within the I/O drawer. SPCN capabilities are integrated in the optical cables.

2.10 External disk subsystems

This section describes the following external disk subsystems that can be attached to the Power E870 and Power E880 system:

- EXP24S SFF Gen2-bay Drawer for high-density storage (#5887)
- IBM System Storage®

Note: The EXP30 Ultra SSD Drawer (#EDR1 or #5888), the EXP12S SAS Disk Drawer (#5886), and the EXP24 SCSI Disk Drawer (#5786) are not supported on the Power E870 and Power E880 server.

2.10.1 EXP24S SFF Gen2-bay Drawer

The EXP24S SFF Gen2-bay Drawer (#5887) is an expansion drawer with twenty-four 2.5-inch form-factor SAS bays. The EXP24S supports up to 24 hot-swap SFF-2 SAS hard disk drives (HDDs) or solid-state drives (SSDs). It uses only 2 EIA of space in a 19-inch rack. The EXP24S includes redundant ac power supplies and uses two power cords.

To maximize configuration flexibility and space utilization, the system node of Power E870 and Power E880 system does not have integrated SAS bays or integrated SAS controllers. PCIe SAS adapters and the EXP24S can be used to provide direct access storage.

To further reduce possible single points of failure, EXP24S configuration rules consistent with previous Power Systems are used. IBM i configurations require the drives to be protected (RAID or mirroring). Protecting the drives is highly advised, but not required for other operating systems. All Power operating system environments that are using SAS adapters with write cache require the cache to be protected by using pairs of adapters.

With AIX, Linux, and VIOS, you can order the EXP24S with four sets of six bays, two sets of 12 bays, or one set of 24 bays (mode 4, 2, or 1). With IBM i, you can order the EXP24S as one set of 24 bays (mode 1). Figure 2-31 shows the front of the unit and the groups of disks on each mode.

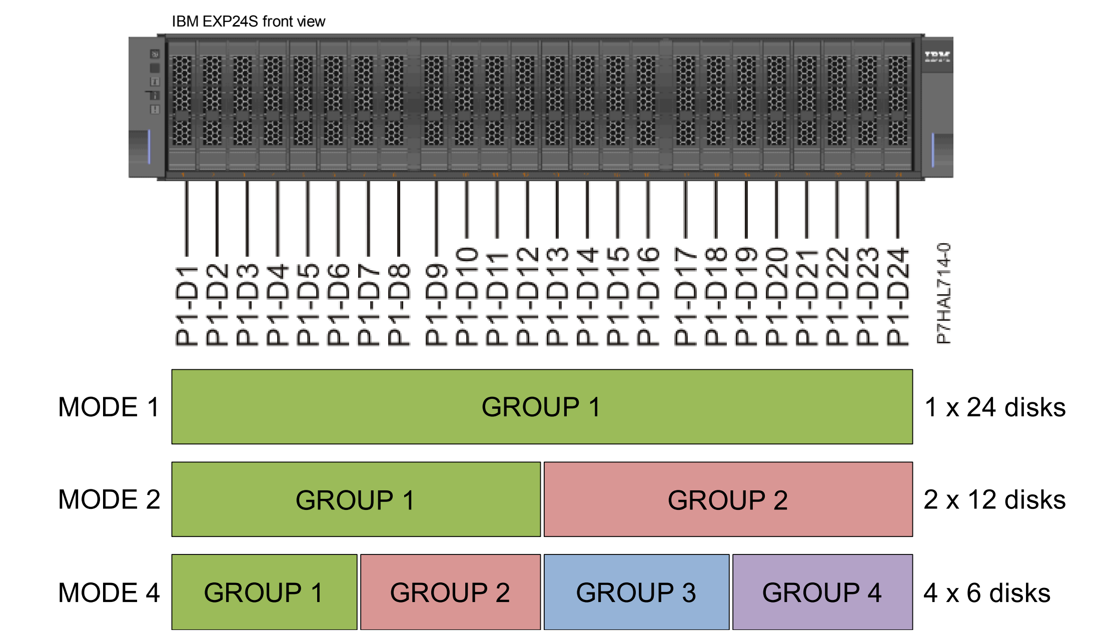


Figure 2-31 EXP24S front view with location codes and disk groups depending on its mode of operation

Mode setting is done by IBM manufacturing. If you need to change the mode after installation, ask your IBM support representative to refer to:

http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5121

The stickers indicate whether the enclosure is set to mode 1, mode 2, or mode 4. They are attached to the lower-left shelf of the chassis (A) and the center support between the enclosure services manager modules (B).

Figure 2-32 shows the mode stickers.

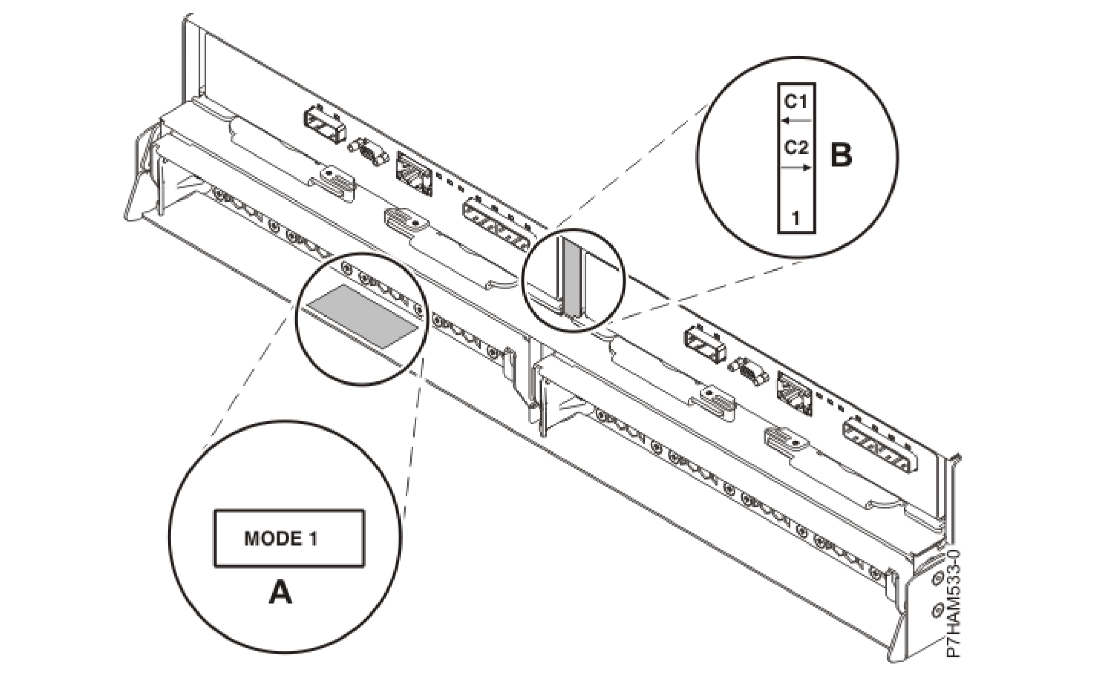


Figure 2-32 Mode sticker locations at the rear of the 5887 disk drive enclosure

The EXP24S SAS ports are attached to a SAS PCIe adapter or pair of adapters using SAS YO or X cables. Cable length varies depending on the feature code, and proper length should be calculated considering routing for proper airflow and ease of handling. A diagram of both types of SAS cables can be seen in Figure 2-33.

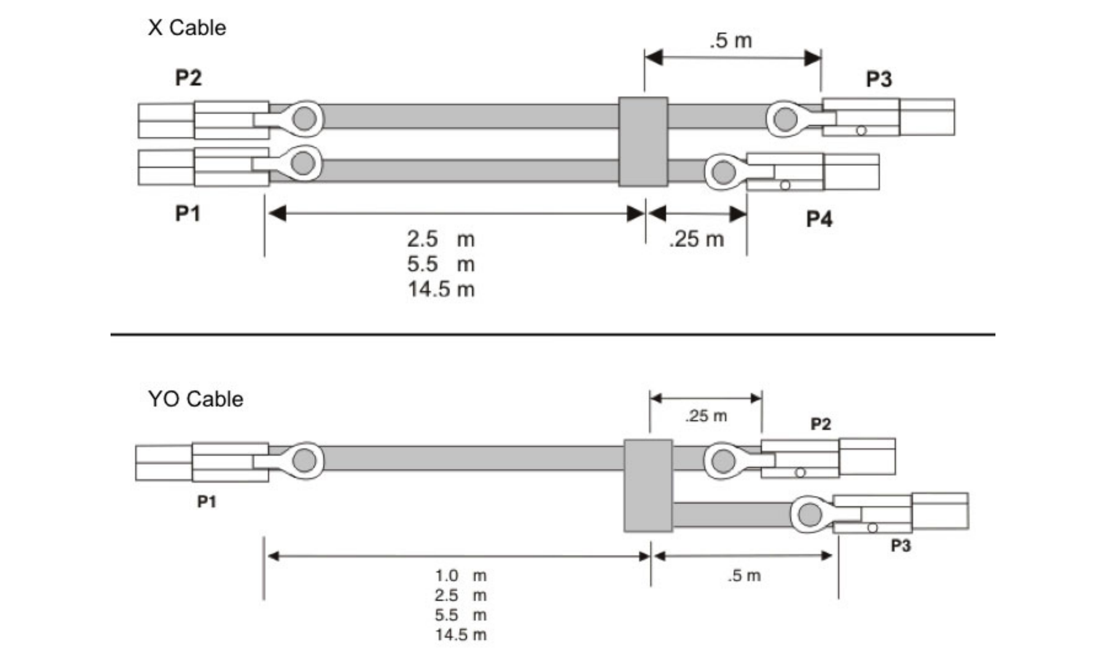


Figure 2-33 Diagram of SAS cables types X and YO

The following SAS adapters support the EXP24S:

- PCIe 380MB Cache Dual-x4 3Gb SAS RAID Adapter (#5805)
- PCIe Dual-x4 SAS Adapter (#5901)
- PCIe2 1.8GB Cache RAID SAS Adapter Tri-port 6Gb (#5913)
- PCIe2 1.8GB Cache RAID SAS Adapter Tri-port 6Gb CR (#ESA3)
- PCIe3 RAID SAS Adapter Quad-port 6Gb x8 (#EJ0J)
- PCIe3 12GB Cache RAID SAS Adapter Quad-port 6Gb x8 (#EJ0L)
- PCIe3 12GB Cache RAID Plus SAS Adapter Quad-port 6Gb x8 (#EJ14)
- PCIe3 LP RAID SAS ADAPTER (#EJ0M)

The EXP24S drawer can support up to 24 SAS SFF Gen-2 disks. Table 2-32 lists the available disk options.

Table 2-32 Available disks for the EXP24S

Feature Code	CCIN	Description	OS support
ES0G		775 GB SFF-2 SSD for AIX/Linux	AIX, Linux
ES0H		775 GB SFF-2 SSD for IBM i	IBM i
ES0Q		387 GB SFF-2 4 K SSD for AIX/Linux	AIX, Linux
ES0R		387 GB SFF-2 4 K SSD for IBM i	IBM i
ES0S		775 GB SFF-2 4 K SSD for AIX/Linux	AIX, Linux
ES0T		775 GB SFF-2 4 K SSD for IBM i	IBM i
ES19		387 GB SFF-2 SSD for AIX/Linux	AIX, Linux
ES1A		387 GB SFF-2 SSD for IBM i	IBM i
ES2C		387 GB SFF-2 SSD for AIX/Linux	AIX, Linux
ES2D		387 GB SFF-2 SSD for IBM i	IBM i
ES78		387 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	AIX, Linux
ES79		387 GB SFF-2 SSD 5xx eMLC4 for IBM i	IBM i
ES7E		775 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	AIX, Linux
ES7F		775 GB SFF-2 SSD 5xx eMLC4 for IBM i	IBM i
ES80		1.9 TB Read Intensive SAS 4k SFF-2 SSD for AIX/Linux	AIX, Linux
ES81		1.9 TB Read Intensive SAS 4k SFF-2 SSD for IBM i	IBM i
ES85		387 GB SFF-2 SSD 4k eMLC4 for AIX/Linux	AIX, Linux
ES86		387 GB SFF-2 SSD 4k eMLC4 for IBM i	IBM i
ES8C		775 GB SFF-2 SSD 4k eMLC4 for AIX/Linux	AIX, Linux
ES8D		775 GB SFF-2 SSD 4k eMLC4 for IBM i	IBM i
ES8F		1.55 TB SFF-2 SSD 4k eMLC4 for AIX/Linux	AIX, Linux
ES8G		1.55 TB SFF-2 SSD 4k eMLC4 for IBM i	IBM i

Feature Code	CCIN	Description	OS support
1948	19B1	283 GB 15 K RPM SAS SFF-2 Disk Drive (IBM i)	IBM i
1953		300 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX, Linux
1962	19B3	571 GB 10 K RPM SAS SFF-2 Disk Drive (IBM i)	IBM i
1964		600 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX, Linux
ES62		3.86-4.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)	AIX, Linux
ES64		7.72-8.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)	AIX, Linux
ESD2	59CD	1.1 TB 10 K RPM SAS SFF-2 Disk Drive (IBM i)	IBM i
ESD3	0	1.2 TB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX, Linux
ESDN		571 GB 15 K RPM SAS SFF-2 Disk Drive - 528 Block (IBM i)	IBM i
ESDP		600 GB 15 K RPM SAS SFF-2 Disk Drive - 5xx Block (AIX/Linux)	AIX, Linux
ESEU	59D2	571 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224	IBM i
ESEV	0	600 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX, Linux
ESEY	0	283 GB 15 K RPM SAS SFF-2 4 K Block - 4224 Disk Drive	IBM i
ESEZ	0	300 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	AIX, Linux
ESF2	58DA	1.1 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224	IBM i
ESF3	0	1.2 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX, Linux
ESFN	0	571 GB 15 K RPM SAS SFF-2 4 K Block - 4224 Disk Drive	IBM i
ESFP	0	600 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	AIX, Linux
ESFS	0	1.7 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224	IBM i
ESFT	0	1.8 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX, Linux

There are six SAS connectors on the rear of the EXP24S drawer to which two SAS adapters or controllers are attached. They are labeled T1, T2, and T3; there are two T1, two T2, and two T3 connectors. While configuring the drawer, special configuration feature codes will indicate for the plant the mode of operation in which the disks and ports will be split:

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

Figure 2-34 shows the rear connectors of the EXP24S drawer, how they relate with the modes of operation, and disk grouping.

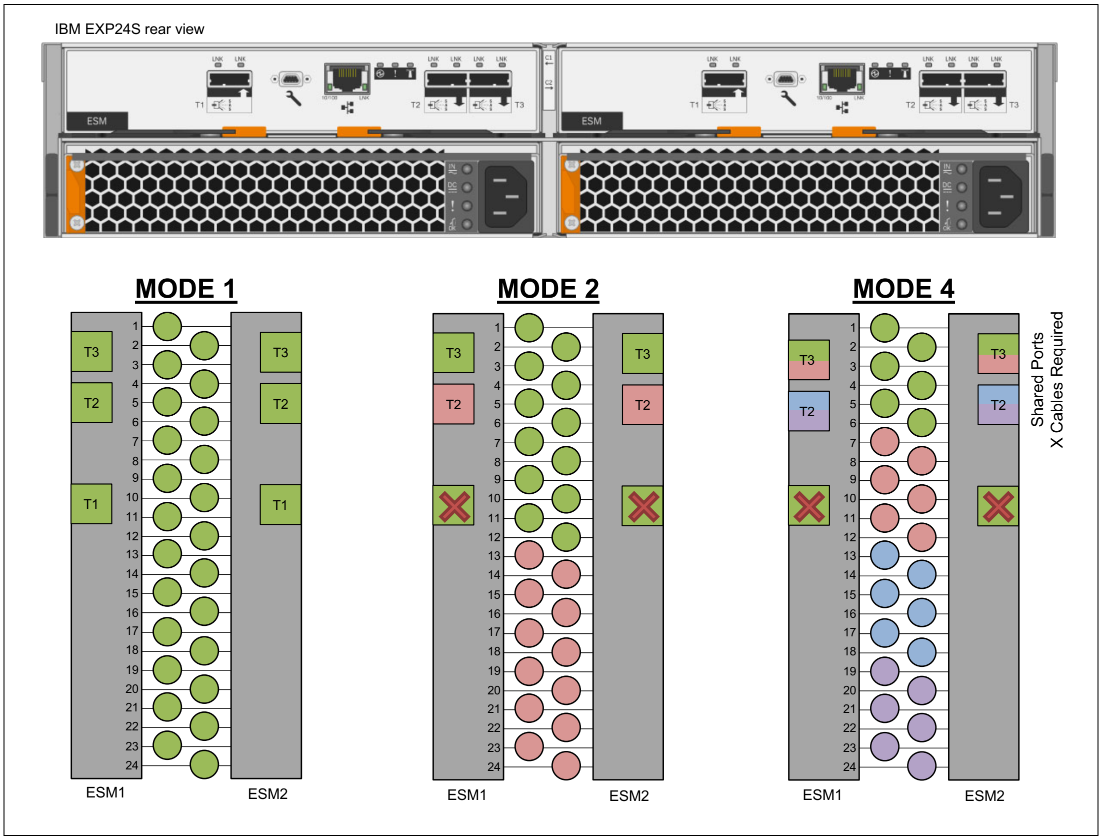


Figure 2-34 Rear view of EXP24S with the 3 modes of operation and the disks assigned to each port

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

The most common configurations for EXP24S with Power Systems are detailed in 2.10.2, “EXP24S common usage scenarios” on page 97. Not all possible scenarios are included. For more information about SAS cabling and cabling configurations, search “Planning for serial-attached SCSI cables” in the IBM Knowledge Center, which can be accessed at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm?cp=9119-MME

2.10.2 EXP24S common usage scenarios

The EXP24S drawer is very versatile in the ways that it can be attached to Power Systems. This section describes the most common usage scenarios for EXP24S and Virtual I/O Servers, using standard PCIe SAS adapters #5901.

Note: Not all possible scenarios are included. Refer to the “Planning for serial-attached SCSI cables” guide in the IBM Knowledge Center to see more supported scenarios.

Scenario 1: Basic non-redundant connection

This scenario assumes a single Virtual I/O Server with a single PCIe SAS adapter #5901 and an EXP24S set on mode 1, allowing for up to 24 disks to be attached to the server. Figure 2-35 shows the connection diagram and components of the solution.

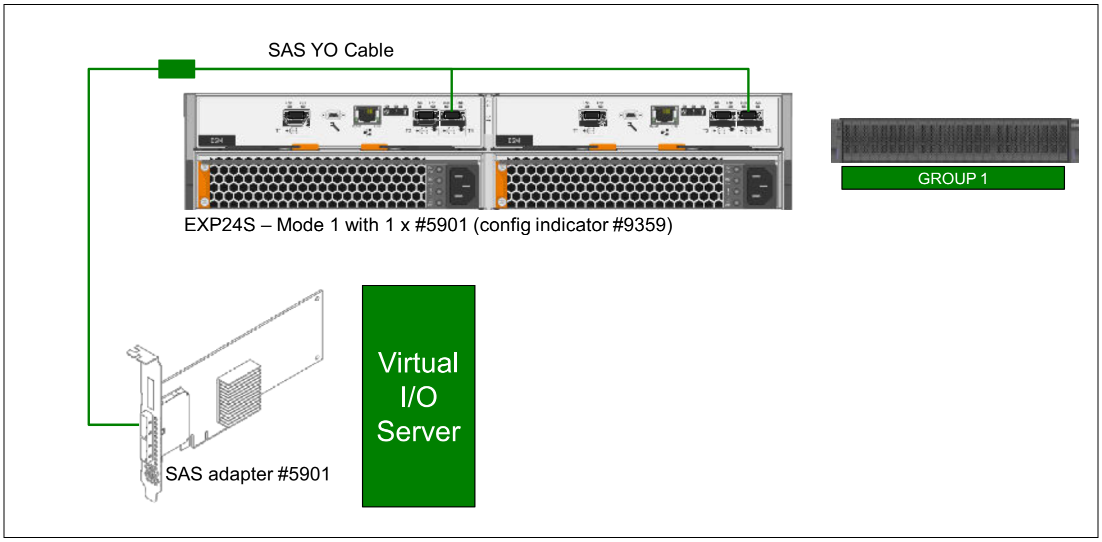


Figure 2-35 Scenario 1 - Basic non-redundant connection

For this scenario, these are the required feature codes:

- One EXP24S drawer #5887 with indicator feature #9359 (mode 1 with single #5901)
- One PCIe SAS adapter #5901
- One SAS YO cable 3 Gbps with proper length

Scenario 2: Basic redundant connection

This scenario assumes a single Virtual I/O Server with two PCIe SAS adapters #5901 and an EXP24S set on mode 1, allowing for up to 24 disks to be attached to the server. Figure 2-36 shows the connection diagram and components of the solution.

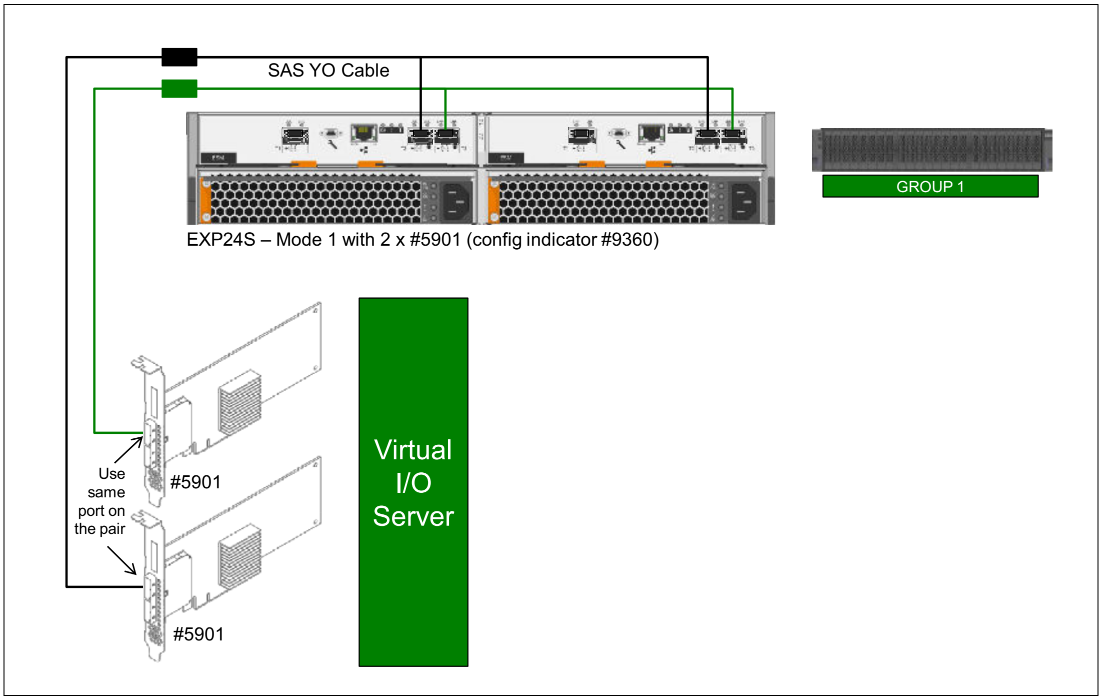

SAS YO Cable
GROUP 1
EXP24S – Mode 1 with 2 x #5901 (config indicator #9360)
Use same port on the pair
#5901
#5901
Virtual I/O Server

Figure 2-36 Scenario 2- Basic redundant connection

For this scenario, these are the required feature codes:

- One EXP24S drawer #5887 with indicator feature #9360 (mode 1 with dual #5901)
- Two PCIe SAS adapter #5901
- Two SAS YO cables 3 Gbps with proper length

The ports used on the SAS adapters must be the same for both adapters of the pair. There is no SSD support on this scenario.

Scenario 3: Dual Virtual I/O Servers sharing a single EXP24S

This scenario assumes a dual Virtual I/O Server with two PCIe SAS adapters #5901 each and an EXP24S set on mode 2, allowing for up to 12 disks to be attached to the each Virtual I/O Server. Figure 2-37 shows the connection diagram and components of the solution.

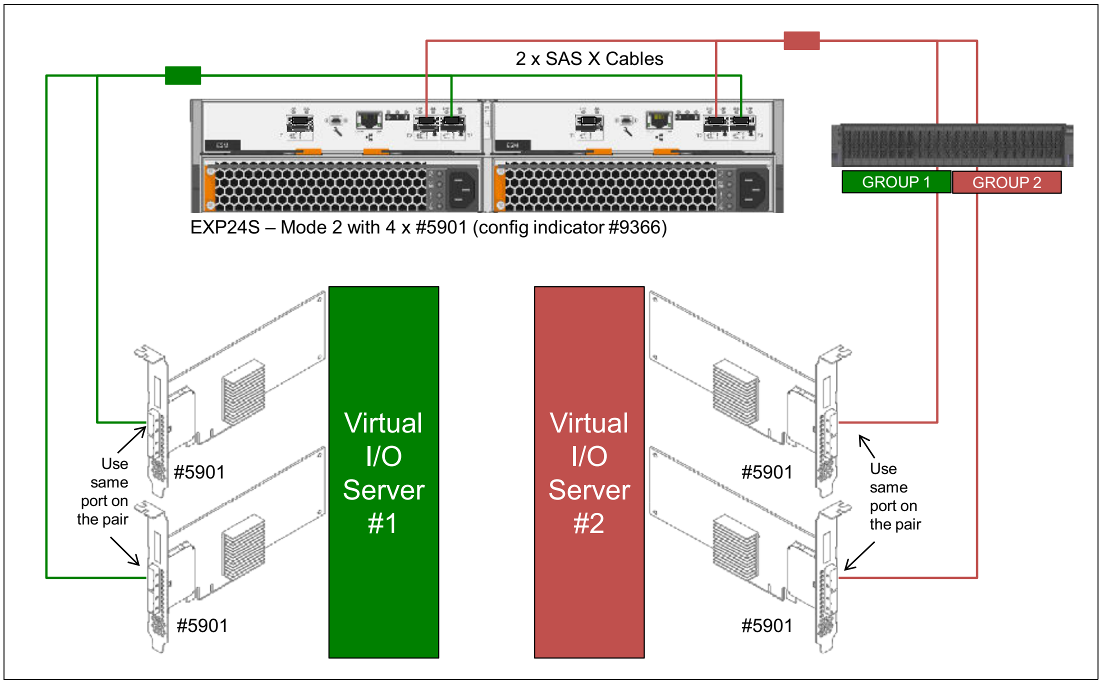

2 x SAS X Cables
GROUP 1
GROUP 2
EXP24S – Mode 2 with 4 x #5901 (config indicator #9366)
Use same port on the pair
#5901
#5901
Virtual I/O Server #1
Virtual I/O Server #2
#5901
#5901
Use same port on the pair

Figure 2-37 Dual Virtual I/O Servers sharing a single EXP24S

For this scenario, these are the required feature codes:

- One EXP24S drawer #5887 with indicator feature #9366 (mode 2 with quad #5901)
- Four PCIe SAS adapter #5901
- Two SAS X cables 3 Gbps with proper length

The ports used on the SAS adapters must be the same for both adapters of the pair. There is no SSD support on this scenario.

Scenario 4: Dual Virtual I/O Servers sharing two EXP24S

This scenario assumes a dual Virtual I/O Server with two PCIe SAS adapters #5901 each and two EXP24S set on mode 2, allowing for up to 24 disks to be attached to the each Virtual I/O Server (2 per drawer). If compared to scenario 3, this scenario has the benefit to allow disks from different EXP24S drawers to be mirrored, allowing for hot maintenance of the whole EXP24S drawers if all data is properly mirrored. Figure 2-38 shows the connection diagram and components of the solution.

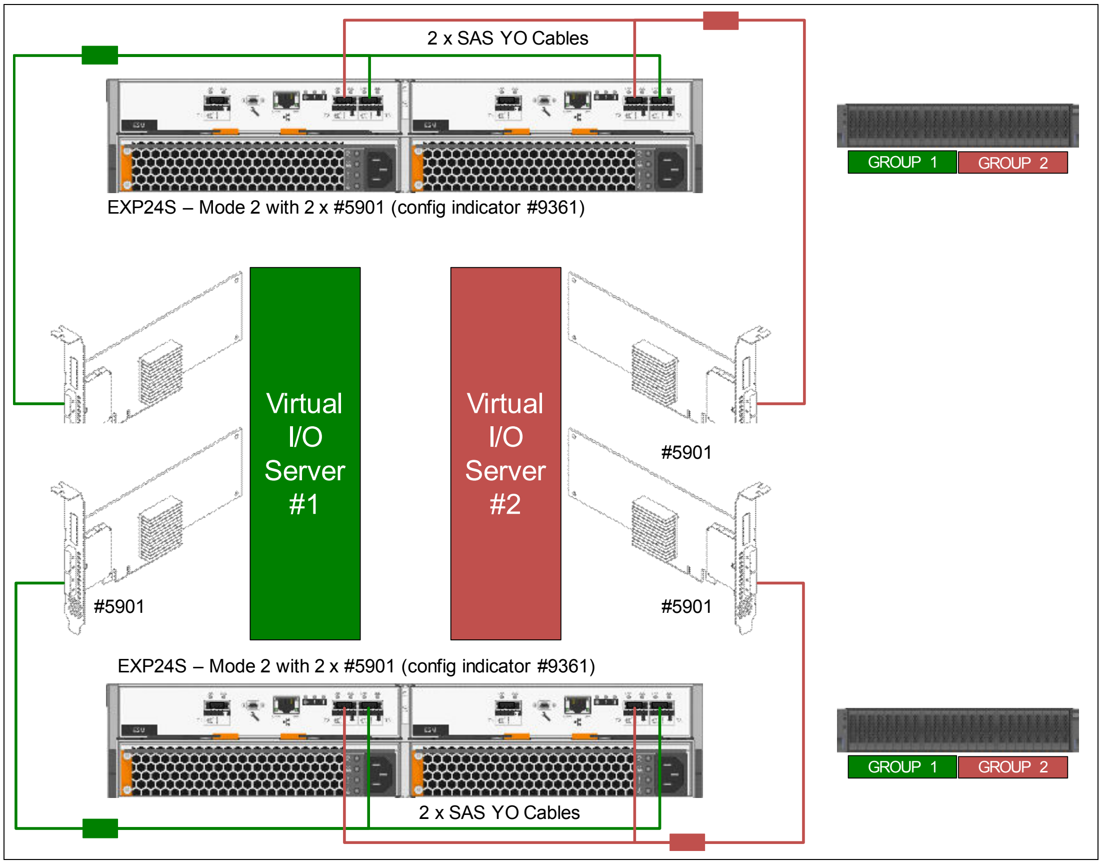


Figure 2-38 Dual Virtual I/O Servers sharing two EXP24S

For this scenario, these are the required feature codes:

- Two EXP24S drawers #5887 with indicator feature #9361 (mode 2 with dual #5901)
- Four PCIe SAS adapter #5901
- Four SAS YO cables 3 Gbps with proper length.

There is no SSD support on this scenario.

Scenario 5: Four Virtual I/O Servers sharing two EXP24S

This scenario assumes four Virtual I/O Servers with two PCIe SAS adapters #5901 each and two EXP24S set on mode 4, allowing for up to 12 disks to be attached to the each Virtual I/O Server (6 per drawer). This scenario has the benefit to allow disks from different EXP24S drawers to be mirrored, allowing for hot maintenance of the whole EXP24S drawers if all data is properly mirrored. Figure 2-39 shows the connection diagram and components of the solution.

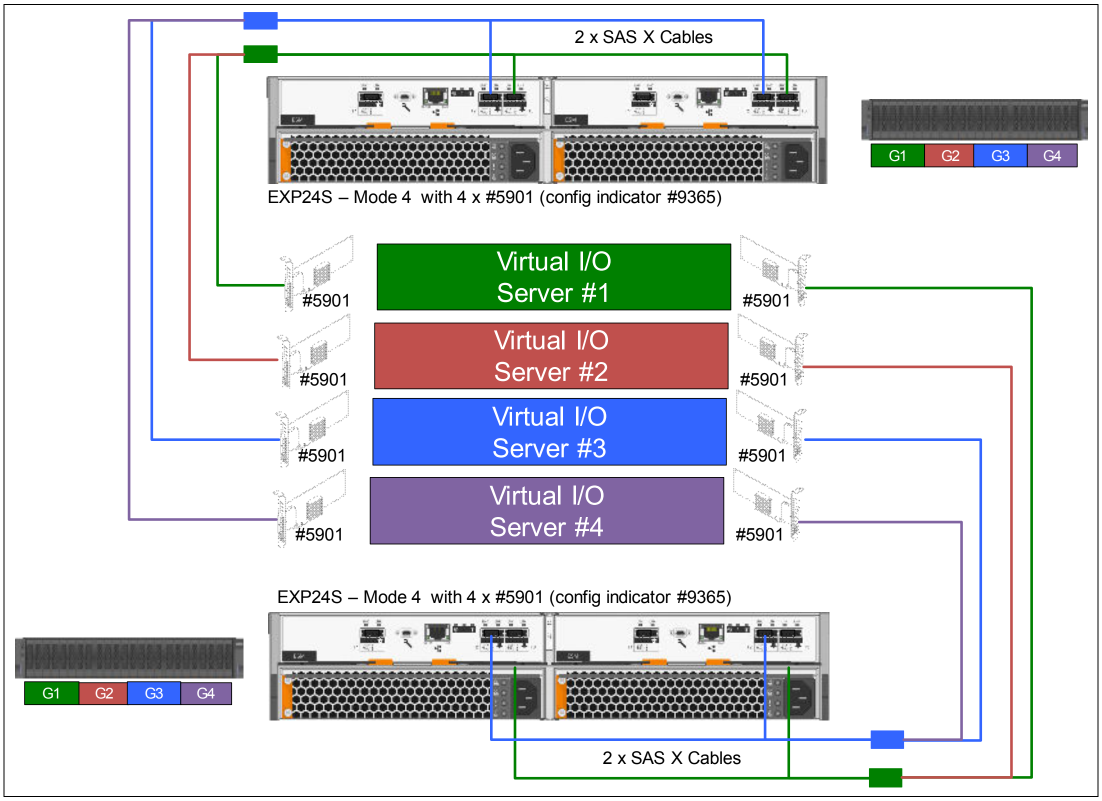


Figure 2-39 Four Virtual I/O Servers sharing two EXP24S

For this scenario, these are the required feature codes:

- Two EXP24S drawers #5887 with indicator feature #9365 (mode 4 with four #5901)
- Eight PCIe SAS adapter #5901
- Four SAS X cables 3 Gbps with proper length.

There is no SSD support on this scenario.

Other scenarios

For direct connection to logical partitions, different adapters, and cables, see “5887 disk drive enclosure” in the IBM Knowledge Center:

http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

2.10.3 IBM System Storage

The IBM System Storage Disk Systems products and offerings provide compelling storage solutions with superior value for all levels of business, from entry-level to high-end storage systems. For more information about the various offerings, see the following website:

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

The following section highlights a few of the offerings.

IBM Network-Attached Storage

IBM Network-Attached Storage (NAS) products provide a wide range of network attachment capabilities to a broad range of host and client systems, such as IBM Scale Out Network Attached Storage and the IBM System Storage Nxxx series. For more information about the hardware and software, see the following website:

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

IBM Storwize family

The IBM Storwize® family is the ideal solution to optimize the data architecture for business flexibility and data storage efficiency. Different models, such as the Storwize V3700, V5000, and V7000, offer storage virtualization, IBM Real-time Compression™, IBM Easy Tier®, and many other functions. For more information, see the following website:

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

IBM Flash Storage

IBM Flash Storage delivers extreme performance to derive measurable economic value across the data architecture (servers, software, applications, and storage). IBM offers a comprehensive flash portfolio with the IBM FlashSystem™ family. For more information, see the following website:

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

IBM XIV Storage System

IBM XIV® is a high-end disk storage system, helping thousands of enterprises meet the challenge of data growth with hotspot-free performance and ease of use. Simple scaling, high service levels for dynamic, heterogeneous workloads, and tight integration with hypervisors and the OpenStack platform enable optimal storage agility for cloud environments.

XIV extends ease of use with integrated management for large and multi-site XIV deployments, reducing operational complexity and enhancing capacity planning. For more information, see the following website:

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

IBM System Storage DS8000

The IBM System Storage DS8000 is a high-performance, high-capacity, and secure storage system that is designed to deliver the highest levels of performance, flexibility, scalability, resiliency, and total overall value for the most demanding, heterogeneous storage environments. The system is designed to manage a broad scope of storage workloads that exist in today’s complex data center, doing it effectively and efficiently.

Additionally, the IBM System Storage DS8000 includes a range of features that automate performance optimization and application quality of service, and also provide the highest levels of reliability and system uptime. For more information, see the following website:

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

2.11 Hardware Management Console

The Hardware Management Console (HMC) is a dedicated appliance that allows administrators to configure and manage system resources on IBM Power Systems servers that use IBM POWER6, POWER6+ POWER7, POWER7+, and POWER8 processors and the PowerVM hypervisor. The HMC provides basic virtualization management support for configuring logical partitions (LPARs) and dynamic resource allocation, including processor and memory settings for selected Power Systems servers. The HMC also supports advanced service functions, including guided repair and verification, concurrent firmware updates for managed systems, and around-the-clock error reporting through IBM Electronic Service Agent™ for faster support.

The HMC management features help improve server usage, simplify systems management, and accelerate provisioning of server resources by using the PowerVM virtualization technology.

Requirements:

- When using the HMC with the Power E870 and Power E880 servers with a maximum of two system nodes, the HMC code must be running at V8R8.2.0 level, or later.
- When using the HMC with the Power E880 servers with a 12-core chip option or a maximum of three or four system nodes, the HMC code must be running at V8R8.3.0 level, or later.

The Power E870 and Power E880 platforms support two main service environments:

- Attachment to one or more HMCs. This environment is the common configuration for servers supporting logical partitions with dedicated or virtual I/O. In this case, all servers have at least one logical partition.
- Hardware support for customer-replaceable units. This support comes standard along with the HMC. In addition, users can upgrade this support level to IBM onsite support to be consistent with other Power Systems servers.

2.11.1 HMC code level

HMC V8R8.3.0 contains the following new features:

- Support for Power E880 servers with the new 48-core system node option (#EPBD)
- Support for Power E880 servers with the new option for a third and fourth I/O drawer
- Support for Power E850 servers
- Service Management:

  New access points for Call Home functionality in Europe

- Virtualization Management:
  - Trial / temporary Enterprise Enablement support
  - 52 sessions for local 5250 consoles (currently limited to 26)
- Console Management:

- Full release of the new enhanced user interface
- Browser currency
- Improved log retention (through file system resizing, rotation changes & content reduction)
- Call Home support for modem (dial-in via AT&T Global Network) and VPN will be removed

If you are attaching an HMC to a new server or adding a function to an existing server that requires a firmware update, the HMC machine code might need to be updated to support the firmware level of the server. In a dual HMC configuration, both HMCs must be at the same version and release of the HMC code.

To determine the HMC machine code level that is required for the firmware level on any server, go to the following website to access the Fix Level Recommendation Tool (FLRT) on or after the planned availability date for this product:

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

FLRT identifies the correct HMC machine code for the selected system firmware level.

Note: Access to firmware and machine code updates is conditional on entitlement and license validation in accordance with IBM policy and practice. IBM may verify entitlement through customer number, serial number electronic restrictions, or any other means or methods that are employed by IBM at its discretion.

2.11.2 HMC RAID 1 support

HMCs now offer a high availability feature. The 7042-CR9, by default, includes two HDDs with RAID 1 configured. RAID 1 is also offered on the 7042-CR6, 7042-CR7, 7042-CR8 and 7042-CR9 models (if the feature was removed from the initial order) as an MES upgrade option.

RAID 1 uses data mirroring. Two physical drives are combined into an array, and the same data is written to both drives. This makes the drives mirror images of each other. If one of the drives experiences a failure, it is taken offline and the HMC continues operating with the other drive.

HMC models

To use an existing HMC to manage any POWER8 processor-based server, the HMC must be a model CR5, or later, rack-mounted HMC, or model C08, or later, deskside HMC. The latest HMC model is the 7042-CR9. For your reference, Table 2-33 lists a comparison between the 7042-CR8 and the 7042-CR9 HMC models.

Note: The 7042-CR9 ships with 16 GB of memory, and is expandable to 192 GB with an upgrade feature. 16 GB is advised for large environments or where external utilities, such as PowerVC and other third party monitors, are to be implemented.

Table 2-33 Comparison between 7042-CR7 and 7042-CR8 models

Feature	CR8	CR9
IBM System x model	x3550 M4 7914 PCH	x3550 M5 5463 AC1
HMC model	7042-CR8	7042-CR9

Feature	CR8	CR9
Processor	Intel 8-Core Xeon v2 2.00 GHz	Intel 18-core Xeon v3 2.4 GHz
Memory max:	16 GB (when featured)	16 GB DDR4 expandable to 192 GB
DASD	500 GB	500 GB
RAID 1	Default	Default
USB ports	Two front, four back	Two front, four rear
Integrated network	Four 1 Gb Ethernet	Four 1 Gb Ethernet
I/O slots	One PCI Express 3.0 slot	One PCI Express 3.0 slot

2.11.3 HMC connectivity to the POWER8 processor-based systems

POWER8 processor-based servers, and their predecessor systems, that are managed by an HMC require Ethernet connectivity between the HMC and the server's service processor. In addition, if dynamic LPAR, Live Partition Mobility, or PowerVM Active Memory Sharing operations are required on the managed partitions, Ethernet connectivity is needed between these partitions and the HMC. A minimum of two Ethernet ports are needed on the HMC to provide such connectivity.

For the HMC to communicate properly with the managed server, eth0 of the HMC must be connected to either the HMC1 or HMC2 ports of the managed server, although other network configurations are possible. You may attach a second HMC to the remaining HMC port of the server for redundancy. The two HMC ports must be addressed by two separate subnets.

Figure 2-40 shows a simple network configuration to enable the connection from the HMC to the server and to allow for dynamic LPAR operations. For more information about HMC and the possible network connections, see *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491.

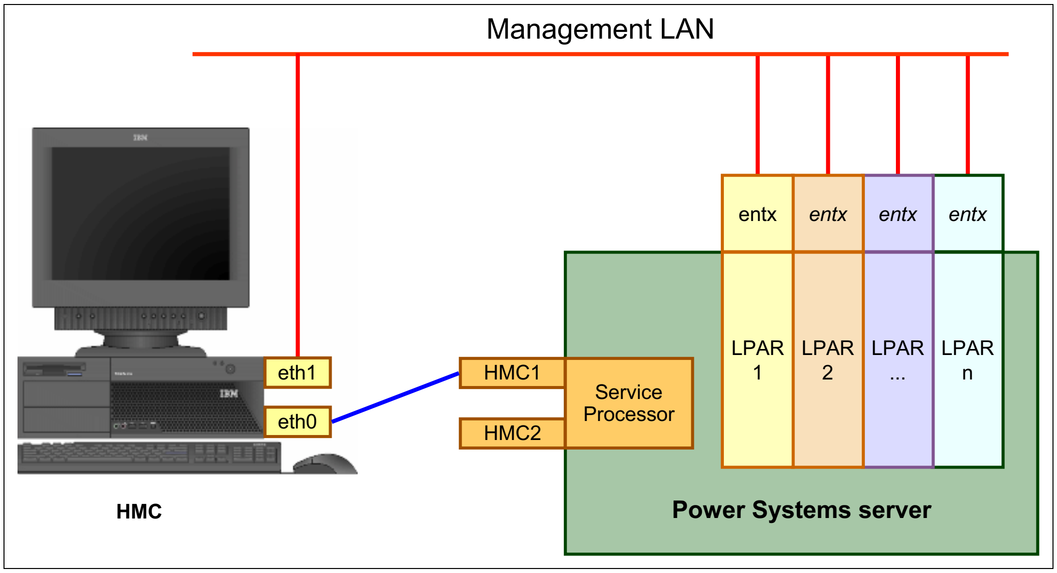


Figure 2-40 Network connections from the HMC to service processor and LPARs

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

If the service processor of the managed server does not receive a DHCP reply before timeout, predefined IP addresses are set up on both ports. Static IP address allocation is also an option and can be configured using the ASMI menus.

Notes: The two service processor HMC ports have the following features:

- Run at a speed of 1 Gbps
- Are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI options from a client directly from a client web browser
- Use the following network configuration if no IP addresses are set:
  - Service processor eth0 (HMC1 port): 169.254.2.147 with netmask 255.255.255.0
  - Service processor eth1 (HMC2 port): 169.254.3.147 with netmask 255.255.255.0

For more information about the service processor, see 2.5.2, “Service Processor Bus” on page 74.

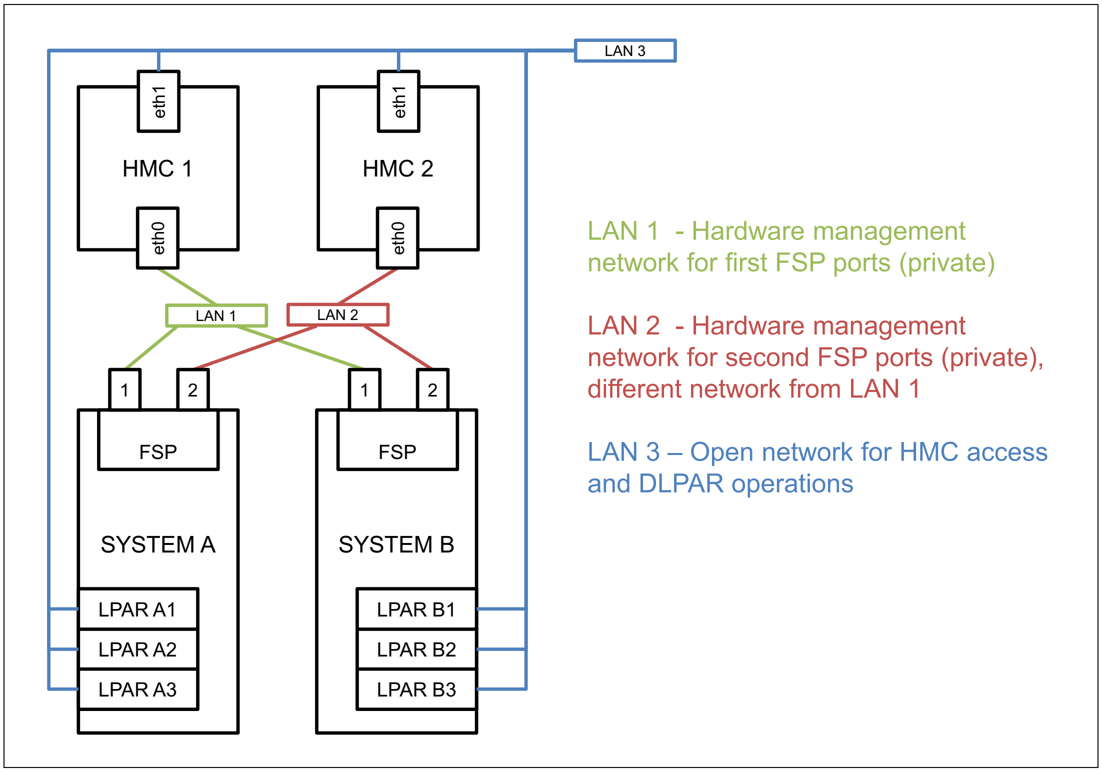
2.11.4 High availability HMC configuration

The HMC is an important hardware component. Although Power Systems servers and their hosted partitions can continue to operate when the managing HMC becomes unavailable, certain operations, such as dynamic LPAR, partition migration using PowerVM Live Partition Mobility, or the creation of a new partition, cannot be performed without the HMC. To avoid such situations, consider installing a second HMC, in a redundant configuration, to be available when the other is not (during maintenance, for example).

To achieve HMC redundancy for a POWER8 processor-based server, the server must be connected to two HMCs:

- The HMCs must be running the same level of HMC code.
- The HMCs must use different subnets to connect to the service processor.
- The HMCs must be able to communicate with the server’s partitions over a public network to allow for full synchronization and functionality.

Figure 2-41 shows one possible highly available HMC configuration that is managing two servers. Each HMC is connected to one FSP port of each managed server.



LAN 3
eth1
HMC 1
eth0
eth1
HMC 2
eth0
LAN 1
LAN 2
1
2
1
2
FSP
FSP
SYSTEM A
SYSTEM B
LPAR A1
LPAR A2
LPAR A3
LPAR B1
LPAR B2
LPAR B3
LAN 1 - Hardware management network for first FSP ports (private)
LAN 2 - Hardware management network for second FSP ports (private), different network from LAN 1
LAN 3 – Open network for HMC access and DLPAR operations

Figure 2-41 Highly available HMC networking example

For simplicity, only the hardware management networks (LAN1 and LAN2) are highly available (Figure 2-41). However, the open network (LAN3) can be made highly available by using a similar concept and adding a second network between the partitions and HMCs.

For more information about redundant HMCs, see *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491.

2.12 Operating system support

The IBM Power E870 and Power E880 systems support the following operating systems:

- AIX
- IBM i
- Linux

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

For details about the software available on IBM Power Systems, visit the IBM Power Systems Software™ website:

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

2.12.1 Virtual I/O Server

The minimum required levels of Virtual I/O Server for both the Power E870 and Power E880 are:

- VIOS 2.2.3.4 with ifix IV63331
- VIOS 2.2.2.6

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

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

2.12.2 IBM AIX operating system

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

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

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

The Fix Central website also provides information about how to obtain the fixes that are included on CD-ROM.

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

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

IBM AIX Version 6.1

A partition that uses AIX 6.1 can run in POWER6, POWER6+, or POWER7 mode. This will limit the partition to SMT-4 among other hardware capabilities.

The minimum level of AIX Version 6.1 supported on the Power E870 and Power E880 depends on the partition having 100% virtualized resources or not.

For partitions that have all of their resources virtualized via Virtual I/O Server, the minimum levels of AIX Version 6.1 supported on the Power E870 and Power E880 are as follows:

- AIX Version 6.1 with the 6100-08 Technology Level and Service Pack 1 or later
- AIX Version 6.1 with the 6100-09 Technology Level and Service Pack 1 or later

For all other partitions, the minimum levels of AIX Version 6.1 supported on the Power E870 and Power E880 are as follows:

- AIX Version 6.1 with the 6100-08 Technology Level and Service Pack 6, or later
- AIX Version 6.1 with the 6100-09 Technology Level and Service Pack 4, and APAR IV63331, or later

IBM AIX Version 7.1

A partition that uses AIX 7.1 can run in POWER6, POWER6+, POWER7, or POWER8 mode. This allows for an easier migration from previous systems and full exploitation of the POWER8 features.

The minimum level of AIX Version 7.1 supported on the Power E870 and Power E880 depends on the partition having 100% virtualized resources or not.

For partitions that have all of their resources virtualized via Virtual I/O Server, the minimum levels of AIX Version 7.1 supported on the Power E870 and Power E880 are as follows:

- AIX Version 7.1 with the 7100-02 Technology Level and Service Pack 1 or later
- AIX Version 7.1 with the 7100-03 Technology Level and Service Pack 1 or later

For all other partitions, the minimum levels of AIX Version 7.1 supported on the Power E870 and Power E880 are as follows:

- AIX Version 7.1 with the 7100-02 Technology Level Service Pack 6, or later
- AIX Version 7.1 with the 7100-03 Technology Level Service Pack 4, and APARs IV63332 and IV69116, or later

2.12.3 IBM i operating system

The IBM i operating system is supported on the Power E870 and Power E880 with the following minimum required levels:

- IBM i 7.1 TR9 or later
- IBM i 7.2 TR1 or later

> **Note:** The minimum supported levels for the Power E880 with the new processor or I/O drawer options are:
>
> - IBM i 7.1 TR10
> - IBM i 7.2 TR2

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

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

Visit the IBM Prerequisite website for compatibility information for hardware features and the corresponding AIX and IBM i Technology Levels:

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

2.12.4 Linux operating systems

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

The supported versions of Linux on Power E870 and Power E880 are as follows:

- Big endian:
  - SUSE Linux Enterprise Server 11 Service Pack 3, or later
  - Red Hat Enterprise Linux 6.5 for POWER, or later

- Little endian:
  - Red Hat Enterprise Linux 7.1, or later
  - SUSE Linux Enterprise Server 12 and later Service Packs
  - Ubuntu 15.04

If you want to configure Linux partitions in virtualized Power Systems, be aware of the following conditions:

- Not all devices and features that are supported by the AIX operating system are supported in logical partitions running the Linux operating system.
- Linux operating system licenses are ordered separately from the hardware. You can acquire Linux operating system licenses from IBM to be included with the POWER8 processor-based servers, or from other Linux distributors.

For information about features and external devices that are supported by Linux, see this site:

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

For information about SUSE Linux Enterprise Server, see this site:

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

For information about Red Hat Enterprise Linux Advanced Server, see this site:

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

2.12.5 Java versions that are supported

Java is supported on POWER8 servers. For best exploitation of the performance capabilities and most recent improvements of POWER8 technology, upgrade Java-based applications to Java 7 or Java 6. For more information, visit these websites:

http://www.ibm.com/developerworks/java/jdk/aix/service.html
http://www.ibm.com/developerworks/java/jdk/linux/download.html

2.12.6 Boosting performance and productivity with IBM compilers

IBM XL C, XL C/C++ and XL Fortran compilers for AIX and for Linux exploit the latest POWER8 processor architecture. Release after release, these compilers continue to deliver application performance improvements and additional capability, exploiting architectural enhancements made available through the advancement of the POWER technology.

IBM compilers are designed to optimize and tune your applications for execution on IBM POWER platforms. Compilers help you unleash the full power of your IT investment. With the XL compilers you can create and maintain critical business and scientific applications, while maximizing application performance and improving developer productivity. The performance gain from years of compiler optimization experience is seen in the continuous release-to-release compiler improvements that support the POWER and POWERPC families of processors. XL compilers support POWER4, POWER4+, POWER5, POWER5+, POWER6, POWER7, and POWER7+ processors, and now add support for the new POWER8 processors. With the support of the latest POWER8 processor chip, IBM advances a more than 20-year investment in the XL compilers for POWER series and IBM PowerPC® series architectures.

XL C, XL C/C++ and XL Fortran features introduced to exploit the latest POWER8 processor include vector unit and vector scalar extension (VSX) instruction set to efficiently manipulate vector operations in your application, vector functions within the Mathematical Acceleration

Subsystem (MASS) libraries for improved application performance, built-in functions or intrinsics and directives for direct control of POWER instructions at the application level, and architecture and tune compiler options to optimize and tune your applications.

XL compilers support application development on big endian distributions. XL C/C++ for Linux, V13.1.1 and XL Fortran for Linux, V15.1.1 deliver new compilers that support application development on the IBM POWER8 servers that run the little endian Linux distributions. With these two releases, compiler support on the Linux distributions Ubuntu 14.04 for IBM POWER8, Ubuntu 14.10 for IBM POWER8, and SUSE Linux Enterprise Server 12 for Power, includes exploitation of the little endian architecture on the POWER8 processor.

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

The performance of IBM PL/I for AIX applications has been improved through both front-end changes and back-end optimizer enhancements. The back-end optimizer, a component common also to the IBM XL compilers, lets your applications leverage the latest industry-leading optimization technology. The PL/I compiler produces code that is intended to perform well across all hardware levels on AIX.

IBM Rational® Developer for AIX and Linux, C/C++ Edition provides a rich set of integrated development tools that support XL C for AIX, XL C/C++ for AIX, and XL C/C++ for Linux compiler. It also supports the GNU compiler and debugger on Linux on x86 architectures to make it possible to do development on other infrastructures and then easily port and optimize the resultant workloads to run on POWER and fully exploit the POWER platform's unique qualities of service. Tool capabilities include file management, searching, smart assistive editing, application analysis, unit test automation, code coverage analysis, a unique expert system Migration and Porting Assistant, a unique expert system Performance Advisor, local build, and cross-language/cross-platform debugger, all integrated into an Eclipse workbench. This solution can greatly improve developers' productivity and initial code quality with resultant benefits to downstream disciplines such as QA and Operations.

IBM Rational Developer for AIX and Linux, AIX COBOL Edition provides a rich set of integrated development tools that support the COBOL for AIX compiler. Capabilities include file management, searching, smart assistive editing, application analysis, local build, and cross-language/cross-platform debugger, all integrated into an Eclipse workbench. This solution can boost developers' productivity by moving from older, text-based, command-line development tools to a rich set of integrated development tools and unlike competing distributed COBOL IDEs, is not dependent upon an expensive companion COBOL runtime environment.

IBM Rational Development Studio for IBM i 7.2 provides programming languages for creating modern business applications such as these:

- The ILE RPG, ILE COBOL, C, and C++ compilers
- The heritage RPG and COBOL compilers
- Application Development ToolSet (ADTS), the text-based development tools (such as SEU and PDM)

The latest release includes support for free-form RPG and better Unicode handling and continues to support Open Access: RPG Edition as an included technology.

IBM Rational Developer for i provides a rich set of integrated development tools that support the IBM i ILE compilers. Tool capabilities include file management, searching, smart assistive editing, application analysis, code coverage analysis, and cross-language/cross-platform debugger, all integrated into an Eclipse workbench. This solution can greatly improve developers' productivity as compared to using text-based and command-line development tools, and is more appealing than such tools when you need to recruit new IBM i development talent.

2.13 Energy management

The Power E870 and Power E880 systems have features to help clients become more energy efficient. EnergyScale technology enables advanced energy management features to conserve power dramatically and dynamically and further improve energy efficiency. Intelligent Energy optimization capabilities enable the POWER8 processor to operate at a higher frequency for increased performance and performance per watt, or dramatically reduce frequency to save energy.

2.13.1 IBM EnergyScale technology

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

EnergyScale uses power and thermal information that is collected from the system to implement policies that can lead to better performance or better energy usage. IBM EnergyScale has the following features:

- Power trending

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

- Power saver mode

  Power saver mode lowers the processor frequency and voltage on a fixed amount, reducing the energy consumption of the system while still delivering predictable performance. This percentage is predetermined to be within a safe operating limit and is not user configurable. The server is designed for a fixed frequency drop of almost 50% down from nominal frequency (the actual value depends on the server type and configuration).

  Power saver mode is not supported during system start, although it is a persistent condition that is sustained after the boot when the system starts running instructions.

- Dynamic power saver mode

  Dynamic power saver mode varies processor frequency and voltage based on the usage of the POWER8 processors. Processor frequency and usage are inversely proportional for most workloads, implying that as the frequency of a processor increases, its usage decreases, given a constant workload. Dynamic power saver mode takes advantage of this relationship to detect opportunities to save power, based on measured real-time system usage.

  When a system is idle, the system firmware lowers the frequency and voltage to power energy saver mode values. When fully 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 is designed to reduce the maximum frequency to about 95% of nominal values. If performance is favored over energy consumption, the maximum frequency can be increased above the nominal frequency for extra performance. Table 2-34 shows the maximum frequency boost available for different speed processors in the Power E870 and E880.

  Table 2-34 Maximum frequency boosts for Power E870 and E880 processors

System	Cores per chip	Nominal speed	Maximum boost speed
Power E870	8	4.024 GHz	4.123 GHz
Power E870	10	4.190 GHz	4.456 GHz
Power E880	8	4.356 GHz	4.522 GHz
Power E880	12	4.024 GHz	4.256 GHz

  The frequency boost figures are maximums and will depend on the environment where the servers are installed. Maximum boost frequencies may not be reached if the server is installed in higher temperatures or at altitude.

  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 must never be reached but that frees extra power that was never used in the data center. The *margined* power is this amount of extra power that is allocated to a server during its installation in a data center. It is based on the server environmental specifications that usually are never reached because server specifications are always based on maximum configurations and worst-case scenarios.

- Soft power capping

  There are two power ranges into which the power cap can be set: power capping, as described previously, and soft power capping. Soft power capping extends the allowed energy capping range further, beyond a region that can be ensured in all configurations and conditions. If the energy management goal is to meet a particular consumption limit, then soft power capping is the mechanism to use.

- Processor core nap mode

  IBM POWER8 processor uses a low-power mode that is called *nap* that stops processor execution when there is no work to do on that processor core. The latency of exiting nap mode is small, typically not generating any impact on applications running. Therefore, the IBM POWER Hypervisor™ can use nap mode as a general-purpose idle state. When the operating system detects that a processor thread is idle, it yields control of a hardware thread to the POWER Hypervisor. The POWER Hypervisor immediately puts the thread into nap mode. Nap mode allows the hardware to turn off the clock on most of the circuits in 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 and causes a cumulative effect. Nap mode saves 10 - 15% of power consumption in the processor core.

- Processor core sleep mode

  To save even more energy, the POWER8 processor has an even lower power mode referred to as *sleep*. Before a core and its associated private L2 cache enter sleep mode, the cache is flushed, transition lookaside buffers (TLB) are invalidated, and the hardware clock is turned off in the core and in the cache. Voltage is reduced to minimize leakage current. Processor cores that are inactive in the system (such as capacity on demand (CoD) processor cores) are kept in sleep mode. Sleep mode saves about 80% power consumption in the processor core and its associated private L2 cache.

- Processor chip winkle mode

  The most amount of energy can be saved when a whole POWER8 chiplet enters the *winkle* mode. In this mode, the entire chiplet is turned off, including the L3 cache. This mode can save more than 95% power consumption.

- Fan control and altitude input

  System firmware dynamically adjusts fan speed based on energy consumption, altitude, ambient temperature, and energy savings modes. Power Systems are designed to operate in worst-case environments, in hot ambient temperatures, at high altitudes, and with high-power components. In a typical case, one or more of these constraints are not valid. When no power savings setting is enabled, fan speed is based on ambient temperature and assumes a high-altitude environment. When a power savings setting is enforced (either Power Energy Saver Mode or Dynamic Power Saver Mode), the fan speed varies based on power consumption and ambient temperature.

- Processor folding

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

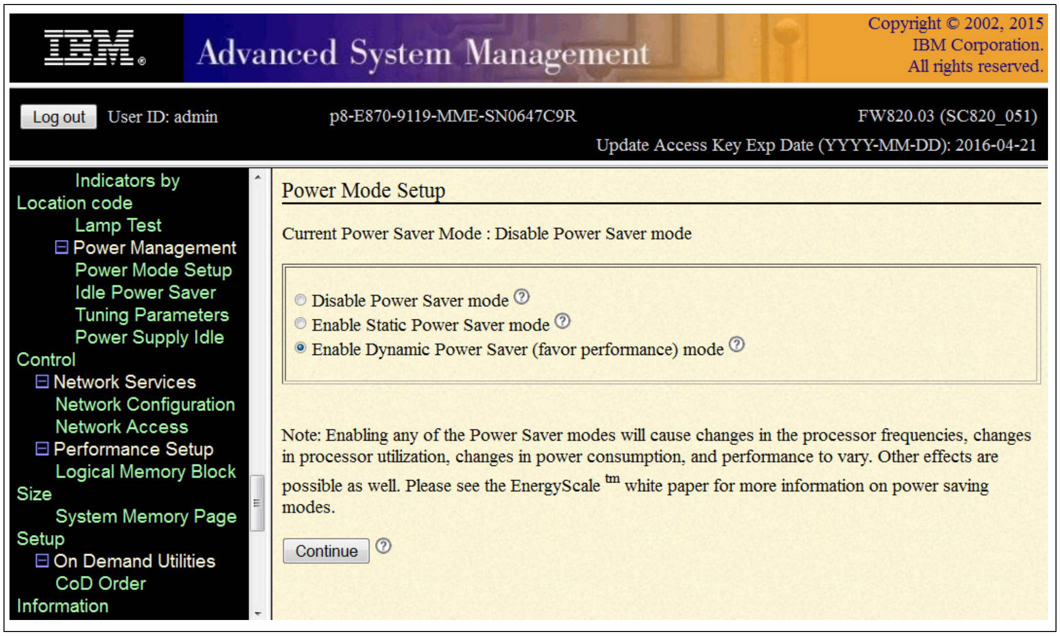
- EnergyScale for I/O

  IBM POWER8 processor-based systems automatically power off hot-pluggable PCI adapter slots that are empty or not being used. System firmware automatically scans all pluggable PCI slots at regular intervals, looking for those that meet the criteria for being not in use and powering them off. This support is available for all POWER8 processor-based servers and the expansion units that they support.

- Server power down

  If overall data center processor usage is low, workloads can be consolidated on fewer numbers of servers so that some servers can be turned off completely. Consolidation makes sense when there are long periods of low usage, such as weekends. Live Partition Mobility can be used to move workloads to consolidate partitions onto fewer systems, reducing the number of servers that are powered on and therefore reducing the power usage.

  On POWER8 processor-based systems, several EnergyScale technologies are embedded in the hardware and do not require an operating system or external management component. Fan control, environmental monitoring, and system energy management are controlled by the On Chip Controller (OCC) and associated components. The power mode can also be set up without external tools, by using the ASMI interface, as shown in Figure 2-42.



Power Mode Setup

Current Power Saver Mode : Disable Power Saver mode

○ Disable Power Saver mode
○ Enable Static Power Saver mode
● Enable Dynamic Power Saver (favor performance) mode

Note: Enabling any of the Power Saver modes will cause changes in the processor frequencies, changes in processor utilization, changes in power consumption, and performance to vary. Other effects are possible as well. Please see the EnergyScale ^{tm} white paper for more information on power saving modes.

Figure 2-42 Setting the power mode in ASMI

2.13.2 On Chip Controller

To maintain the power dissipation of POWER7+ with its large increase in performance and bandwidth, POWER8 invested significantly in power management innovations. A new On Chip Controller (OCC) using an embedded IBM PowerPC core with 512 KB of SRAM runs real-time control firmware to respond to workload variations by adjusting the per-core frequency and voltage based on activity, thermal, voltage, and current sensors.

The on-die nature of the OCC allows for approximately 100× speedup in response to workload changes over POWER7+, enabling reaction under the timescale of a typical OS time slice and allowing for multi-socket, scalable systems to be supported. It also enables more granularity in controlling the energy parameters in the processor, and increases reliability in energy management by having one controller in each processor that can perform certain functions independently of the others.

POWER8 also includes an internal voltage regulation capability that enables each core to run at a different voltage. Optimizing both voltage and frequency for workload variation enables better increase in power savings versus optimizing frequency only.

2.13.3 Energy consumption estimation

Often, for Power Systems, various energy-related values are important:

- Maximum power consumption and power source loading values

  These values are important for site planning and are described in the IBM Knowledge Center, found at the following website:

  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp

  Search for type and model number and “server specifications”. For example, for the Power E870 and Power E880 system, search for “9119-MME and 9119-MHE server specifications”.

- An estimation of the energy consumption for a certain configuration

  The calculation of the energy consumption for a certain configuration can be done in the IBM Systems Energy Estimator, found at the following website:

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

  In that tool, select the type and model for the system, and enter some details about the configuration and wanted CPU usage. As a result, the tool shows the estimated energy consumption and the waste heat at the wanted usage and also at full usage.

3

Virtualization

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

IBM Power Systems combined with PowerVM technology offer key capabilities that can help you consolidate and simplify your IT environment:

- Improve server usage and share I/O resources to reduce total cost of ownership (TCO) and make better usage of IT assets.
- Improve business responsiveness and operational speed by dynamically reallocating resources to applications as needed, to better match changing business needs or handle unexpected changes in demand.
- Simplify IT infrastructure management by making workloads independent of hardware resources, so you can make business-driven policies to deliver resources based on time, cost, and service-level requirements.

PowerVM license: Enterprise edition is default on Power E870 and Power E880.

Single Root I/O Virtualization (SR-IOV) is now also supported on the Power E870 and Power E880 server. For more information about SR-IOV see chapter 3.4, “Single root I/O virtualization (SR-IOV)” on page 126.

Note: PowerKVM is not supported on the Power E870 and Power E880. PowerVM is the only virtualization technology available for these systems.

3.1 IBM POWER Hypervisor

Combined with features in the POWER8 processors, the IBM POWER Hypervisor delivers functions that enable other system technologies, including logical partitioning technology, virtualized processors, IEEE VLAN-compatible virtual switch, virtual SCSI adapters, virtual Fibre Channel adapters, and virtual consoles. The POWER Hypervisor is a basic component of the system's firmware and offers the following functions:

- Provides an abstraction between the physical hardware resources and the logical partitions that use them.
- Enforces partition integrity by providing a security layer between logical partitions.
- Controls the dispatch of virtual processors to physical processors (see "Processing mode" on page 132).
- Saves and restores all processor state information during a logical processor context switch.
- Controls hardware I/O interrupt management facilities for logical partitions.
- Provides virtual LAN channels between logical partitions that help reduce the need for physical Ethernet adapters for inter-partition communication.
- Monitors the service processor and performs a reset or reload if it detects the loss of the service processor, notifying the operating system if the problem is not corrected.

The POWER Hypervisor is always active, regardless of the system configuration and also when not connected to the managed console. It requires memory to support the resource assignment to the logical partitions on the server. The amount of memory that is required by the POWER Hypervisor firmware varies according to several factors:

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

The minimum amount of physical memory that is required 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 that is configured in the system (see Table 3-1).

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

Configurable Central Electronics Complex memory	Default logical memory block
Up to 32 GB	128 MB
Greater than 32 GB	256 MB

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

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

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

3.1.1 Virtual SCSI

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

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

A Virtual I/O Server (VIOS) partition or an IBM i partition can define virtual SCSI server adapters. Other partitions are *client* partitions. The VIOS partition is a special logical partition, which is described in 3.5.4, “Virtual I/O Server” on page 134. The VIOS software is included on all PowerVM editions. When using the PowerVM Standard Edition and PowerVM Enterprise Edition, dual VIOSes can be deployed to provide maximum availability for client partitions when performing VIOS maintenance.

3.1.2 Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions on the same server to use fast and secure communication without any need for physical interconnection. The virtual Ethernet allows a transmission speed up to 20 Gbps, depending on the maximum transmission unit (MTU) size, type of communication, and CPU entitlement. Virtual Ethernet support began with IBM AIX Version 5.3, IBM i V5R3, Red Hat Enterprise Linux 4, and SUSE Linux Enterprise Server, 9, and it is supported on all later versions. (For more information, see 3.5.8, “Operating system support for PowerVM” on page 144). The virtual Ethernet is part of the base system configuration.

Virtual Ethernet has the following major features:

- The virtual Ethernet adapters can be used for both IPv4 and IPv6 communication and can transmit packets with a size up to 65,408 bytes. Therefore, the maximum MTU for the corresponding interface can be up to 65,394 (or 65,390 if VLAN tagging is used).
- The POWER Hypervisor presents itself to partitions as a virtual 802.1Q-compliant switch. The maximum number of VLANs is 4096. Virtual Ethernet adapters can be configured as either untagged or tagged (following the IEEE 802.1Q VLAN standard).
- A partition can support 256 virtual Ethernet adapters. Besides a default port VLAN ID, the number of additional VLAN ID values that can be assigned per virtual Ethernet adapter is 20, which implies that each virtual Ethernet adapter can be used to access 21 virtual networks.
- Each partition operating system detects the virtual local area network (VLAN) switch as an Ethernet adapter without the physical link properties and asynchronous data transmit operations.

Any virtual Ethernet can also have connectivity outside of the server if a Layer 2 bridge to a physical Ethernet adapter is set in one VIOS partition, also known as Shared Ethernet Adapter. For more information about shared Ethernet, see 3.5.4, “Virtual I/O Server” on page 134.

Adapter and access: 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.

3.1.3 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. Figure 3-1 shows the connections between the client partition virtual Fibre Channel adapters and the external storage. For more information, see 3.5.8, “Operating system support for PowerVM” on page 144.

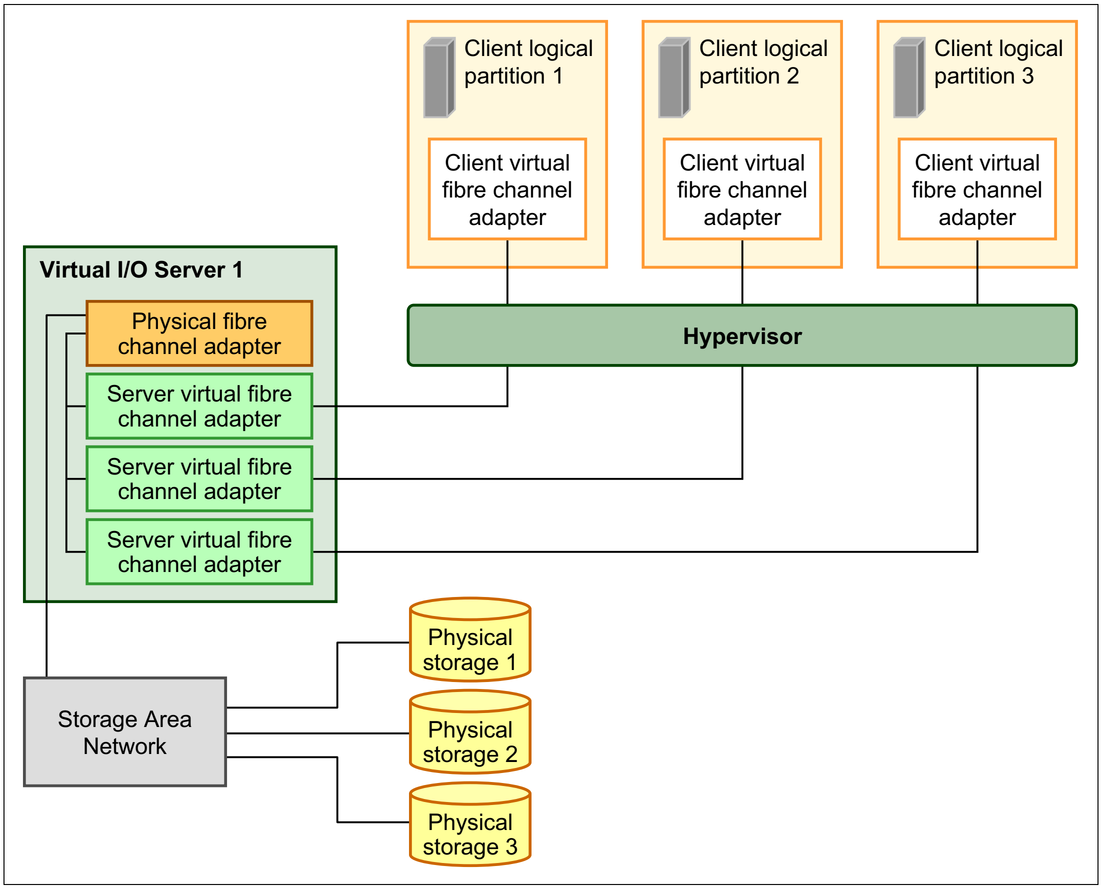

Client logical partition 1
Client virtual fibre channel adapter
Client logical partition 2
Client virtual fibre channel adapter
Client logical partition 3
Client virtual fibre channel adapter
Virtual I/O Server 1
Physical fibre channel adapter
Server virtual fibre channel adapter
Server virtual fibre channel adapter
Server virtual fibre channel adapter
Hypervisor
Physical storage 1
Physical storage 2
Physical storage 3
Storage Area Network

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

3.1.4 Virtual (TTY) console

Each partition must have access to a system console. Tasks such as operating system installation, network setup, and various problem analysis activities require a dedicated system console. The POWER Hypervisor provides the virtual console by using a virtual TTY or serial adapter and a set of Hypervisor calls to operate on them. Virtual TTY does not require the purchase of any additional features or software, such as the PowerVM Edition features.

Depending on the system configuration, the operating system console can be provided by the Hardware Management Console (HMC) virtual TTY, IVM virtual TTY, or from a terminal emulator that is connected to a system port.

3.2 POWER processor modes

Although they are not virtualization features, the POWER processor modes are described here because they affect various virtualization features.

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

- POWER6 compatibility mode

  This execution mode is compatible with Version 2.05 of the Power Instruction Set Architecture (ISA). For more information, visit the following website:

  http://power.org/wp-content/uploads/2012/07/PowerISA_V2.05.pdf

- POWER6+ compatibility mode

  This mode is similar to POWER6, with eight more storage protection keys.

- POWER7 mode

  This is the mode for POWER7+ and POWER7 processors, implementing Version 2.06 of the Power Instruction Set Architecture. For more information, visit the following website:

  http://power.org/wp-content/uploads/2012/07/PowerISA_V2.06B_V2_PUBLIC.pdf

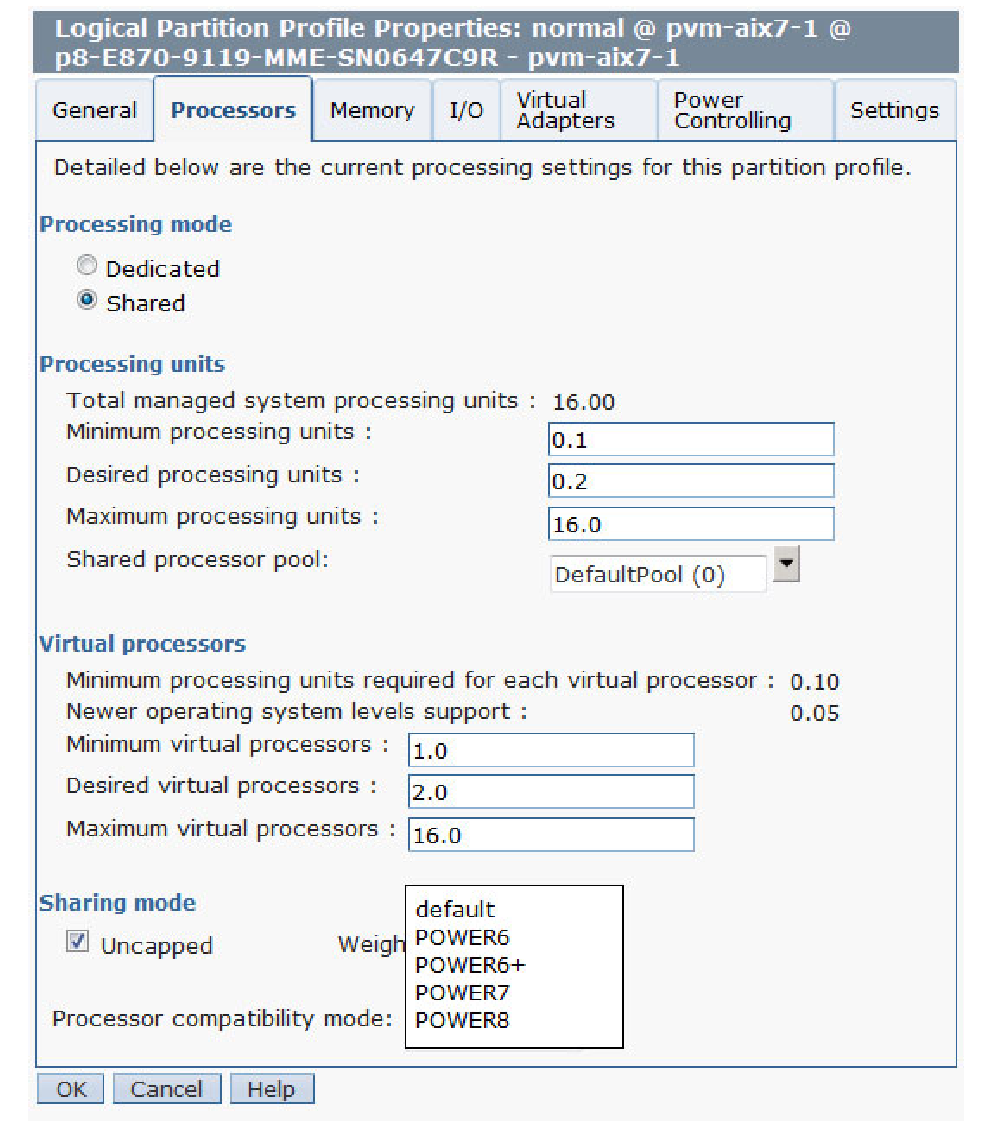
- POWER8 mode

  This is the native mode for POWER8 processors implementing Version 2.07 of the Power Instruction Set Architecture. For more information, visit the following address:

  https://www.power.org/documentation/power-isa-version-2-07/

The selection of the mode is made on a per-partition basis, from the managed console, by editing the partition profile.

Figure 3-2 shows the compatibility modes within the LPAR profile.



Logical Partition Profile Properties: normal @ pvm-aix7-1 @ p8-E870-9119-MME-SN0647C9R - pvm-aix7-1

Detailed below are the current processing settings for this partition profile.

Processing mode

○ Dedicated
● Shared

Processing units

Total managed system processing units : 16.00
Minimum processing units : 0.1
Desired processing units : 0.2
Maximum processing units : 16.0
Shared processor pool: DefaultPool (0)

Virtual processors

Minimum processing units required for each virtual processor : 0.10
Newer operating system levels support : 0.05
Minimum virtual processors : 1.0
Desired virtual processors : 2.0
Maximum virtual processors : 16.0

Sharing mode

☑ Uncapped Weigh

Processor compatibility mode:

default
POWER6
POWER6+
POWER7
POWER8

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

Table 3-2 lists the differences between the processor modes.

Table 3-2 Differences between POWER6, POWER7, and POWER8 compatibility modes

POWER6 and POWER6+ mode	POWER7 mode	POWER8 mode	Customer value
2-thread SMT	4-thread SMT	8-thread SMT	Throughput performance, and processor core usage
Vector Multimedia Extension/ AltiVec (VMX)	Vector scalar extension (VSX)	VSX2 In-Core Encryption Acceleration	High-performance computing

POWER6 and POWER6+ mode	POWER7 mode	POWER8 mode	Customer value
Affinity off by default	3-tier memory, micropartition affinity, and dynamic platform optimizer	► HW memory affinity tracking assists ► Micropartition prefetch ► Concurrent LPARs per core	Improved system performance for system images spanning sockets and nodes
64-core and 128-thread scaling	► 32-core and 128-thread scaling ► 64-core and 256-thread scaling ► 128-core and 512-thread scaling ► 256-core and 1024-thread scaling	► 192-core and 1536-thread scaling ► Hybrid threads ► Transactional memory ► Active system optimization hardware assists	Performance and scalability for large scale-up single system image workloads (such as OLTP, ERP scale-up, and WPAR consolidation)
EnergyScale CPU Idle	EnergyScale CPU Idle and Folding with NAP and SLEEP	WINKLE, NAP, SLEEP, and Idle power saver	Improved energy efficiency

3.3 Active Memory Expansion

Active Memory Expansion is an optional feature for the enterprise servers Power E870 and Power E880.

This feature enables memory expansion on the system. By using compression and decompression of memory, content can effectively expand the maximum memory capacity, providing additional server workload capacity and performance.

Active Memory Expansion is a technology that allows the effective maximum memory capacity to be much larger than the true physical memory maximum. Compression and decompression of memory content can allow memory expansion up to 125% for AIX partitions, which in turn enables a partition to perform more work or support more users with the same physical amount of memory. Similarly, it can allow a server to run more partitions and do more work for the same physical amount of memory.

Note: The Active Memory Expansion feature is not supported by IBM i and the Linux operating system.

Active Memory Expansion uses the CPU resource of a partition to compress and decompress the memory contents of this same partition. The trade-off of memory capacity for processor cycles can be an excellent choice, but the degree of expansion varies based on how compressible the memory content is, and it also depends on having adequate spare CPU capacity available for this compression and decompression.

The POWER8 processor includes Active Memory Expansion on the processor chip to provide dramatic improvement in performance and greater processor efficiency. To take advantage of the hardware compression offload, AIX 6.1 Technology Level 9 or later is required, for node1 and node2 with HMCV8R8.2.0 and for node2 and node3 with HMCV8R8.3.0+FW830.

Tests in IBM laboratories, using sample work loads, showed excellent results for many workloads in terms of memory expansion per additional CPU used. Other test workloads had more modest results. The ideal scenario is when there are many cold pages, that is, infrequently referenced pages. However, if many memory pages are referenced frequently, Active Memory Expansion might not be a preferred choice.

> **Tip:** If the workload is based on Java, the garbage collector must be tuned so that it does not access the memory pages so often, that is, turning cold pages to hot.

For more information about Active Memory Expansion, download the document *Active Memory Expansion: Overview and Usage Guide*, found at:

http://public.dhe.ibm.com/common/ssi/ecm/en/pow03037usen/POW03037USEN.PDF

3.4 Single root I/O virtualization (SR-IOV)

Single root I/O virtualization (SR-IOV) is an extension to the PCI Express (PCIe) specification that allows multiple operating systems to simultaneously share a PCIe adapter with little or no runtime involvement from a hypervisor or other virtualization intermediary.

SR-IOV is PCI standard architecture that enables PCIe adapters to become self-virtualizing. It enables adapter consolidation, through sharing, much like logical partitioning enables server consolidation. With an adapter capable of SR-IOV, you can assign virtual *slices* of a single physical adapter to multiple partitions through logical ports; all of this is done without the need for a Virtual I/O Server (VIOS).

Initial SR-IOV deployment supports up to 48 logical ports per adapter, depending on the adapter. You can provide additional fan-out for more partitions by assigning a logical port to a VIOS, and then using that logical port as the physical device for a Shared Ethernet Adapter (SEA). VIOS clients can then use that SEA through a traditional virtual Ethernet configuration.

Overall, SR-IOV provides integrated virtualization without VIOS and with greater server efficiency as more of the virtualization work is done in the hardware and less in the software.

The following are the hardware requirements to enable SR-IOV:

- One of the following pluggable PCIe adapters:
  - PCIe2 4-port (10Gb FCoE and 1GbE) SR&RJ45 Adapter (#EN0H)
  - PCIe2 4-port (10Gb FCoE and 1GbE) SFP+Copper and RJ4 Adapter (#EN0K)
  - PCIe2 LP 4-port(10Gb FCoE & 1GbE) SFP+Copper&RJ45 (#EN0L)
  - PCIe2 4-port (10Gb FCoE and 1GbE) LR&RJ45 Adapter (#EN0M)
  - PCIe2 LP 4-port(10Gb FCoE & 1GbE) LR&RJ45 Adapter (#EN0N)
  - PCIe3 4-port (10Gb) SR Adapter (#EN15)
  - PCIe3 LP 4-port (10Gb) SR Adapter (#EN16)
  - PCIe3 4-port (10Gb) SFP+Copper Adapter (#EN17)
  - PCIe3 LP 4-port (10Gb) SFP+Copper Adapter (#EN18)

The minimum operating system requirements, related to SR-IOV functions, are as follows:

- VIOS
  - Virtual I/O Server Version 2.2.3.51
- AIX
  - AIX 6.1 Technology Level 9 with Service Pack 5 and APAR IV68443 or later
  - AIX 7.1 Technology Level 3 with Service Pack 5 and APAR IV68444 or later

- Linux
  - SUSE Linux Enterprise Server 11 SP3, or later
  - SUSE Linux Enterprise Server 12, or later
  - Red Hat Enterprise Linux 6.5, or later
  - Red Hat Enterprise Linux 7, or later
- IBM i
  - IBM i 7.1 TR10 or later
  - IBM i 7.2 TR2 or later

> **Firmware level:** SR-IOV is supported from firmware level SC820_067 (FW820.10) for POWER8 processor-based servers. Check the Fix Central portal to verify the specific firmware level for your type of the machine at:
>
> `https://www.ibm.com/support/fixcentral/`

The entire adapter (all four ports) is configured for SR-IOV or none of the ports is. (FCoE not supported when using SR-IOV).

SR-IOV provides significant performance and usability benefits, as described in the following sections.

3.4.1 Direct access I/O and performance

The primary benefit of allocating adapter functions directly to a partition, as opposed to using a virtual intermediary (VI) like VIOS, is performance. The processing overhead involved in passing client instructions through a VI, to the adapter and back, are substantial.

With direct access I/O, SR-IOV capable adapters running in shared mode allow the operating system to directly access the slice of the adapter that has been assigned to its partition, so there is no control or data flow through the hypervisor. From the partition perspective, the adapter appears to be physical I/O. With respect to CPU and latency, it exhibits the characteristics of physical I/O; and because the operating system is directly accessing the adapter, if the adapter has special features, like multiple queue support or receive side scaling (RSS), the partition can leverage those also, if the operating system has the capability in its device drivers.

3.4.2 Adapter sharing

The current trend of consolidating servers to reduce cost and improve efficiency is increasing the number of partitions per system, driving a requirement for more I/O adapters per system to accommodate them. SR-IOV addresses and simplifies that requirement by enabling the sharing of SR-IOV capable adapters. Because each adapter can be shared and directly accessed by up to 48 partitions, depending on the adapter, the partition to PCI slot ratio can be significantly improved without adding the overhead of a virtual intermediary.

3.4.3 Adapter resource provisioning (QoS)

Power Systems SR-IOV provides QoS controls to specify a capacity value for each logical port, improving the ability to share adapter ports effectively and efficiently. The capacity value determines the desired minimum percentage of the physical port's resources that should be applied to the logical port.

The exact resource represented by the capacity value can vary based on the physical port type and protocol. In the case of Ethernet physical ports, capacity determines the minimum percentage of the physical port's transmit bandwidth that the user desires for the logical port.

For example, consider Partitions A, B, and C, with logical ports on the same physical port. If Partition A is assigned an Ethernet logical port with a capacity value of 20%, Partitions B and C cannot use more than 80% of the physical port's transmission bandwidth unless Partition A is using less than 20%. Partition A can use more than 20% if bandwidth is available. This ensures that, although the adapter is being shared, the partitions maintain their portion of the physical port resources when needed.

3.4.4 Flexible deployment

Power Systems SR-IOV enables flexible deployment configurations, ranging from a simple, single-partition deployment, to a complex, multi-partition deployment involving VIOS partitions and VIOS clients running different operating systems.

In a single-partition deployment, the SR-IOV capable adapter in shared mode is wholly owned by a single partition, and no adapter sharing takes place. This scenario offers no practical benefit over traditional I/O adapter configuration, but the option is available.

In a more complex deployment scenario, an SR-IOV capable adapter could be shared by both VIOS and non-VIOS partitions, and the VIOS partitions could further virtualize the logical ports as shared Ethernet adapters for VIOS client partitions. This scenario leverages the benefits of direct access I/O, adapter sharing, and QoS that SR-IOV provides, and also the benefits of higher-level virtualization functions, such as Live Partition Mobility (for the VIOS clients), that VIOS can offer.

3.4.5 Reduced costs

SR-IOV facilitates server consolidation by reducing the number of physical adapters, cables, switch ports, and I/O slots required per system. This translates to reduced cost in terms of physical hardware required, and also reduced associated energy costs for power consumption, cooling, and floor space. You may save additional cost on CPU and memory resources, relative to a VIOS adapter sharing solution, because SR-IOV does not have the resource overhead inherent in using a virtualization intermediary to interface with the adapters.

3.5 PowerVM

The PowerVM platform is the family of technologies, capabilities, and offerings that delivers industry-leading virtualization on the IBM Power Systems. It is the umbrella branding term for Power Systems virtualization (logical partitioning, IBM Micro-Partitioning, POWER Hypervisor, VIOS, Live Partition Mobility, and more). As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and software. The licensed features of each of the two separate editions of PowerVM are described here.

3.5.1 PowerVM edition

Power VM Enterprise Edition (#5228) is standard on Power E870 and Power E880 servers. Power VM Standard Edition is not supported on the Power E870 and E880. PowerVM Enterprise Edition offers all PowerVM features including Advanced Memory Sharing (AMS) and Live Partition Mobility (LPM) for specific operating systems. To verify the details about what is specifically supported by the operating system, see Table 3-3 on page 144.

3.5.2 Logical partitions

Logical partitions (LPARs) and virtualization increase the usage of system resources and add a level of configuration possibilities.

Logical partitioning

Logical partitioning was introduced with the POWER4 processor-based product line and AIX Version 5.1, Red Hat Enterprise Linux 3.0, and SUSE Linux Enterprise Server 9.0 operating systems. This technology was able to divide an IBM eServer™ pSeries (now IBM System p) system into separate logical systems, allowing each LPAR to run an operating environment on dedicated attached devices, such as processors, memory, and I/O components.

Later, dynamic logical partitioning increased the flexibility, allowing selected system resources, such as processors, memory, and I/O components, to be added and deleted from logical partitions while they are running. AIX Version 5.2, with all the necessary enhancements to enable dynamic LPAR, was introduced in 2002. At the same time, Red Hat Enterprise Linux 5 and SUSE Linux Enterprise 9.0 were also able to support dynamic logical partitioning. The ability to reconfigure dynamic LPARs encourages system administrators to dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

Micro-Partitioning

When you use the Micro-Partitioning technology, you can allocate fractions of processors to a logical partition. This technology was introduced with POWER5 processor-based systems. A logical partition using fractions of processors is also known as a *shared processor partition* or *micropartition*. Micropartitions run over a set of processors that are called a *shared processor pool*, and virtual processors are used to let the operating system manage the fractions of processing power that are assigned to the logical partition. From an operating system perspective, a virtual processor cannot be distinguished from a physical processor, unless the operating system is enhanced to determine 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:

- The minimum, wanted, and maximum processing units

  Processing units are defined as processing power, or the fraction of time that the partition is dispatched on physical processors. Processing units define the capacity entitlement of the partition.

- The shared processor pool

  Select a pool from the list with the names of each configured shared processor pool. This list also shows, in parentheses, the pool ID of each configured shared processor pool. If the name of the wanted shared processor pool is not available here, you must first configure the shared processor pool by using the shared processor pool Management window. Shared processor partitions use the default shared processor pool, called DefaultPool by default. For more information about multiple shared processor pools, see 3.5.3, “Multiple shared processor pools” on page 133.

- Whether the partition can access extra processing power to “fill up” its virtual processors above its capacity entitlement (you select either to cap or uncap your partition)

  If spare processing power is available in the shared processor pool or other partitions are not using their entitlement, an uncapped partition can use additional processing units if its entitlement is not enough to satisfy its application processing demand.

- The weight (preference) if there is an uncapped partition
- The minimum, wanted, and maximum number of virtual processors

The POWER Hypervisor calculates partition processing power based on minimum, wanted, and maximum values, processing mode, and the requirements of other active partitions. The actual entitlement is never smaller than the processing unit’s wanted value, but can exceed that value if it is an uncapped partition and up to the number of virtual processors that are allocated.

On the POWER8 processors, a partition can be defined with a processor capacity as small as 0.05 processing units. This number represents 0.05 of a physical core. Each physical core can be shared by up to 20 shared processor partitions, and the partition’s entitlement can be incremented fractionally by as little as 0.01 of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC.

The Power E870 supports up to 80 cores in a single system. Here are the maximum numbers:

- 80 dedicated partitions
- 1024 micropartitions

The Power E880 has statement of direction to support up to 192 cores in a single system. Here are the maximum numbers:

- 192 dedicated partitions
- 1024 micropartitions

An important point is that the maximum amounts are supported by the hardware, but the practical limits depend on application workload demands.

Consider the following additional information about virtual processors:

- A virtual processor can be running (dispatched) either on a physical core or as standby waiting for a physical core to became available.
- Virtual processors do not introduce any additional abstraction level. They are only a dispatch entity. When running on a physical processor, virtual processors run at the same speed as the physical processor.
- Each partition's profile defines a CPU entitlement that determines how much processing power any given partition should receive. The total sum of CPU entitlement of all partitions cannot exceed the number of available physical processors in a shared processor pool.
- The number of virtual processors can be changed dynamically through a dynamic LPAR operation.

Processing mode

When you create a logical partition, you can assign entire processors for dedicated use, or you can assign partial processing units from a shared processor pool. This setting defines the processing mode of the logical partition. Figure 3-3 shows a diagram of the concepts that are described in this section.

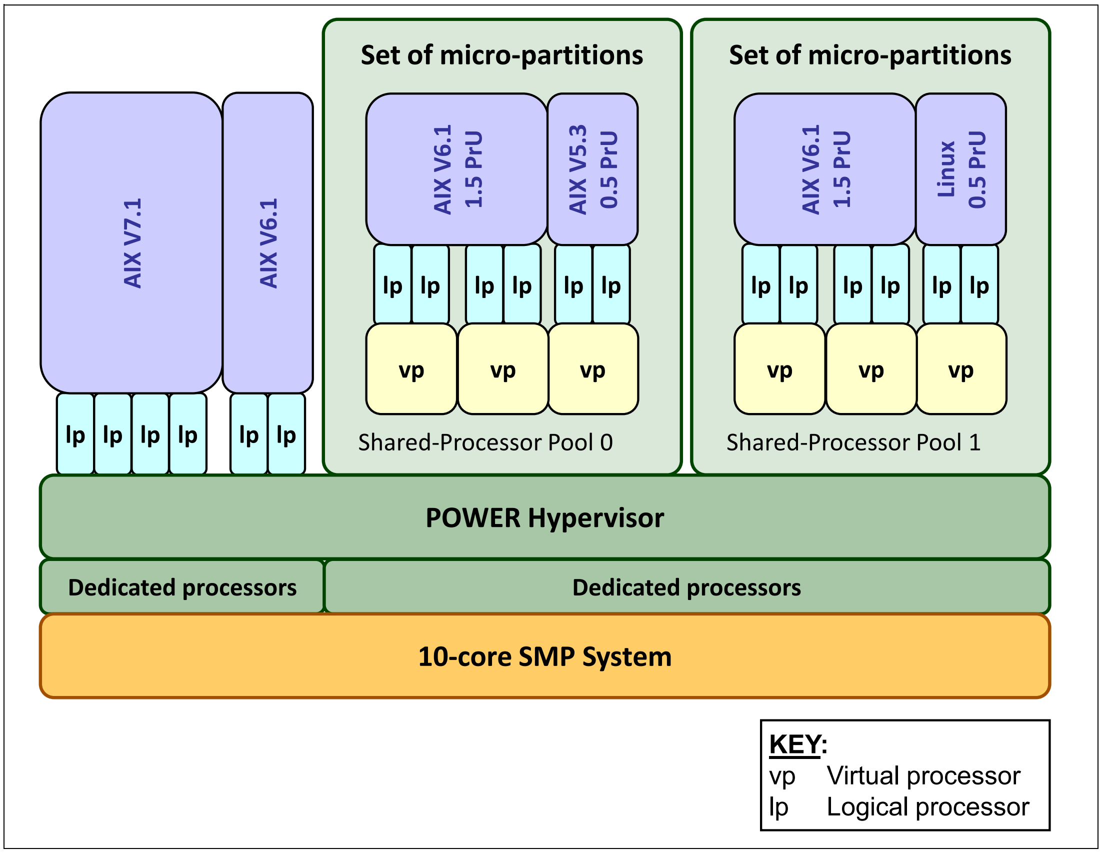


Figure 3-3 Logical partitioning concepts

Dedicated mode

In dedicated mode, physical processors are assigned as a whole to partitions. The simultaneous multithreading feature in the POWER8 processor core allows the core to run instructions from two, four, or eight independent software threads simultaneously.

To support this feature, consider the concept of *logical processors*. The operating system (AIX or Linux) sees one physical core as two, four, or eight logical processors if the simultaneous multithreading feature is on. It can be turned off and on dynamically while the operating system is running (for AIX, run **smtctl**, for Linux, run **ppc64_cpu --smt**, and for IBM i, use **QPRCMLTTSK** system value). If simultaneous multithreading is off, each physical core is presented as one logical processor in AIX or Linux, and thus only one thread.

Shared dedicated mode

On POWER8 processor technology-based servers, you can configure dedicated partitions to become processor donors for idle processors that they own, allowing for the donation of spare CPU cycles from dedicated processor partitions to a shared processor pool. The dedicated partition maintains absolute priority for dedicated CPU cycles. Enabling this feature can help increase system usage without compromising the computing power for critical workloads in a dedicated processor.

Shared mode

In shared mode, logical partitions use virtual processors to access fractions of physical processors. Shared partitions can define any number of virtual processors (the maximum number is 20 times the number of processing units that are 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 the partition's processing units entitlement. One processing unit represents one physical processor's processing capacity.

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

3.5.3 Multiple shared processor pools

Multiple shared processor pools (MSPPs) are supported on POWER8 processor-based servers. This capability allows a system administrator to create a set of micropartitions with the purpose of controlling the processor capacity that can be consumed from the physical shared processor pool.

Implementing MSPPs depends on a set of underlying techniques and technologies. Figure 3-4 shows an overview of the architecture of multiple shared processor pools.

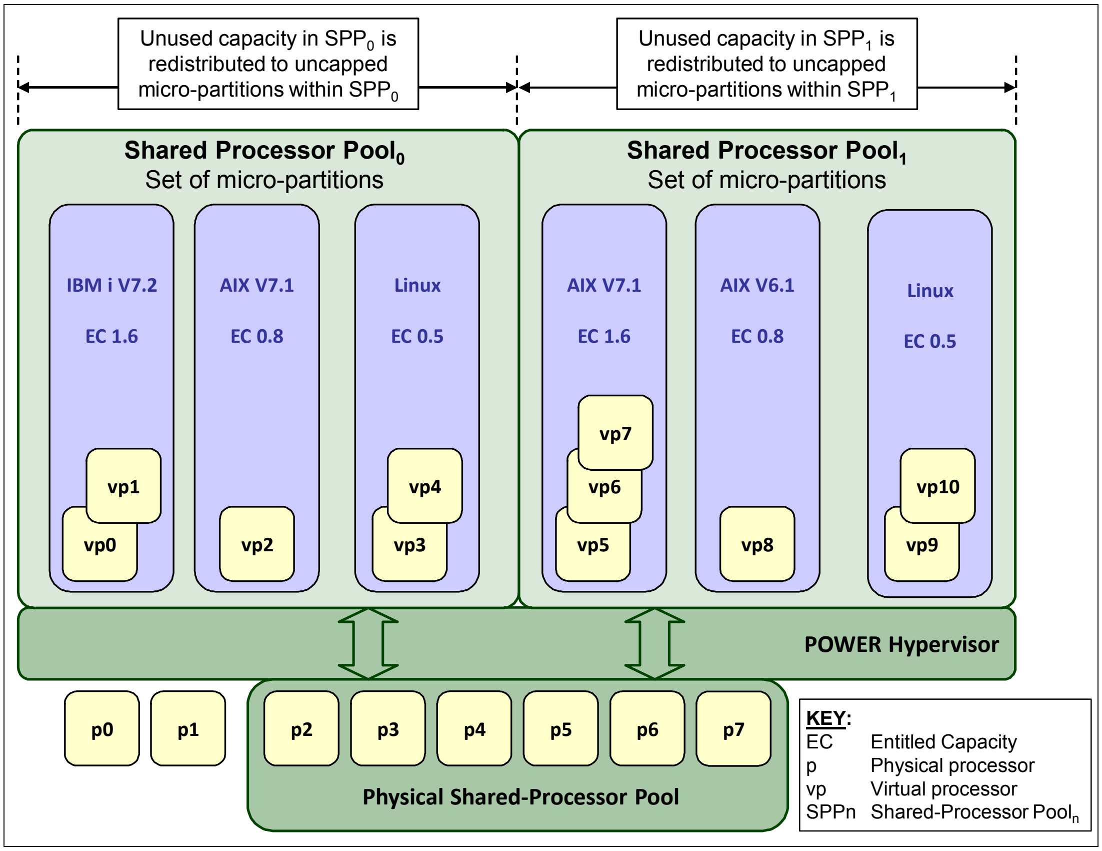

Unused capacity in SPP_0 is redistributed to uncapped micro-partitions within SPP_0
Unused capacity in SPP_1 is redistributed to uncapped micro-partitions within SPP_1
Shared Processor $Pool_0$
Set of micro-partitions
Shared Processor $Pool_1$
Set of micro-partitions
IBM i V7.2
EC 1.6
AIX V7.1
EC 0.8
Linux
EC 0.5
AIX V7.1
EC 1.6
AIX V6.1
EC 0.8
Linux
EC 0.5
vp0
vp1
vp2
vp3
vp4
vp5
vp6
vp7
vp8
vp9
vp10
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 3-4 Overview of the architecture of multiple shared processor pools

Micropartitions are created and then identified as members of either the default shared processor pool$_0$ or a user-defined shared processor pool$_n$. The virtual processors that exist within the set of micropartitions are monitored by the POWER Hypervisor, and processor capacity is managed according to user-defined attributes.

If the Power Systems server is under heavy load, each micropartition within a shared processor pool is ensured its processor entitlement plus any capacity that it might be allocated from the reserved pool capacity if the micropartition is uncapped.

If certain micropartitions in a shared processor pool do not use their capacity entitlement, the unused capacity is ceded and other uncapped micropartitions within the same shared processor pool are allocated the additional capacity according to their uncapped weighting. In this way, the entitled pool capacity of a shared processor pool is distributed to the set of micropartitions within that shared processor pool.

All Power Systems servers that support the multiple shared processor pools capability have a minimum of one (the default) shared processor pool and up to a maximum of 64 shared processor pools.

3.5.4 Virtual I/O Server

The virtual I/O Server (VIOS) is part of all PowerVM editions. It is a special-purpose partition that allows the sharing of physical resources between logical partitions to allow more efficient usage (for example, consolidation). In this case, the VIOS owns the physical resources (SCSI, Fibre Channel, network adapters, and optical devices) and allows client partitions to share access to them, thus minimizing the number of physical adapters in the system.

The 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 3-5 shows an overview of a VIOS configuration.

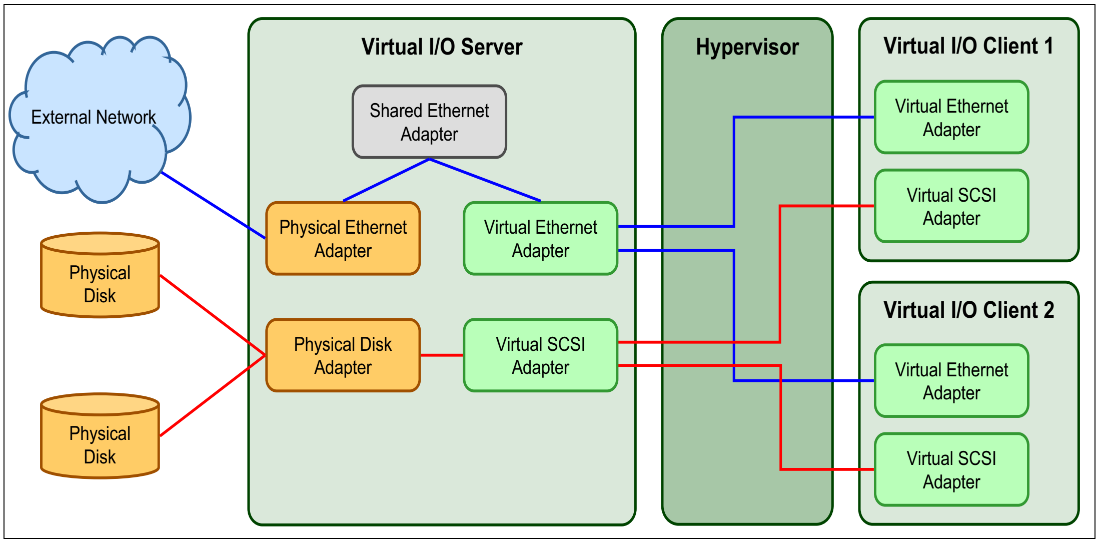


Figure 3-5 Architectural view of the VIOS

Because the VIOS is an operating system-based appliance server, redundancy for physical devices that are 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 is supported only in special VIOS partitions. Three major virtual devices are supported by the VIOS:

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

The Virtual Fibre Channel adapter is used with the NPIV feature, as described in 3.5.8, “Operating system support for PowerVM” on page 144.

Shared Ethernet Adapter

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

The SEA also provides the ability for several client partitions to share one physical adapter. With an SEA, you can connect internal and external VLANs by using a physical adapter. The Shared Ethernet Adapter service can be hosted only 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.

Figure 3-6 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 shares the physical access.

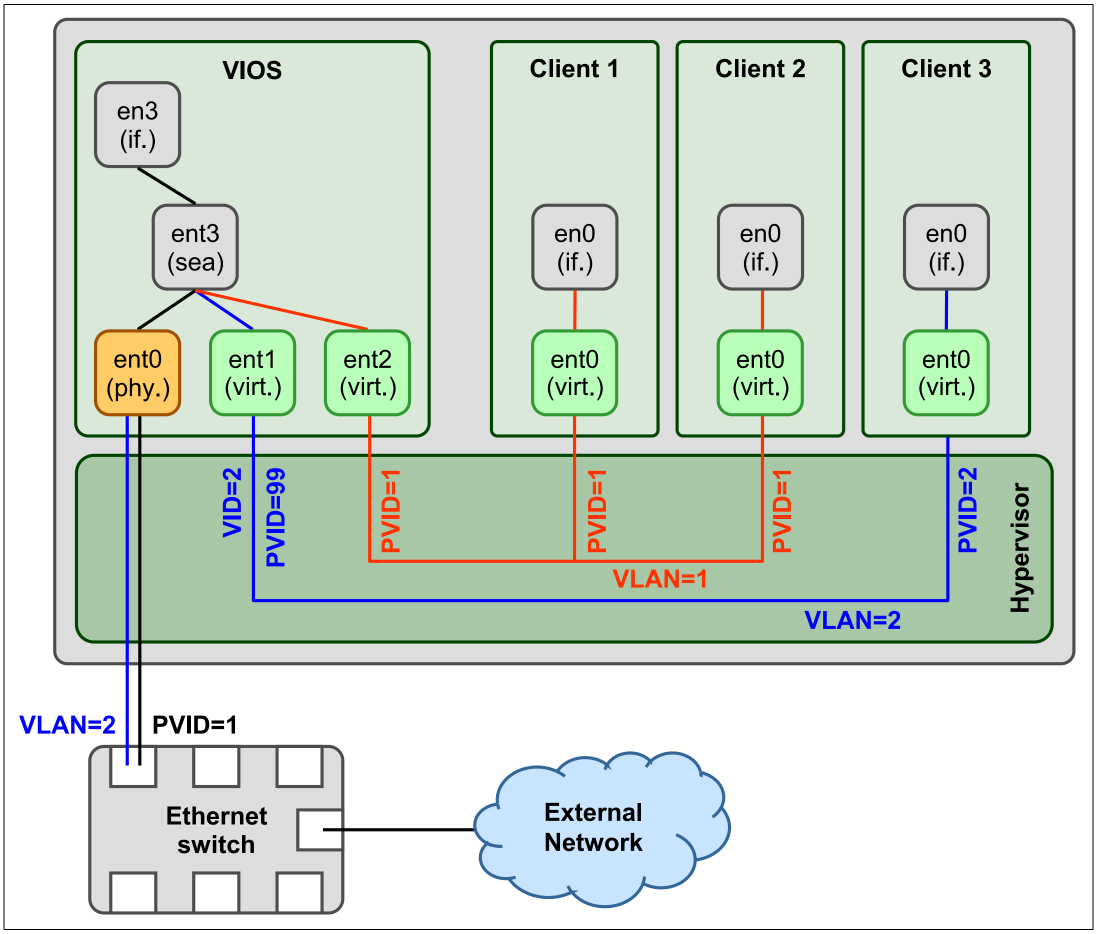


Figure 3-6 Architectural view of a Shared Ethernet Adapter

A single SEA setup can have up to 16 virtual Ethernet trunk adapters and each virtual Ethernet trunk adapter can support up to 20 VLAN networks. Therefore, a possibility is for a single physical Ethernet to be shared between 320 internal VLAN networks. The number of shared Ethernet adapters that can be set up in a 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 on an SEA.

Virtual SCSI

Virtual SCSI is used to see a virtualized implementation of the SCSI protocol. Virtual SCSI is based on a client/server relationship. The VIOS logical partition owns the physical resources and acts as a server or, in SCSI terms, a target device. The client logical partitions access the virtual SCSI backing storage devices that are provided by the VIOS as clients.

The virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured by using a managed console or through the Integrated Virtualization Manager on smaller systems. The virtual SCSI server (target) adapter is responsible for running any SCSI commands that 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 that are presented to the Virtual/O Server can be exported and assigned to a client partition in various ways:

- The entire disk is presented to the client partition.
- The disk is divided into several logical volumes, which can be presented to a single client or multiple clients.
- As of VIOS 1.5, files can be created on these disks, and file-backed storage devices can be created.

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

For more information about specific storage devices that are supported for VIOS, see the following website:

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

N_Port ID Virtualization

N_Port ID Virtualization (NPIV) is a technology that allows multiple logical partitions to access independent physical storage through the same physical Fibre Channel adapter. This adapter is attached to a VIOS partition that acts only as a pass-through, managing the data transfer through the POWER Hypervisor.

Each partition that uses NPIV is identified by a pair of unique worldwide port names, enabling you to connect each partition to independent physical storage on a SAN. Unlike virtual SCSI, only the client partitions see the disk.

For more information and requirements for NPIV, see the following resources:

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

Virtual I/O Server functions

The Virtual I/O Server (VIOS) has many features, including monitoring solutions and the following capabilities:

- Support for Live Partition Mobility starting on POWER6 processor-based systems with the PowerVM Enterprise Edition. For more information about Live Partition Mobility, see 3.5.6, “Active Memory Sharing” on page 142.
- Support for virtual SCSI devices that are backed by a file, which are then accessed as standard SCSI-compliant LUNs.
- Support for virtual Fibre Channel devices that are used with the NPIV feature.
- Virtual I/O Server Expansion Pack with additional security functions, such as Kerberos (Network Authentication Service for users and client and server applications), Simple Network Management Protocol (SNMP) v3, and Lightweight Directory Access Protocol (LDAP) client function.

- System Planning Tool (SPT) and Workload Estimator, which are designed to ease the deployment of a virtualized infrastructure. For more information about the System Planning Tool, see 3.6, “System Planning Tool” on page 152.
- IBM Systems Director agent and several preinstalled IBM Tivoli® agents, such as the following examples:
  - IBM Tivoli Identity Manager, which allows easy integration into an existing Tivoli Systems Management infrastructure
  - IBM Tivoli Application Dependency Discovery Manager (ADDM), which creates and automatically maintains application infrastructure maps, including dependencies, change histories, and deep configuration values
- vSCSI enterprise reliability, availability, and serviceability (eRAS).
- Additional CLI statistics in **svmon**, **vmstat**, **fcstat**, and **topas**.
- The VIOS Performance Advisor tool provides advisory reports based on key performance metrics for various partition resources that are collected from the VIOS environment.
- Monitoring solutions to help manage and monitor the VIOS and shared resources. Commands and views provide additional metrics for memory, paging, processes, Fibre Channel HBA statistics, and virtualization.

For more information about the VIOS and its implementation, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940

3.5.5 Live Partition Mobility

Live Partition Mobility (LPM) is a technique that allows a partition running on one server to be migrated dynamically to another server.

This feature can be extremely useful in a situation where a system needs to be evacuated for maintenance but its partitions do not allow for downtime. LPM allows for all the partitions to be moved while running to another server so the system can be properly shut down without impacts for the applications.

In simplified terms, LPM typically works in an environment where all of the I/O from one partition is virtualized through PowerVM and VIOS and all partition data is stored in a Storage Area Network (SAN) accessed by both servers.

To migrate a partition from one server to another, a partition is identified on the new server and configured to have the same virtual resources as the primary server including access to the same logical volumes as the primary using the SAN.

When an LPM migration is initiated on a server for a partition, PowerVM in the first system starts copying the state of memory in the first partition over to a destination partition in another server through each system PowerVM. This is done across a LAN while the initial partition continues to run. PowerVM has control of I/O operations through I/O virtualization and keeps track of memory changes that occur throughout the process.

At some point, when all of the memory state is copied from the primary partition, the primary partition is paused. PowerVM in the second server takes over control of the shared storage resources and allows the partition now running in that server to resume processing at the point where the first server left off.

Thinking in terms of using LPM for hardware repairs, if all of the workloads on a server are migrated by LPM to other servers, then after all have been migrated, the first server could be turned off to repair components.

LPM can also be used for doing firmware upgrades or adding additional hardware to a server when the hardware cannot be added concurrently in addition to software maintenance within individual partitions.

In successful LPM situations, while there may be a short time when applications are not processing new workload, the applications do not fail or crash and do not need to be restarted. Roughly speaking then, LPM allows for planned outages to occur on a server without suffering downtime that would otherwise be required.

Minimum configuration

For LPM to work, it is necessary that the system containing a partition to be migrated, and the system being migrated to, both have a local LAN connection using a virtualized LAN adapter. The LAN adapter should be high speed for better migration performance. The LAN used should be a local network and should be private and have only two uses. The first is for communication between servers; the second is for communication between partitions on each server and the HMC for resource monitoring and control functions (RMC.)

LPM also needs all systems in the LPM cluster to be attached to the same SAN (when using SAN for required common storage), which typically requires use of Fibre Channel adapters.

If a single HMC is used to manage both systems in the cluster, connectivity to the HMC also needs to be provided by an Ethernet connection to each service processor.

The LAN and SAN adapters used by the partition must be assigned to a Virtual I/O server and the partitions access to these would be by virtual LAN (VLAN) and virtual SCSI (vSCSI) connections within each partition to the VIOS.

Each partition to be migrated must only use virtualized I/O through a VIOS; there can be no non-virtualized adapters assigned to such partitions.

A diagram with the minimum requirements can be seen in Figure 3-7.

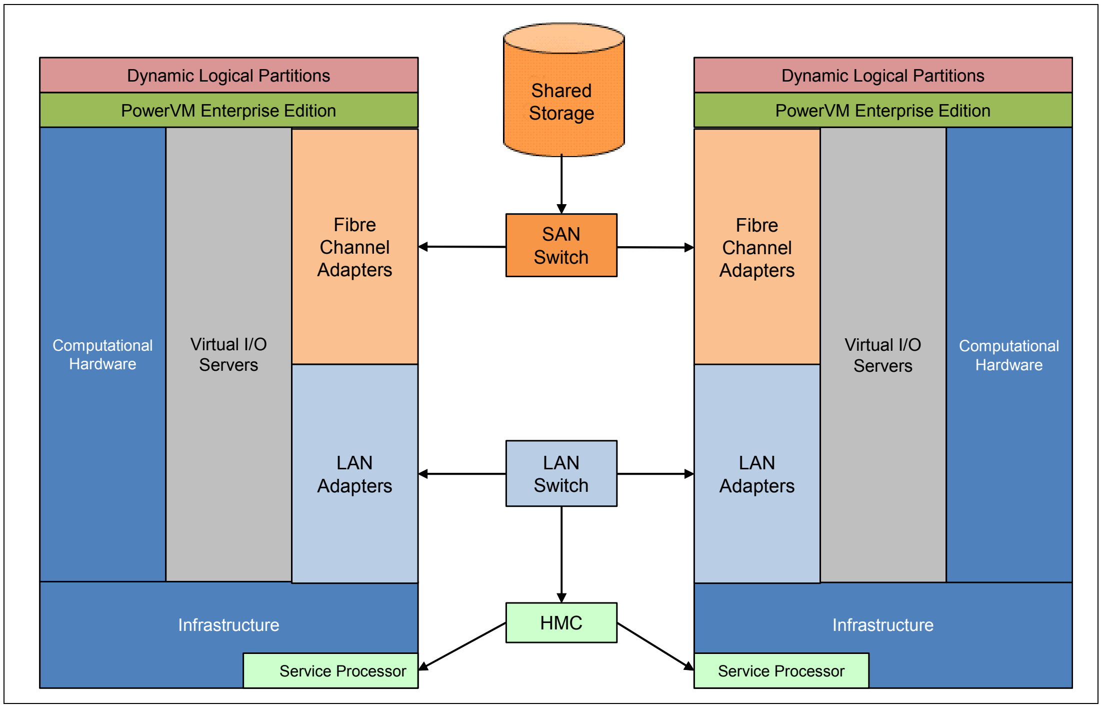


Figure 3-7 Minimum Live Partition Mobility requirements

Suggested configuration

LPM connectivity in the minimum configuration discussion is vulnerable to a number of different hardware and firmware faults that would lead to the inability to migrate partitions. Multiple paths to networks and SANs are therefore advised. To accomplish this, a VIOS server can be configured to use dual Fibre Channel and LAN adapters.

Externally to each system, redundant hardware management consoles (HMCS) can be utilized for greater availability. There can also be options to maintain redundancy in SANs and local network hardware. A diagram with the suggested scenario can be seen in Figure 3-8.

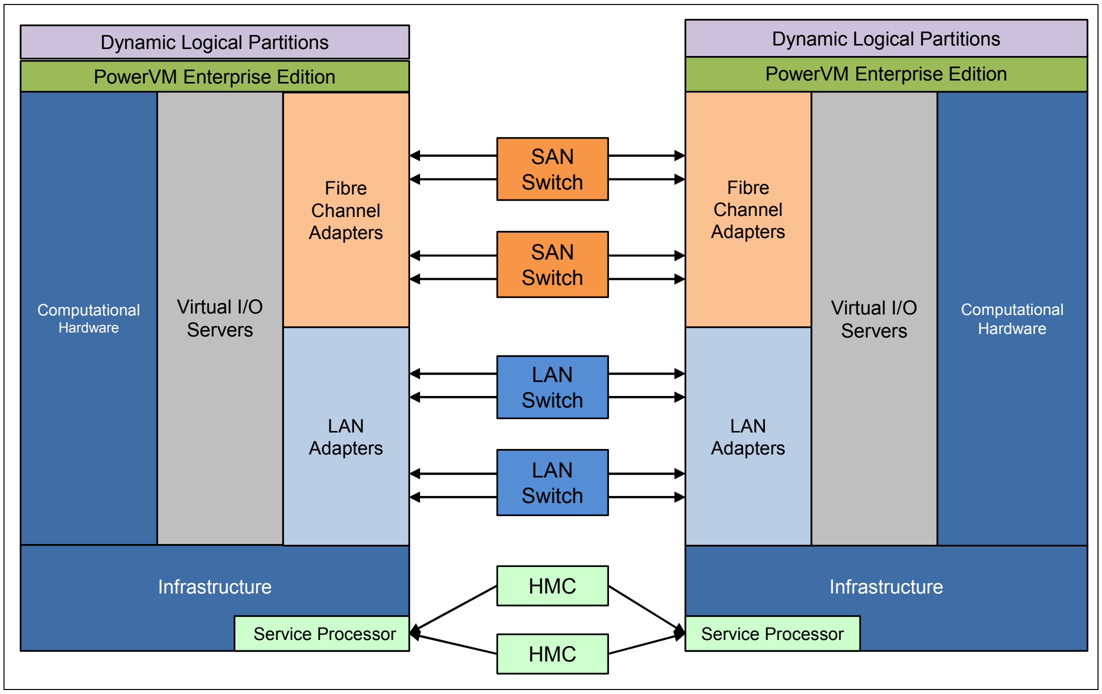


Figure 3-8 Redundant infrastructure for LPM

PCIe slot selection

POWER8 processor has PCIe controllers integrated on the chip allowing for a single processor to have two or more PCIe gen3 slots directly attached to it. In case of a complete processor failure, these slots might become unusable.

When this affects availability of PCIe slots, it must be taken into consideration while selecting the slot placement for the adapters on the VIrtual I/O Servers.

For scale-out class systems, each processor module directly controls two 16 lane (x16) PCIe slots and additional I/O capabilities are provided by x8 connections to a PCIe switch integrated in the processor board. All the slots connected to the PCIe switch belong to a single processor.

Figure 3-9 illustrates how such a system could be configured to maximize redundancy in a VIOS environment, presuming that the rootvgs for each VIOS are accessed from storage area networks.

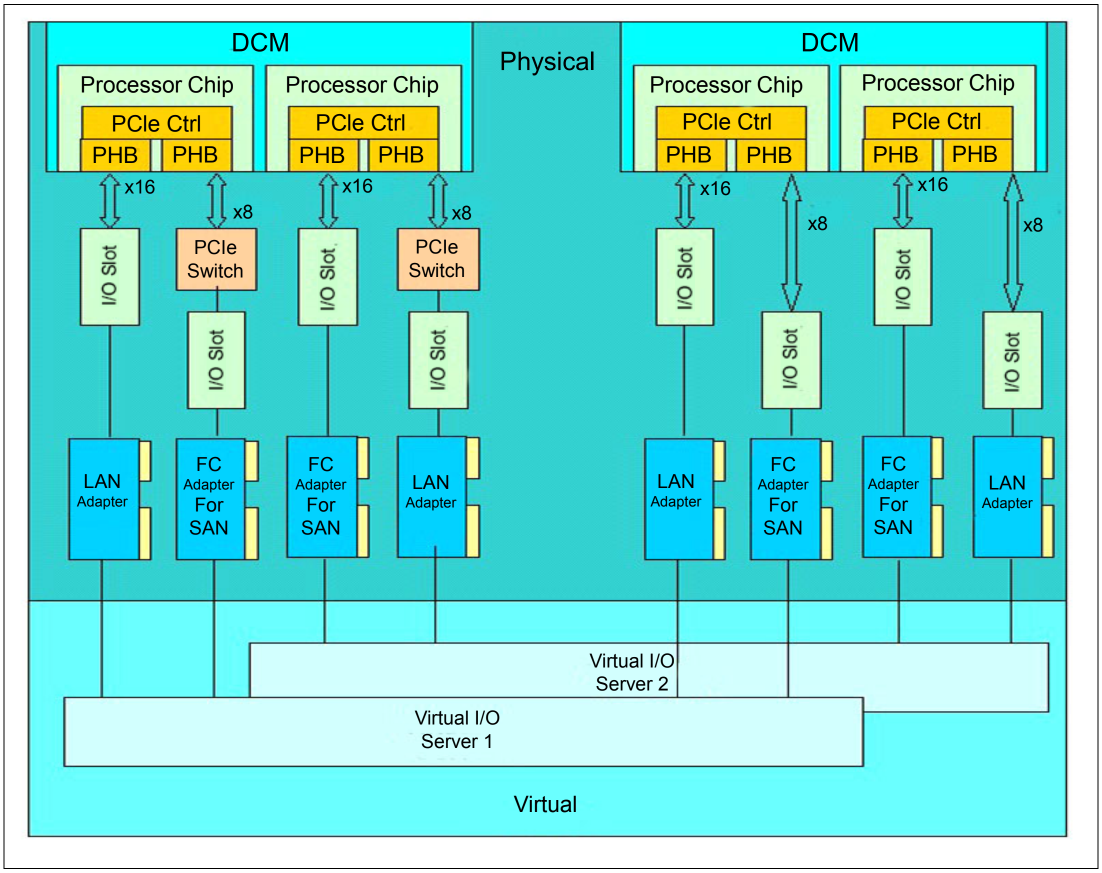


Figure 3-9 I/O Subsystem of a POWER8 2-socket system

For enterprise class systems, I/O drawers for expanding I/O are supported. When these drawers are used, all the slots connected to a PCIe Adapter for Expansion Drawer are bound to a given processor. A similar concept for I/O redundancy can be used to maximize availability of I/O access using two I/O drawers, one connected to each processor socket in the system. A logical diagram with the PCIe slots and its processors can be seen in Figure 3-10.

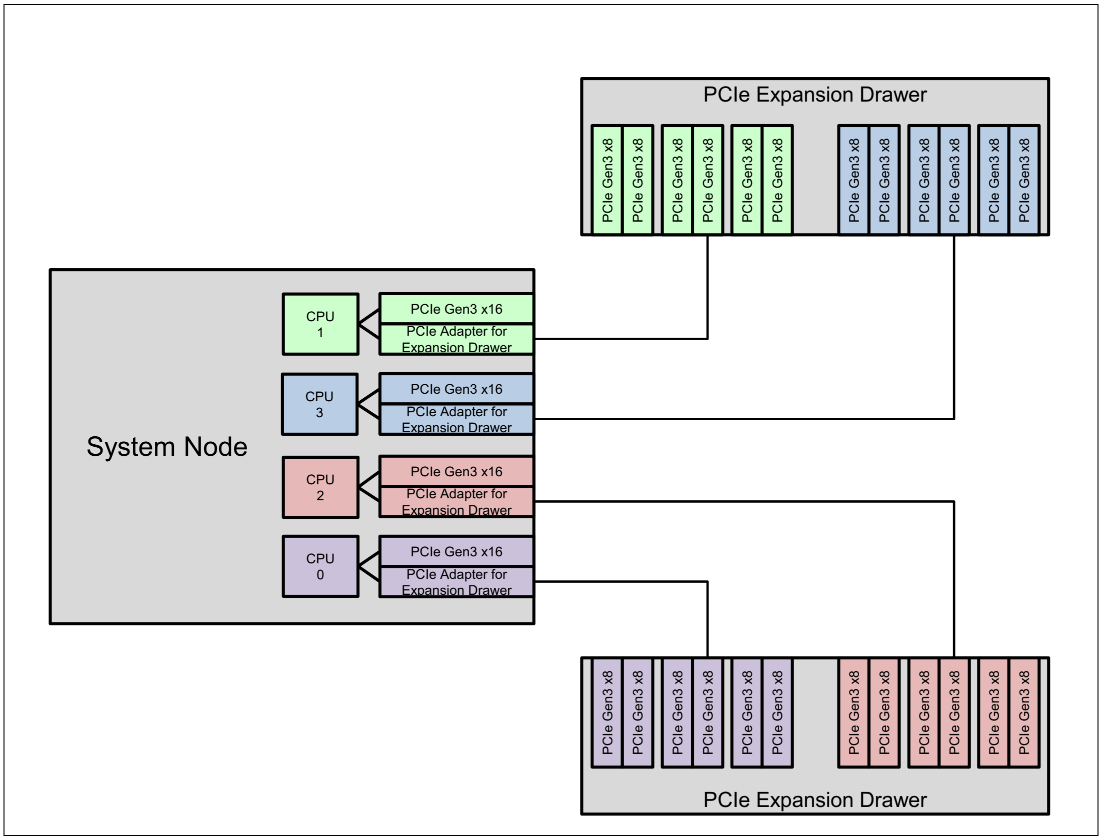


Figure 3-10 Logical diagram of processors and its associated PCIe slots

More information

For more information about PowerVM and Live Partition Mobility, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940:

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

3.5.6 Active Memory Sharing

Active Memory Sharing is an IBM PowerVM advanced memory virtualization technology that provides system memory virtualization capabilities to IBM Power Systems, allowing multiple partitions to share a common pool of physical memory.

Active Memory Sharing is available only with the Enterprise version of PowerVM.

The physical memory of an IBM Power System can be assigned to multiple partitions in either dedicated or shared mode. The system administrator can assign some physical memory to a partition and some physical memory to a pool that is shared by other partitions. A single partition can have either dedicated or shared memory:

- With a pure dedicated memory model, the system administrator's task is to optimize available memory distribution among partitions. When a partition suffers degradation because of memory constraints and other partitions have unused memory, the administrator can manually issue a dynamic memory reconfiguration.
- With a shared memory model, the system automatically decides the optimal distribution of the physical memory to partitions and adjusts the memory assignment based on partition load. The administrator reserves physical memory for the shared memory pool, assigns partitions to the pool, and provides access limits to the pool.

Active Memory Sharing can be used to increase memory usage on the system either by decreasing the global memory requirement or by allowing the creation of additional partitions on an existing system. Active Memory Sharing can be used in parallel with Active Memory Expansion on a system running a mixed workload of several operating systems. For example, AIX partitions can take advantage of Active Memory Expansion. Other operating systems take advantage of Active Memory Sharing also.

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

3.5.7 Active Memory Deduplication

In a virtualized environment, the systems might have a considerable amount of duplicated information that is stored on RAM after each partition has its own operating system, and some of them might even share the same kinds of applications. On heavily loaded systems, this behavior might lead to a shortage of the available memory resources, forcing paging by the Active Memory Sharing partition operating systems, the Active Memory Deduplication pool, or both, which might decrease overall system performance.

Active Memory Deduplication allows the POWER Hypervisor to map dynamically identical partition memory pages to a single physical memory page within a shared memory pool. This way enables a better usage of the Active Memory Sharing shared memory pool, increasing the system's overall performance by avoiding paging. Deduplication can cause the hardware to incur fewer cache misses, which also leads to improved performance.

Active Memory Deduplication depends on the Active Memory Sharing feature being available, and it consumes CPU cycles that are donated by the Active Memory Sharing pool's VIOS partitions to identify deduplicated pages. The operating systems that are running on the Active Memory Sharing partitions can "hint" to the POWER Hypervisor that some pages (such as frequently referenced read-only code pages) are good for deduplication.

To perform deduplication, the hypervisor cannot compare every memory page in the Active Memory Sharing pool with every other page. Instead, it computes a small signature for each page that it visits and stores the signatures in an internal table. Each time that a page is inspected, a look-up of its signature is done in the known signatures in the table. If a match is found, the memory pages are compared to be sure that the pages are really duplicates. When a duplicate is found, the hypervisor remaps the partition memory to the existing memory page and returns the duplicate page to the Active Memory Sharing pool.

From the LPAR perspective, the Active Memory Deduplication feature is not apparent. If an LPAR attempts to modify a deduplicated page, the Power hypervisor grabs a free page from the Active Memory Sharing pool, copies the duplicate page contents into the new page, and maps the LPAR's reference to the new page so the LPAR can modify its own unique page.

For more information regarding Active Memory Deduplication, see *Power Systems Memory Deduplication*, REDP-4827.

3.5.8 Operating system support for PowerVM

Table 3-3 shows operating system support for virtualization features.

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

Feature	AIX 6.1 TL9 SP1	AIX 7.1 TL03 SP1	IBM i 7.1 TR 9	IBM i 7.2 TR 1	RHEL 6.6	RHEL 7.1	SLES 11 SP3	SLES 12	Ubuntu 15.04
Virtual SCSI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Ethernet	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared Ethernet Adapter	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Fibre Channel	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Tape	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Logical partitioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
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	Yes	Yes	Yes	Yes
Micro-Partitioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared dedicated capacity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Shared Processor Pools	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
VIOS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Integrated Virtualization Manager	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Suspend/resume and hibernation[a]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared Storage Pools	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Thin provisioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Active Memory Sharing[b]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Active Memory Deduplication	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Live Partition Mobility	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Simultaneous multithreading (SMT)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Active Memory Expansion	Yes	Yes	No	No	No	No	No	No	No

a. At the time of writing, Suspend/Resume is not available. Check with your IBM System Services Representative (SSR) for availability on POWER8 platforms.

b. At the time of writing, Active Memory Sharing when used with Live Partition Mobility is not supported. Check with your IBM SSR for availability on POWER8 platforms.

For more information about specific features for Linux, see the following website:

`http://pic.dhe.ibm.com/infocenter/lnxinfo/v3r0m0/index.jsp?topic=%2Fliaam%2Fsupportedfeaturesforlinuxonpowersystemsservers.htm`

3.5.9 Linux support

The IBM Linux Technology Center (LTC) contributes to the development of Linux by providing support for IBM hardware in Linux distributions. In particular, the LTC has available tools and code for the Linux communities to take advantage of the POWER8 technology and develop POWER8 optimized software.

For more information about specific Linux distributions, see the following website:

`http://pic.dhe.ibm.com/infocenter/lnxinfo/v3r0m0/index.jsp?topic=%2Fliaam%2Fliaamdistros.htm`

3.5.10 PowerVM simplification

Starting with HMC V8R8.2.0 PowerVM simplification was introduced. Figure 3-11 shows new options starting from the HMC firmware V8R1 that are available for PowerVM simplification. Selecting **Manage PowerVM** opens a new window, where a user can view and manage all the aspects of a PowerVM configuration, such as SEA, virtual networks, and virtual storage using the graphical interface only.

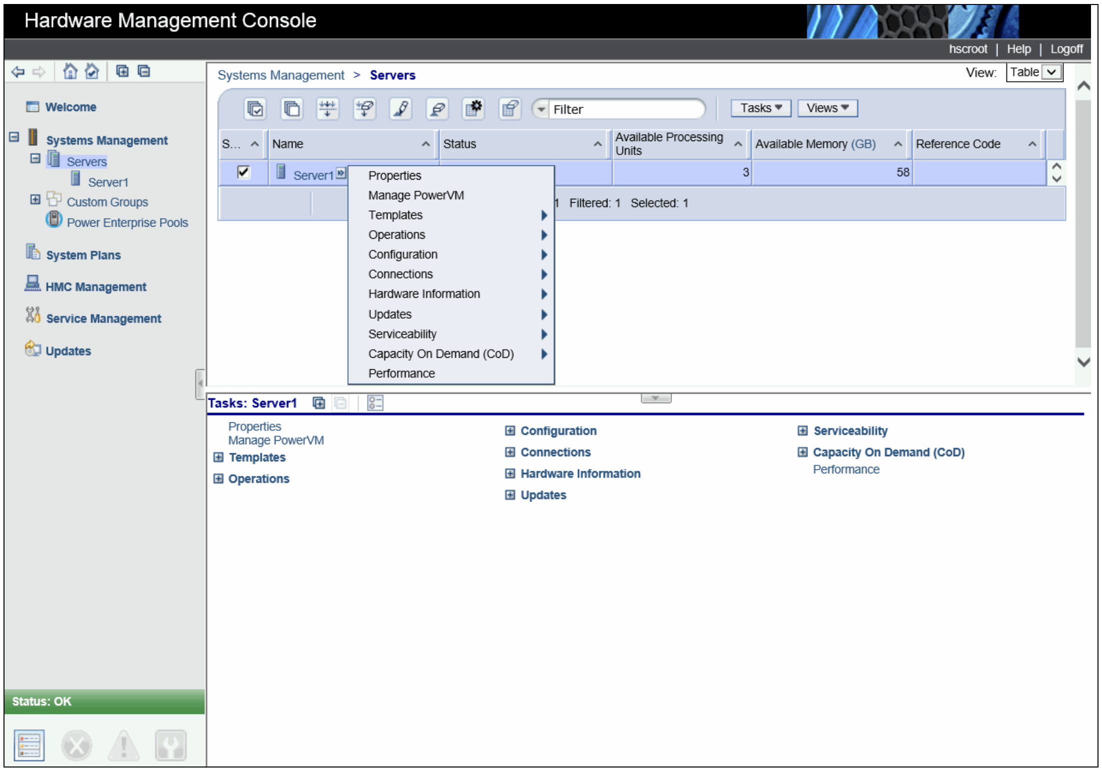


Figure 3-11 PowerVM new level tasks

Templates

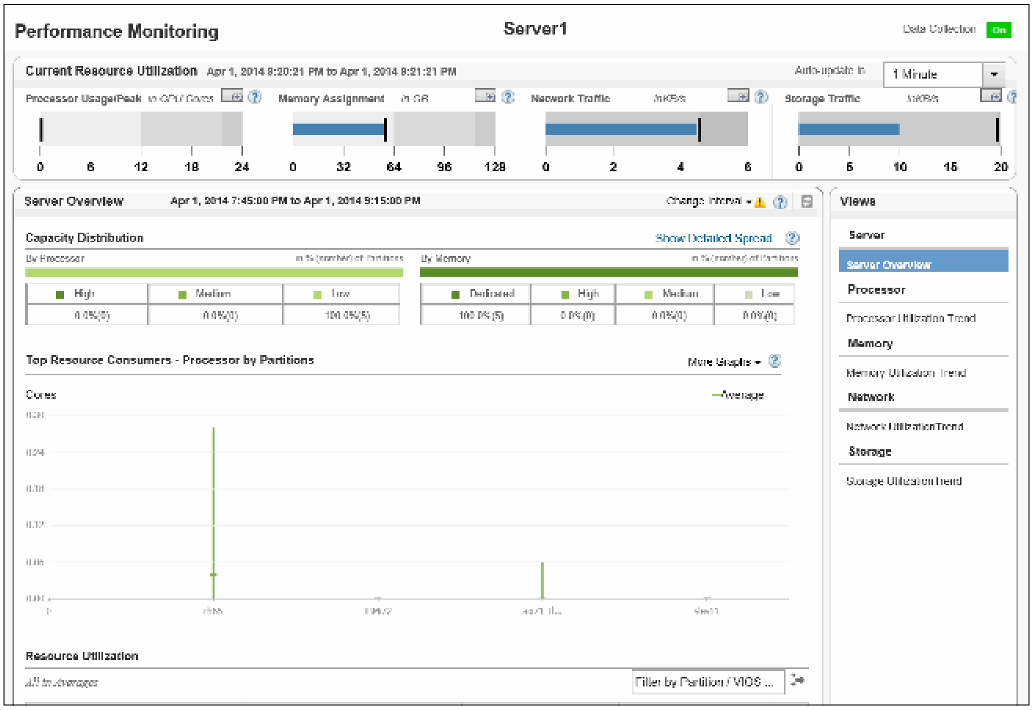
Templates allow you to specify the configuration details for the system I/O, memory, storage, network, processor, and other partition resources. A user can use the predefined or captured templates that are available in the template library to deploy a system. Two types of templates are available in the template library:

- The System Template contains configuration details for system resources, such as system I/O, memory, storage, processors, network, and physical I/O adapters. You can use these system templates to deploy the system.
- The Partition Template contains configuration details for the partition resources. A user can configure logical partitions with the predefined templates or by creating a custom templates.

Using these options, a user can deploy a system, select a System Template, and click **Deploy**. After deploying a system, a user can choose a Partition Template from the library.

Performance

The Performance function opens the Performance and Capacity Monitoring window, as shown in Figure 3-12 and Figure 3-13 on page 147.

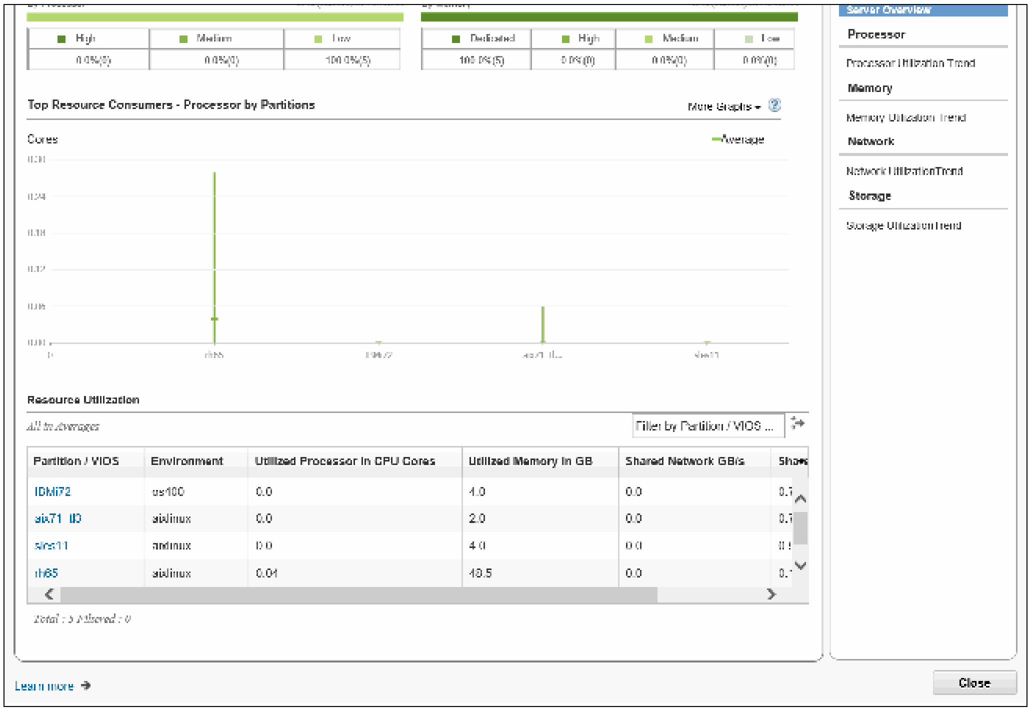


🟩 High	🟩 Medium	🟩 Low
0.0%(0)	0.0%(0)	100.0%(5)

🟩 Dedicated	🟩 High	🟩 Medium	🟩 Low
100.0% (5)	0.0% (0)	0.0%(0)	0.0%(0)

Figure 3-12 HMC performance monitoring CPU Memory assignment (top of the window)

Figure 3-13 shows the bottom of the window where performance and capacity data are presented in a graphical format.



🟩 High	🟩 Medium	🟩 Low
0.0%(0)	0.0%(0)	100.0%(5)

🟩 Dedicated	🟩 High	🟩 Medium	🟩 Low
100.0%(5)	0.0%(0)	0.0%(0)	0.0%(0)

Resource Utilization

All in Averages

Filter by Partition / VIOS ...

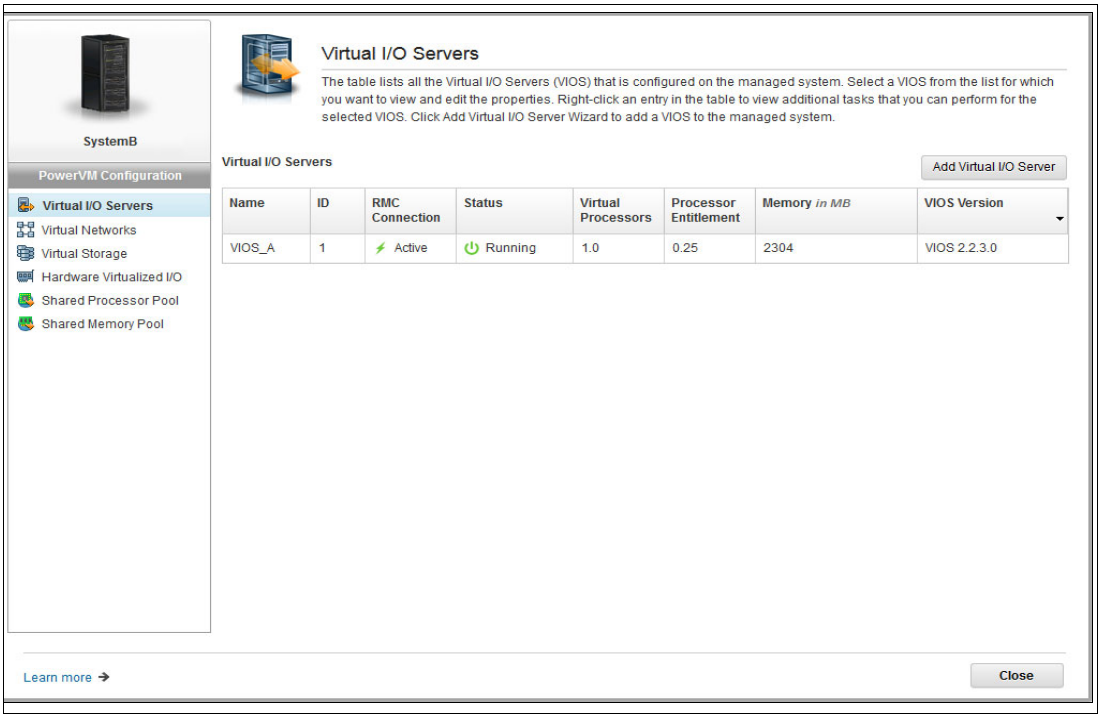
Partition / VIOS	Environment	Utilized Processor in CPU Cores	Utilized Memory in GB	Shared Network GB/s	Sha
IBMi72	os400	0.0	4.0	0.0	0.7
aix71_ID	aixlinux	0.0	2.0	0.0	0.7
sles11	aixlinux	0.0	4.0	0.0	0.5
rh65	aixlinux	0.01	40.5	0.0	0.1

Total : 5 Filtered : 0

Figure 3-13 HMC performance monitoring CPU Memory assignment (bottom of the window)

New GUI functions

As part of the enhancements introduced in this version, this new capability aids in the management of PowerVM from a single UI. Figure 3-14 shows how all virtualization components are now displayed in a single UI.



Virtual I/O Servers

The table lists all the Virtual I/O Servers (VIOS) that is configured on the managed system. Select a VIOS from the list for which you want to view and edit the properties. Right-click an entry in the table to view additional tasks that you can perform for the selected VIOS. Click Add Virtual I/O Server Wizard to add a VIOS to the managed system.

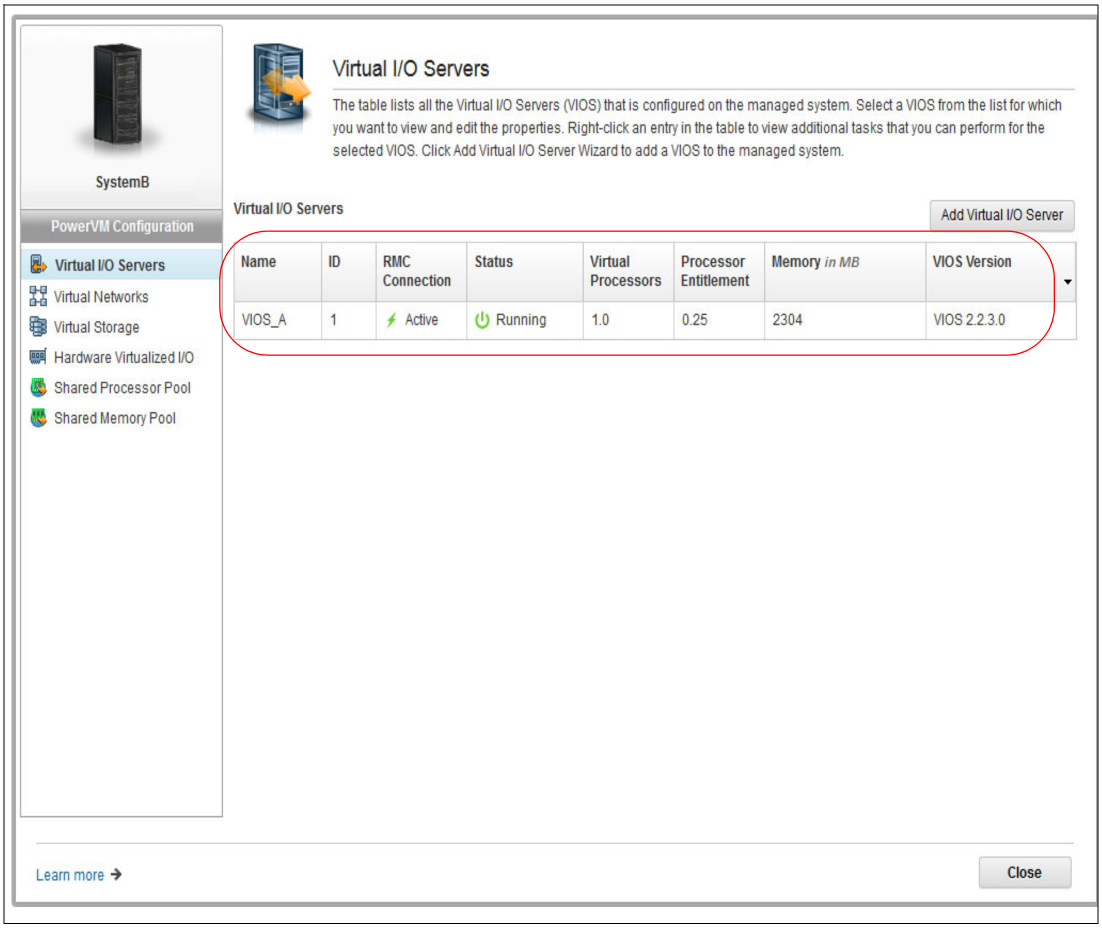
Virtual I/O Servers

Name	ID	RMC Connection	Status	Virtual Processors	Processor Entitlement	Memory *in MB*	VIOS Version
VIOS_A	1	Active	Running	1.0	0.25	2304	VIOS 2.2.3.0

Figure 3-14 VIO servers, Virtual Network, and Virtual Storage are accessible from the UI

3.5.11 Single view of all Virtual I/O servers

The *Manage Virtual I/O Servers* function displays a list of VIOS that are configured in the managed system. It also displays information about each VIOS configuration such as allocated memory, allocated processing units, allocated virtual processors, and status. This can be seen in Figure 3-15.



The table lists all the Virtual I/O Servers (VIOS) that is configured on the managed system. Select a VIOS from the list for which you want to view and edit the properties. Right-click an entry in the table to view additional tasks that you can perform for the selected VIOS. Click Add Virtual I/O Server Wizard to add a VIOS to the managed system.

Virtual I/O Servers

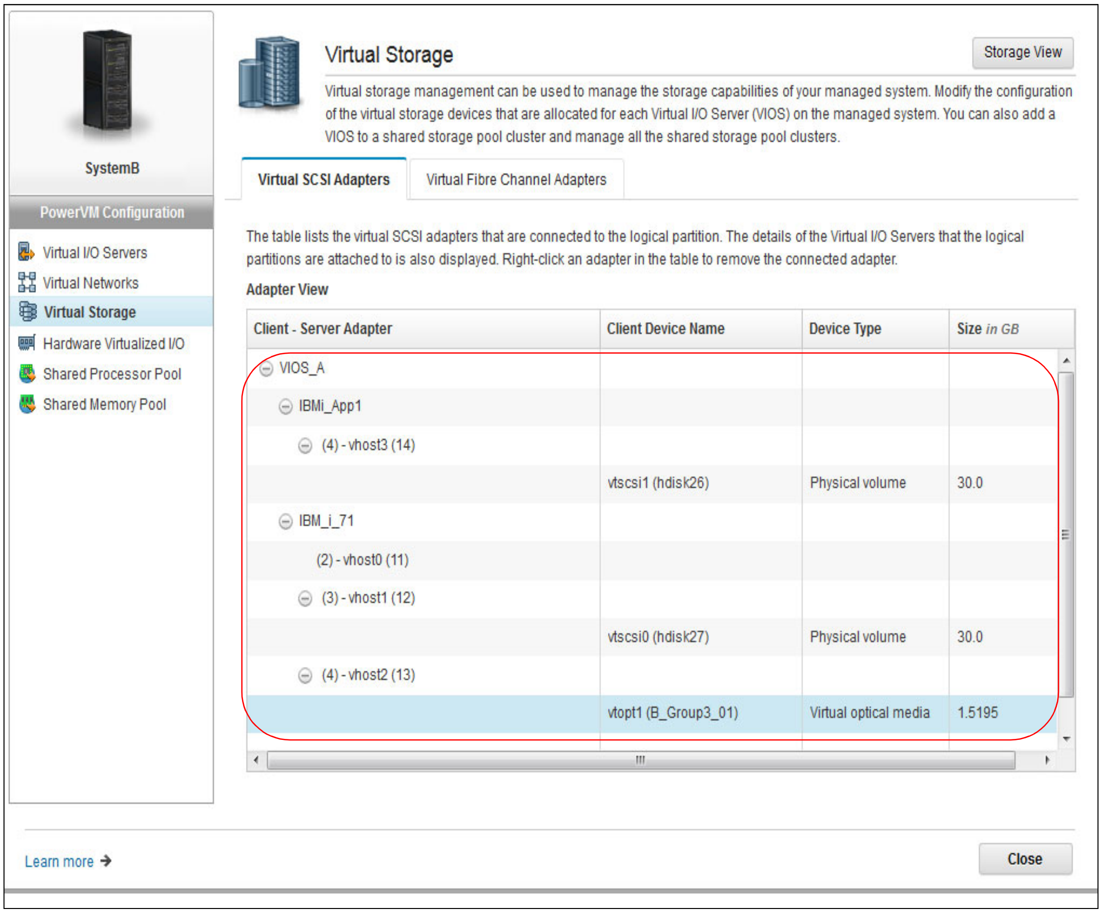
Name	ID	RMC Connection	Status	Virtual Processors	Processor Entitlement	Memory *in MB*	VIOS Version
VIOS_A	1	Active	Running	1.0	0.25	2304	VIOS 2.2.3.0

Figure 3-15 Virtual I/O server properties displayed in a single UI

3.5.12 Virtual Storage Management

HMC V8.8.1.0 allows you to manage and monitor storage devices in a PowerVM virtual storage environment. Is possible to change the configuration of the virtual storage devices that are allocated to each VIOS on the managed system. Also it lets you add a VIOS to a shared storage pool cluster and manage all the shared storage pool clusters.

The new UI lets the user view the adapter configuration of the virtual storage devices that are allocated to the VIOS. The adapter view provides a mapping of the adapters to the physical storage device. By selecting a VIOS, you can manage the virtual storage devices that are configured to a particular partition and select and view all the partitions with storage provisioned by the VIOS. Figure 3-16 shows a single VIOS scenario managing multiple Logical Partitions and its virtual SCSI devices.



Virtual Storage

Virtual storage management can be used to manage the storage capabilities of your managed system. Modify the configuration of the virtual storage devices that are allocated for each Virtual I/O Server (VIOS) on the managed system. You can also add a VIOS to a shared storage pool cluster and manage all the shared storage pool clusters.

The table lists the virtual SCSI adapters that are connected to the logical partition. The details of the Virtual I/O Servers that the logical partitions are attached to is also displayed. Right-click an adapter in the table to remove the connected adapter.

Adapter View

Client - Server Adapter	Client Device Name	Device Type	Size *in GB*
⊖ VIOS_A			
⊖ IBMi_App1			
⊖ (4) - vhost3 (14)			
	vtscsi1 (hdisk26)	Physical volume	30.0
⊖ IBM_i_71			
(2) - vhost0 (11)			
⊖ (3) - vhost1 (12)			
	vtscsi0 (hdisk27)	Physical volume	30.0
⊖ (4) - vhost2 (13)			
	vtopt1 (B_Group3_01)	Virtual optical media	1.5195

Figure 3-16 Adapter assignment to each partition

3.5.13 Virtual Network Management

HMC V8.8.1.0.1 helps to manage PowerVM virtual networks through a user interface (UI). This UI uses a defined set of concepts about networking technologies with specific terminology introduced by IBM Power Architecture®.

As part of Virtual Network Management, new functions are added to perform actions on the virtual network devices such as these:

- Add Virtual Network wizard
- Network Bridge Management
- Link Aggregation Management
- Power VM networking concepts review

Add Virtual Network wizard

Use the **Add Virtual Network** wizard button in the HMC to add an existing virtual network or a new virtual network to the server.

The following tasks can be completed by using the Add Virtual Network wizard:

- Create internal or bridged networks
- Create tagged or untagged virtual networks
- Create a virtual network on an existing or a new virtual switch
- Create a load group or select an existing load group

Network Bridge Management

From a server that is managed by the HMC, it is possible to change the PowerVM virtual network bridge properties, the changes allowed are:

- Enable or disable network failover in the **Failover** field.
- Enable or disable load balancing in the **Load Balance** field.
- Change the primary Virtual I/O Server (VIOS) and the physical adapter location from the table.
- Enable Jumbo Frame on the network bridge for the virtual Ethernet adapter to communicate to an external network.
- Enable QoS on the network bridge to check the priority value of all tagged packets and arrange those packets in the corresponding queue.

Link Aggregation Management

A link aggregation device can be added on the VIOS by using the Add Link Aggregation device wizard. The same wizard can be used to change a Link Aggregation device's properties or remove a Link Aggregation device.

PowerVM networking concepts review

PowerVM includes extensive and powerful networking tools and technologies, which can be used to enable more flexibility, better security, and enhanced usage of hardware resources. Some of these terms and concepts are unique to the Power Architecture. Table 3-4 introduces the PowerVM virtual networking technologies.

Table 3-4 PowerVM Network technologies

PowerVM technology	Definition
Virtual Network	Enables interpartition communication without assigning a physical network adapter to each partition. If the virtual network is bridged, partitions can communicate with external networks. A virtual network is identified by its name or VLAN ID and the associated virtual switch.
Virtual Ethernet adapter	Enables a client logical partition to send and receive network traffic without a physical Ethernet adapter.
Virtual switch	An in-memory, hypervisor implementation of a layer-2 switch.
Network bridge	A software adapter that bridges physical and virtual networks to enable communication. A network bridge can be configured for failover or load sharing.
Link aggregation device	A link aggregation (also known as Etherchannel) device is a network port-aggregation technology that allows several Ethernet adapters to be aggregated.

For more information regarding PowerVM Simplification Enhancements, see *IBM Power Systems Hardware Management Console Version 8 Release 8.1.0 Enhancements*, SG24-8232.

3.6 System Planning Tool

The IBM System Planning Tool (SPT) helps you design systems to be partitioned with logical partitions. You can also plan for and design non-partitioned systems by using the SPT. The resulting output of your design is called a *system plan*, which is stored in a `.sysplan` file. This file can contain plans for a single system or multiple systems. The `.sysplan` file can be used for the following reasons:

- To create reports
- As input to the IBM configuration tool (e-Config)
- To create and deploy partitions on your system (or systems) automatically

System plans that are generated by the SPT can be deployed on the system by the HMC.

> **Automatically deploy:** Ask your IBM SSR or IBM Business Partner to use the Customer Specified Placement manufacturing option if you want to automatically deploy your partitioning environment on a new machine. SPT looks for the resource's allocation that is the same as what is specified in your `.sysplan` file.

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

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

Integration between the System Planning Tool and both the Workload Estimator and IBM Performance Management allows you to create a system that is based on performance and capacity data from an existing system or that is based on new workloads that you specify.

You can use the SPT before you order a system to determine what you must order to support your workload. You can also use the SPT to determine how you can partition a system that you have.

Using the SPT is an effective way of documenting and backing up key system settings and partition definitions. With it, the user can create records of systems and export them to their personal workstation or backup system of choice. These same backups can then be imported back on to the same managed console when needed. This step can be useful when cloning systems, enabling the user to import the system plan to any managed console multiple times.

The SPT and its supporting documentation can be found at the IBM System Planning Tool website:

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

3.7 IBM Power Virtualization Center

IBM Power Virtualization Center (IBM PowerVC) is designed to simplify the management of virtual resources in your Power Systems environment.

After the product code is loaded, the IBM PowerVC no-menus interface guides you through three simple configuration steps to register physical hosts, storage providers, and network resources, and start capturing and intelligently deploying your virtual machines (VMs), among the other tasks shown in the following list:

- Create virtual machines and then resize and attach volumes to them.
- Import existing virtual machines and volumes so they can be managed by IBM PowerVC.
- Monitor the usage of the resources that are in your environment.
- Migrate virtual machines while they are running (hot migration).
- Deploy images quickly to create virtual machines that meet the demands of your ever-changing business needs.

IBM PowerVC is built on OpenStack. OpenStack is an open source software that controls large pools of server, storage, and networking resources throughout a data center.

IBM PowerVC is available in two editions:

- IBM Power Virtualization Center Express Edition
- IBM Power Virtualization Center Standard Edition

Table 3-5 shows an overview of the key features that are included with IBM PowerVC Editions.

Table 3-5 IBM PowerVC Editions key features overview.

IBM Power Virtualization Center Express Edition	IBM Power Virtualization Center Standard Edition
► Supports IBM Power Systems hosts that are managed by the Integrated Virtualization Manager (IVM) ► Supports storage area networks, local storage, and a combination in the same environment ► Supports a single VIOS virtual machine on each host	► Supports IBM Power Systems hosts that are managed by an HMC ► Supports storage area networks ► Supports multiple VIOSs virtual machines on each host

For more information about IBM PowerVC, see *IBM PowerVC Version 1.2 Introduction and Configuration*, SG24-8199.

3.8 IBM Power Virtualization Performance

IBM Power Virtualization Performance (IBM PowerVP) for Power Systems is a new product that offers a performance view into an IBM PowerVM virtualized environment running on the latest firmware of IBM Power Systems. It can show which virtual workloads are using specific physical resources on an IBM Power Systems server.

IBM PowerVP helps reduce time and complexity to find and display performance bottlenecks through a simple dashboard that shows the performance health of the system. It can help simplify both prevention and troubleshooting and thus reduce the cost of performance management.

It assists you in the following ways:

- Shows workloads in real time, which highlights possible problems or bottlenecks (overcommitted resources)
- Helps better use virtualized IBM Power System servers by showing the distribution of workloads
- Can replay saved historical data
- Helps with the resolution of performance-related issues
- Helps to proactively address future issues that can affect performance

IBM PowerVP is integrated with the POWER Hypervisor and collects performance data directly from PowerVM Hypervisor, which offers the most accurate performance information about virtual machines running on IBM Power Systems. This performance information is displayed on a real-time, continuous GUI dashboard and it is also available for historical review.

Features of IBM PowerVP include these:

- Real-time, continuous graphical monitor (dashboard) that delivers an easy-to-read display showing the overall performance health of the Power Systems server.
- Customizable performance thresholds that enable you to customize the dashboard to match your monitoring requirements.
- Historical statistics that enable you to go back in time and replay performance data sequences to discover performance bottlenecks.
- System-level performance views that show all LPARs (VMs) and how they are using real system resources.
- Virtual machine drill down, which gives you more performance details for each VM, displaying detailed information about various resources, such as CPU, memory, and disk activity.
- Support for all virtual machine types, including AIX, IBM i, and Linux.
- Background data collection, which enables performance data to be collected when the GUI is not active.

IBM PowerVP even allows an administrator to drill down and view specific adapter, bus, or CPU usage. An administrator can see the hardware adapters and how much workload is placed on them. IBM PowerVP provides both an overall and detailed view of IBM Power System server hardware, so it is easy to see how virtual machines are consuming resources. For more information about this topic, see the following website:

http://www.ibm.com/systems/power/software/performance/

3.9 VIOS 2.2.3.51 features

The enterprise systems require the VIOS version 2.2.3.51 with APAR IV68443 and IV68444. This release provides the same functionality as the previous release and is updated to support the Power E870 and Power E880 and other new hardware.

VIOS 2.2.3.51 features include these:

- Simplified Shared Ethernet Adapter Failover configuration setup
- Shared Storage Pools enhancements.
- With VIOS version 2.2.3.51, or later, the maximum number of virtual I/O slots that are supported on the AIX, IBM i, and Linux partition is increased up to 32.
- VIOS support:
  - Supports IBM Power E870 and Power E880
  - Supports 8 GB quad port PCIe generation 2 Fibre Channels adapter
  - Supports PCIe Gen3 Fibre Channels adapter
- There is support for Live Partition Mobility performance enhancements to better use the 10 Gb Ethernet Adapters that are used within the mobility process, and PowerVM server evacuation function.
- Shared Ethernet Adapter by default uses the largesend attribute.

The VIOS update packages can be downloaded from the IBM Fix Central website:

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

3.10 Dynamic Partition Remote Restart

Partition Remote Restart is a function designed to enhance availability of a partition on another server when its original host server fails. This is a High Availability function of PowerVM Enterprise Edition.

Starting from IBM Power Systems HMC V8.8.1.0, the requirement of enabling Remote Restart of an LPAR only at creation time has been removed. Dynamic Partition Remote Restart allows for the dynamic toggle of Remote Restart capability when an LPAR is deactivated.

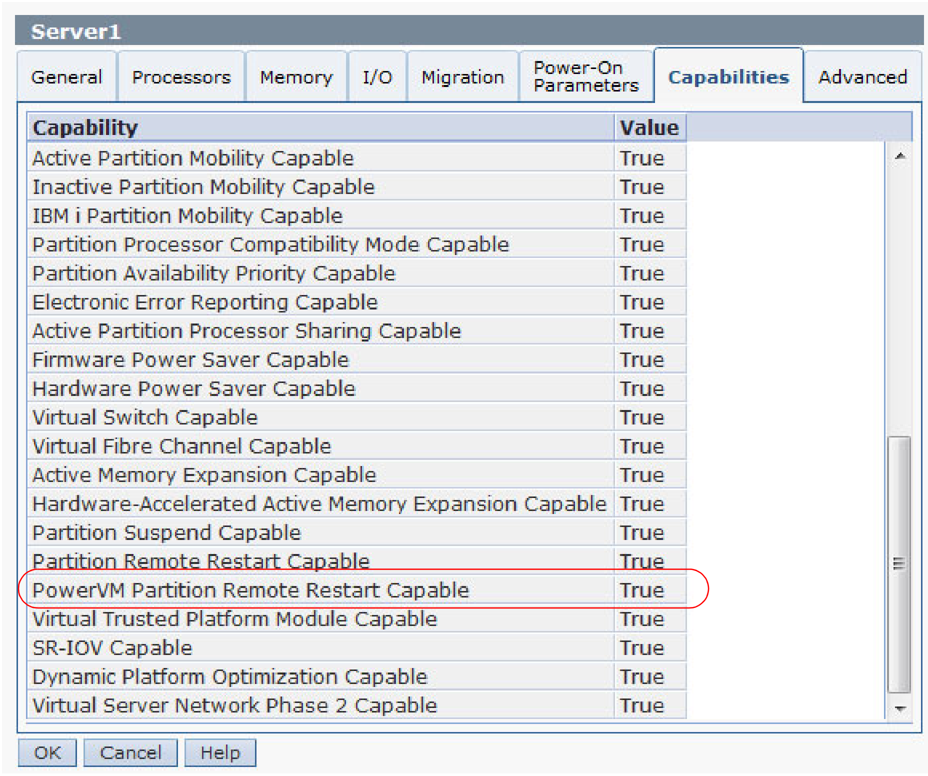
To verify that your managed system can support this capability, enter the following command as shown in Example 3-1.

Example 3-1 Power VM remote restart capable

```
hscroot@slcb27a:~>lssyscfg -r sys -m Server1 -F capabilities
"active_lpar_mobility_capable,inactive_lpar_mobility_capable,os400_lpar_mobility_c
apable,active_lpar_share_idle_procs_capable,active_mem_dedup_capable,active_mem_ex
pansion_capable,hardware_active_mem_expansion_capable,active_mem_mirroring_hypervi
sor_capable,active_mem_sharing_capable,autorecovery_power_on_capable,bsr_capable,c
od_mem_capable,cod_proc_capable,custom_mac_addr_capable,dynamic_platform_optimizat
ion_capable,dynamic_platform_optimization_lpar_score_capable,electronic_err_report
ing_capable,firmware_power_saver_capable,hardware_power_saver_capable,hardware_dis
covery_capable,hardware_encryption_capable,hca_capable,huge_page_mem_capable,lpar_
affinity_group_capable,lpar_avail_priority_capable,lpar_proc_compat_mode_capable,l
par_remote_restart_capable,powervm_lpar_remote_restart_capable,lpar_suspend_capabl
e,os400_lpar_suspend_capable,micro_lpar_capable,os400_capable,5250_application_cap
able,os400_net_install_capable,os400_restricted_io_mode_capable,redundant_err_path
_reporting_capable,shared_eth_auto_control_channel_capable,shared_eth_failover_cap
able,sp_failover_capable,sriov_capable,vet_activation_capable,virtual_eth_disable_
capable,virtual_eth_dlpar_capable,virtual_eth_qos_capable,virtual_fc_capable,virtu
al_io_server_capable,virtual_switch_capable,vlan_stat_capable,vtpm_capable,vsi_on_
veth_capable,vsn_phase2_capable"
```

In the previous example, the highlighted text is indicating that the managed system is capable of remotely restarting a partition.

From the HMC, select the Managed System **Properties** → **Capabilities** tab to display all of the managed system capabilities as shown in Figure 3-17.



Capability	Value
Active Partition Mobility Capable	True
Inactive Partition Mobility Capable	True
IBM i Partition Mobility Capable	True
Partition Processor Compatibility Mode Capable	True
Partition Availability Priority Capable	True
Electronic Error Reporting Capable	True
Active Partition Processor Sharing Capable	True
Firmware Power Saver Capable	True
Hardware Power Saver Capable	True
Virtual Switch Capable	True
Virtual Fibre Channel Capable	True
Active Memory Expansion Capable	True
Hardware-Accelerated Active Memory Expansion Capable	True
Partition Suspend Capable	True
Partition Remote Restart Capable	True
PowerVM Partition Remote Restart Capable	True
Virtual Trusted Platform Module Capable	True
SR-IOV Capable	True
Dynamic Platform Optimization Capable	True
Virtual Server Network Phase 2 Capable	True

Figure 3-17 PowerVM Partition Remote Restart Capable

The capability is only displayed if the managed system supports it.

To activate a partition on a supported system to support Dynamic Partition Remote Restart, enter the following command:

```
chsyscfg -r lpar -m <ManagedSystemName> -i
"name=<PartitionName>,remote_restart_capable=1"
```

To use the Remote Restart feature, the following conditions need to be met:

- The managed system should support *toggle partition remote capability.*
- The partition should be in the inactive state.
- The partition type should be AIX, IBM i, or Linux.
- The reserved storage device pool exists.
- The partition should not own any of the below resources/settings:
  - BSR
  - Time Reference Partition
  - Service Partition
  - Opticonnect
  - HSL
  - Physical I/O
  - HEA
  - Error Reporting Partition
  - Is part of EWLM
  - Huge Page Allocation
  - Owns Virtual Serial Adapters
  - Belongs to I/O Fail Over Pool
  - SR-IOV logical port

For further information. including the use of Partition Remote Restart, see the following website:

http://www.ibm.com/support/knowledgecenter/POWER8/p8hat/p8hat_enadisremres.htm

4

Reliability, availability, and serviceability

This chapter provides information about IBM Power Systems reliability, availability, and serviceability (RAS) design and features.

The elements of RAS can be described as follows:

Reliability Indicates how infrequently a defect or fault in a server occurs

Availability Indicates how infrequently the functioning of a system or application is impacted by a fault or defect

Serviceability Indicates how well faults and their effects are communicated to system managers and how efficiently and nondisruptively the faults are repaired

4.1 Introduction

The POWER8 processor-based servers are available in two different classes:

- Scale-out systems: For environments consisting of multiple systems working in concert. In such environments, application availability is enhanced by the superior availability characteristics of each system.
- Enterprise systems: For environments requiring systems with increased availability. In such environments, mission-critical applications can take full advantage of the scale-up characteristics, increased performance, flexibility to upgrade, and enterprise availability characteristics.

One key differentiator of the IBM POWER8 processor-based servers is that they leverage all the advanced RAS characteristics of the POWER8 processor through the whole portfolio, offering reliability and availability features that often are not seen in other scale-out servers. Some of these features are improvements for POWER8 or features that were found previously only in higher-end Power Systems.

The POWER8 processor modules support an enterprise level of reliability and availability. The processor design has extensive error detection and fault isolation (ED/FI) capabilities to allow for a precise analysis of faults, whether they are hard or soft. They use advanced technology, including stacked latches and Silicon-on-Insulator (SOI) technology, to reduce susceptibility to soft errors, and advanced design features within the processor for correction or try again after soft error events. In addition, the design incorporates spare capacity that is integrated into many elements to tolerate certain faults without requiring an outage or parts replacement. Advanced availability techniques are used to mitigate the impact of other faults that are not directly correctable in the hardware.

Features within the processor and throughout systems are incorporated to support design verification. During the design and development process, subsystems go through rigorous verification and integration testing processes by using these features. During system manufacturing, systems go through a thorough testing process to help ensure high product quality levels, again taking advantage of the designed ED/FI capabilities.

Fault isolation and recovery of the POWER8 processor and memory subsystems are designed to use a dedicated service processor and are meant to be largely independent of any operating system or application deployed.

The Power E870 and Power E880 are enterprise class systems and are designed to support the highest levels of RAS. More details of the specific features exclusive to enterprise class systems can be found in 4.4, “Enterprise systems availability details” on page 170.

4.1.1 RAS enhancements of POWER8 processor-based servers

Several features were included in the whole portfolio of the POWER8 processor-based servers. Some of these features are improvements for POWER8 or features that were found previously only in higher-end Power Systems, leveraging a higher RAS even for scale-out equipment.

Here is a brief summary of these features:

- Processor Enhancements Integration

  POWER8 processor chips are implemented using 22 nm technology and integrated onto SOI modules.

The processor design now supports a spare data lane on each fabric bus, which is used to communicate between processor modules. A spare data lane can be substituted for a failing one dynamically during system operation.

A POWER8 processor module has improved performance compared to POWER7+, including support of a maximum of twelve cores compared to a maximum of eight cores in POWER7+. Doing more work with less hardware in a system provides greater reliability, by concentrating the processing power and reducing the need for additional communication fabrics and components.

The processor module integrates a new On Chip Controller (OCC). This OCC is used to handle Power Management and Thermal Monitoring without the need for a separate controller, which was required in POWER7+. In addition, the OCC can also be programmed to run other RAS-related functions independent of any host processor.

The memory controller within the processor is redesigned. From a RAS standpoint, the ability to use a replay buffer to recover from soft errors is added.

- I/O Subsystem

  The POWER8 processor now integrates PCIe controllers. PCIe slots that are directly driven by PCIe controllers can be used to support I/O adapters directly in the systems or be used to attach external I/O drawers. For greater I/O capacity, the POWER8 processor-based Power E870 and Power E880 servers also support a PCIe switch to provide additional integrated I/O capacity.

- Memory Subsystem

  Custom DIMMs (CDIMMS) are used, which, in addition to the ability to correct a single DRAM fault within an error-correcting code (ECC) word (and then an additional bit fault) to avoid unplanned outages, also contain a spare DRAM module per port (per nine DRAMs for x8 DIMMs), which can be used to avoid replacing memory.

  The Power E870 and Power E880 systems have the option of mirroring the memory used by the hypervisor. This reduces the risk of system outage linked to memory faults, as the hypervisor memory is stored in two distinct memory CDIMMs.

- Power Distribution and Temperature Monitoring

  All systems make use of voltage convertors that transform the voltage level provided by the power supply to the voltage level needed for the various components within the system. The Power E870 and Power E880 systems contain two convertors for each voltage level provided to any given processor or memory DIMM.

  Convertors used for processor voltage levels are configured for redundancy so that when one is detected as failing, it will be called out for repair while the system continues to run with the redundant voltage convertor.

  The convertors used for memory are configured with a form of sparing where when a convertor fails, the system continues operation with another convertor without generating a service event or the need to take any sort of outage for repair.

  The processor module integrates a new On Chip Controller (OCC). This OCC is used to handle Power Management and Thermal Monitoring without the need for a separate controller, as was required in POWER7+. In addition, the OCC can also be programmed to run other RAS-related functions independent of any host processor.

  The E880 and E870 systems make use of triple redundant ambient temperature sensors.

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, and architecture-based approach is designed to improve overall system reliability with each successive generation of system offerings. Reliability can be improved in primarily three ways:

- Reducing the number of components
- Using higher reliability grade parts
- Reducing the stress on the components

In the POWER8 systems, elements of all three are used to improve system reliability.

4.2.1 Designed for reliability

Systems that are 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. The POWER8 chip has more cores per processor module, and the I/O Hub Controller function is integrated in the processor module, which generates a PCIe BUS directly from the Processor module. Parts selection also plays a critical role in overall system reliability.

IBM uses stringent design criteria to select server grade components that are extensively tested and qualified to meet and exceed a minimum design life of seven years. By selecting higher reliability grade components, the frequency of all failures is lowered, and wear-out is not expected within the operating system life. Component failure rates can be further improved by burning in select components or running the system before shipping it to the client. This period of high stress removes the weaker components with higher failure rates, that is, it cuts off the front end of the traditional failure rate bathtub curve (see Figure 4-1).

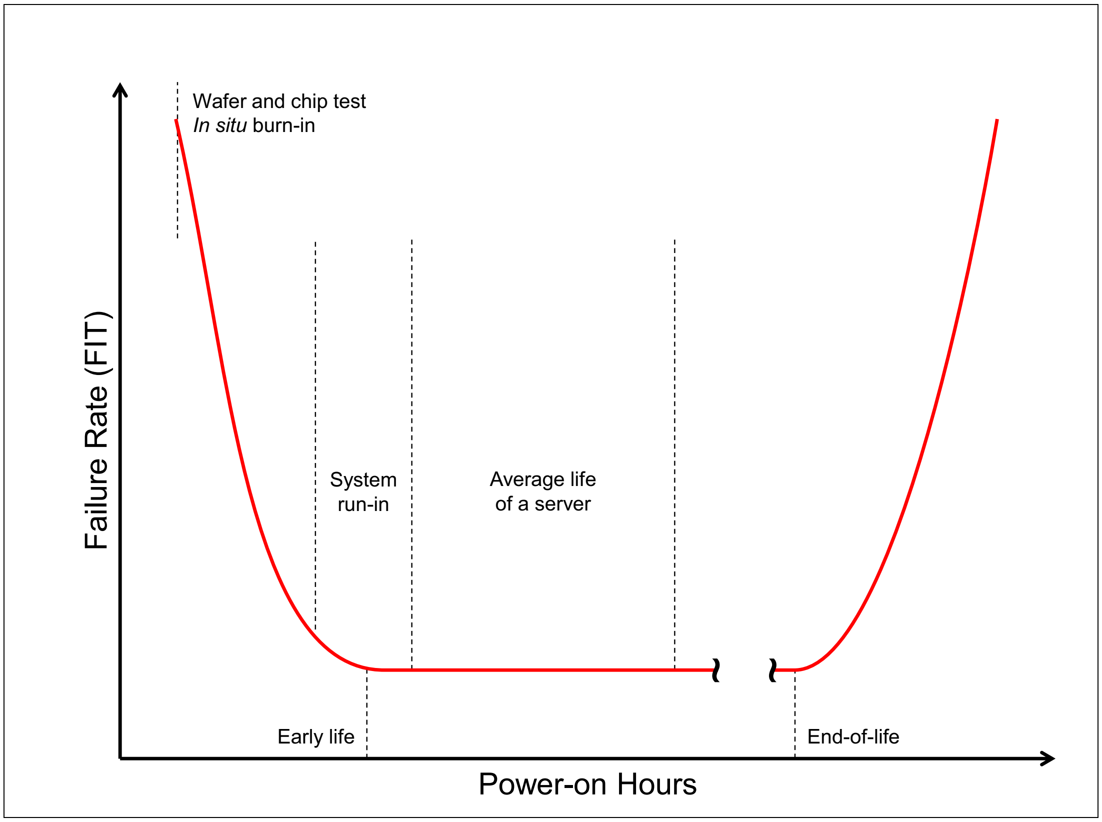


Figure 4-1 Failure rate bathtub curve

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. Large decreases in component reliability are directly correlated to relatively small increases in temperature. All POWER processor-based systems are packaged to ensure adequate cooling. Critical system components, such as the POWER8 processor chips, are positioned on the system board so that they receive clear air flow during operation. POWER8 systems use a premium fan with an extended life to further reduce overall system failure rate and provide adequate cooling for the critical system components.

4.3 Processor/Memory availability details

The more reliable a system or subsystem is, the more available it should be. Nevertheless, considerable effort is made to design systems that can detect faults that do occur and take steps to minimize or eliminate the outages that are associated with them. These design capabilities extend availability beyond what can be obtained through the underlying reliability of the hardware.

This design for availability begins with implementing an architecture for ED/FI.

First-Failure Data Capture (FFDC) is the capability of IBM hardware and microcode to continuously monitor hardware functions. Within the processor and memory subsystem, detailed monitoring is done by circuits within the hardware components themselves. Fault information is gathered into fault isolation registers (FIRs) and reported to the appropriate components for handling.

Processor and memory errors that are recoverable in nature are typically reported to the dedicated service processor built into each system. The dedicated service processor then works with the hardware to determine the course of action to be taken for each fault.

4.3.1 Correctable error introduction

Intermittent or soft errors are typically tolerated within the hardware design by using error correction code or advanced techniques to try operations again after a fault.

Tolerating a correctable solid fault runs the risk that the fault aligns with a soft error and causes an uncorrectable error situation. There is also the risk that a correctable error is predictive of a fault that continues to worsen over time, resulting in an uncorrectable error condition.

You can predictively deallocate a component to prevent correctable errors from aligning with soft errors or other hardware faults and causing uncorrectable errors to avoid such situations. However, unconfiguring components, such as processor cores or entire caches in memory, can reduce the performance or capacity of a system. This in turn typically requires that the failing hardware is replaced in the system. The resulting service action can also temporarily impact system availability.

To avoid such situations in solid faults in POWER8, processors or memory might be candidates for correction by using the “self-healing” features built into the hardware, such as taking advantage of a spare DRAM module within a memory DIMM, a spare data lane on a processor or memory bus, or spare capacity within a cache module.

When such self-healing is successful, the need to replace any hardware for a solid correctable fault is avoided. The ability to predictively unconfigure a processor core is still available for faults that cannot be repaired by self-healing techniques or because the sparing or self-healing capacity is exhausted.

4.3.2 Uncorrectable error introduction

An uncorrectable error can be defined as a fault that can cause incorrect instruction execution within logic functions, or an uncorrectable error in data that is stored in caches, registers, or other data structures. In less sophisticated designs, a detected uncorrectable error nearly always results in the termination of an entire system. More advanced system designs in some cases might be able to terminate just the application by using the hardware that failed. Such designs might require that uncorrectable errors are detected by the hardware and reported to software layers, and the software layers must then be responsible for determining how to minimize the impact of faults.

The advanced RAS features that are built in to POWER8 processor-based systems handle certain “uncorrectable” errors in ways that minimize the impact of the faults, even keeping an entire system up and running after experiencing such a failure.

Depending on the fault, such recovery may use the virtualization capabilities of PowerVM in such a way that the operating system or any applications that are running in the system are not impacted or must participate in the recovery.

4.3.3 Processor Core/Cache correctable error handling

Layer 2 (L2) and Layer 3 (L3) caches and directories can correct single bit errors and detect double bit errors (SEC/DED ECC). Soft errors that are detected in the level 1 caches are also correctable by a try again operation that is handled by the hardware. Internal and external processor “fabric” busses have SEC/DED ECC protection as well.

SEC/DED capabilities are also included in other data arrays that are not directly visible to customers.

Beyond soft error correction, the intent of the POWER8 design is to manage a solid correctable error in an L2 or L3 cache by using techniques to delete a cache line with a persistent issue, or to repair a column of an L3 cache dynamically by using spare capability.

Information about column and row repair operations is stored persistently for processors, so that more permanent repairs can be made during processor reinitialization (during system reboot, or individual Core Power on Reset using the Power On Reset Engine.)

4.3.4 Processor Instruction Retry and other try again techniques

Within the processor core, soft error events might occur that interfere with the various computation units. When such an event can be detected before a failing instruction is completed, the processor hardware might be able to try the operation again by using the advanced RAS feature that is known as *Processor Instruction Retry*.

Processor Instruction Retry allows the system to recover from soft faults that otherwise result in an outage of applications or the entire server.

Try again techniques are used in other parts of the system as well. Faults that are detected on the memory bus that connects processor memory controllers to DIMMs can be tried again. In POWER8 systems, the memory controller is designed with a replay buffer that allows memory transactions to be tried again after certain faults internal to the memory controller faults are detected. This complements the try again abilities of the memory buffer module.

4.3.5 Alternative processor recovery and Partition Availability Priority

If Processor Instruction Retry for a fault within a core occurs multiple times without success, the fault is considered to be a solid failure. In some instances, PowerVM can work with the processor hardware to migrate a workload running on the failing processor to a spare or alternative processor. This migration is accomplished by migrating the pertinent processor core state from one core to another with the new core taking over at the instruction that failed on the faulty core. Successful migration keeps the application running during the migration without needing to terminate the failing application.

Successful migration requires that there is sufficient spare capacity that is available to reduce the overall processing capacity within the system by one processor core. Typically, in highly virtualized environments, the requirements of partitions can be reduced to accomplish this task without any further impact to running applications.

In systems without sufficient reserve capacity, it might be necessary to terminate at least one partition to free resources for the migration. In advance, PowerVM users can identify which partitions have the highest priority and which do not. When you use this Partition Priority feature of PowerVM, if a partition must be terminated for alternative processor recovery to complete, the system can terminate lower priority partitions to keep the higher priority partitions up and running, even when an unrecoverable error occurred on a core running the highest priority workload.

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 rated at 255. The default value is set to 127 for standard partitions and 192 for Virtual I/O Server (VIOS) partitions. Priorities can be modified through the Hardware Management Console (HMC).

4.3.6 Core Contained Checkstops and other PowerVM error recovery

PowerVM can handle certain other hardware faults without terminating applications, such as an error in certain data structures (faults in translation tables or lookaside buffers).

Other core hardware faults that alternative processor recovery or Processor Instruction Retry cannot contain might be handled in PowerVM by a technique called Core Contained Checkstops. This technique allows PowerVM to be signaled when such faults occur and terminate code in use by the failing processor core (typically just a partition, although potentially PowerVM itself if the failing instruction were in a critical area of PowerVM code).

Processor designs without Processor Instruction Retry typically must resort to such techniques for all faults that can be contained to an instruction in a processor core.

4.3.7 Cache uncorrectable error handling

If a fault within a cache occurs that cannot be corrected with SEC/DED ECC, the faulty cache element is unconfigured from the system. Typically, this is done by purging and deleting a single cache line. Such purge and delete operations are contained within the hardware itself, and prevent a faulty cache line from being reused and causing multiple errors.

During the cache purge operation, the data that is stored in the cache line is corrected where possible. If correction is not possible, the associated cache line is marked with a special ECC code that indicates that the cache line itself has bad data.

Nothing within the system terminates just because such an event is encountered. Rather, the hardware monitors the usage of pages with marks. If such data is never used, hardware replacement is requested, but nothing terminates as a result of the operation. Software layers are not required to handle such faults.

Only when data is loaded to be processed by a processor core, or sent out to an I/O adapter, is any further action needed. In such cases, if data is used as owned by a partition, then the partition operating system might be responsible for terminating itself or just the program using the marked page. If data is owned by the hypervisor, then the hypervisor might choose to terminate, resulting in a system-wide outage.

However, the exposure to such events is minimized because cache-lines can be deleted, which eliminates repetition of an uncorrectable fault that is in a particular cache-line.

4.3.8 Other processor chip functions

Within a processor chip, there are other functions besides just processor cores.

POWER8 processors have built-in accelerators that can be used as application resources to handle such functions as random number generation. POWER8 also introduces a controller for attaching cache-coherent adapters that are external to the processor module. The POWER8 design contains a function to “freeze” the function that is associated with some of these elements, without taking a system-wide checkstop. Depending on the code using these features, a “freeze” event might be handled without an application or partition outage.

As indicated elsewhere, single bit errors, even solid faults, within internal or external processor “fabric busses”, are corrected by the error correction code that is used. POWER8 processor-to-processor module fabric busses also use a spare data-lane so that a single failure can be repaired without calling for the replacement of hardware.

4.3.9 Other fault error handling

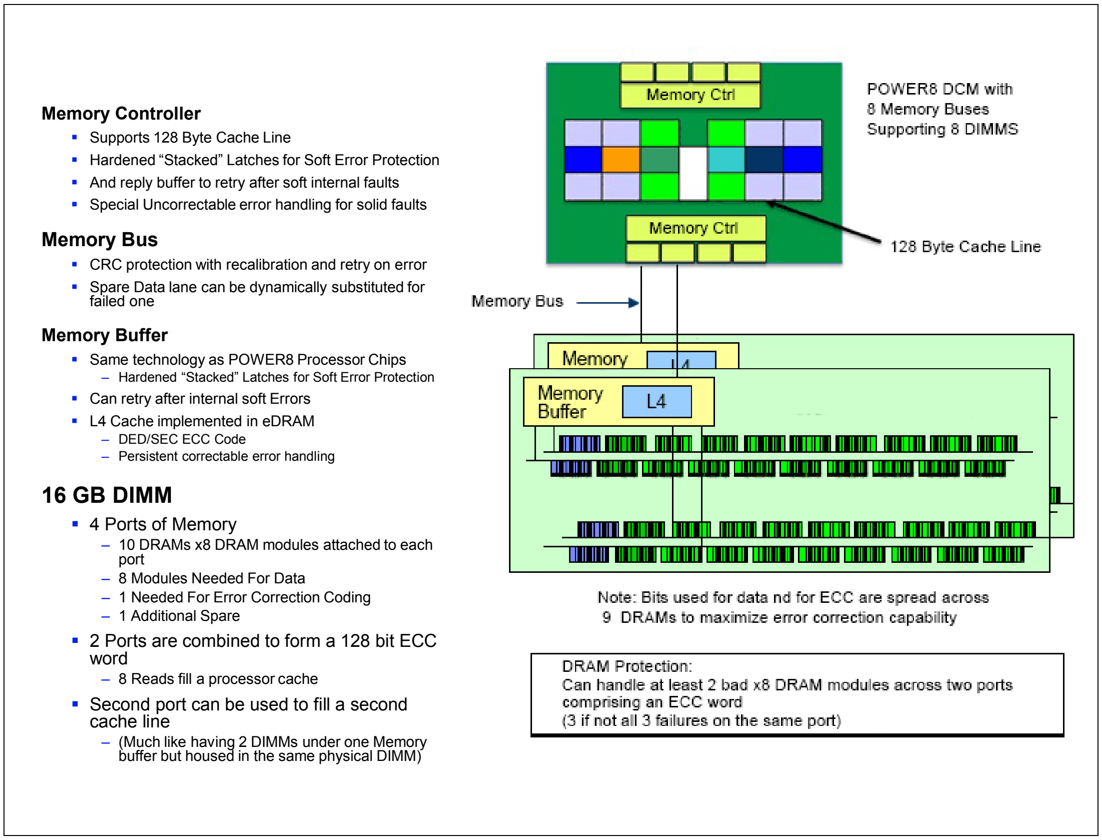
Not all processor module faults can be corrected by these techniques. Therefore, a provision is still made for some faults that cause a system-wide outage. In such a “platform” checkstop event, the ED/FI capabilities that are built in to the hardware and dedicated service processor work to isolate the root cause of the checkstop and unconfigure the faulty element were possible so that the system can reboot with the failed component unconfigured from the system.

The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced (AC power) failure.

The auto-restart (reboot) option must be enabled from the Advanced System Management Interface (ASMI) or from the Control (Operator) Panel.

4.3.10 Memory protection

POWER8 processor-based systems have a three-part memory subsystem design. This design consists of two memory controllers in each processor module, which communicate to buffer modules on memory DIMMS through memory channels and access the DRAM memory modules on DIMMs, as shown in Figure 4-2.



Memory Controller

- Supports 128 Byte Cache Line
- Hardened "Stacked" Latches for Soft Error Protection
- And reply buffer to retry after soft internal faults
- Special Uncorrectable error handling for solid faults

Memory Bus

- CRC protection with recalibration and retry on error
- Spare Data lane can be dynamically substituted for failed one

Memory Buffer

- Same technology as POWER8 Processor Chips
  - Hardened "Stacked" Latches for Soft Error Protection
- Can retry after internal soft Errors
- L4 Cache implemented in eDRAM
  - DED/SEC ECC Code
  - Persistent correctable error handling

16 GB DIMM

- 4 Ports of Memory
  - 10 DRAMs x8 DRAM modules attached to each port
  - 8 Modules Needed For Data
  - 1 Needed For Error Correction Coding
  - 1 Additional Spare
- 2 Ports are combined to form a 128 bit ECC word
  - 8 Reads fill a processor cache
- Second port can be used to fill a second cache line
  - (Much like having 2 DIMMs under one Memory buffer but housed in the same physical DIMM)

Note: Bits used for data nd for ECC are spread across
9 DRAMs to maximize error correction capability

DRAM Protection:
Can handle at least 2 bad x8 DRAM modules across two ports comprising an ECC word
(3 if not all 3 failures on the same port)

Figure 4-2 Memory protection features

The memory buffer chip is made by the same 22 nm technology that is used to make the POWER8 processor chip, and the memory buffer chip incorporates the same features in the technology to avoid soft errors. It implements a try again for many internally detected faults. This function complements a replay buffer in the memory controller in the processor, which also handles internally detected soft errors.

The bus between a processor memory controller and a DIMM uses CRC error detection that is coupled with the ability to try soft errors again. The bus features dynamic recalibration capabilities plus a spare data lane that can be substituted for a failing bus lane through the recalibration process.

The buffer module implements an integrated L4 cache using eDRAM technology (with soft error hardening) and persistent error handling features.

The memory buffer on each DIMM has four ports for communicating with DRAM modules. The 16 GB DIMM, for example, has one rank that is composed of four ports of x8 DRAM modules, each port containing 10 DRAM modules.

For each such port, there are eight DRAM modules worth of data (64 bits) plus another DRAM module's worth of error correction and other such data. There is also a spare DRAM module for each port that can be substituted for a failing port.

Two ports are combined into an ECC word and supply 128 bits of data. The ECC that is deployed can correct the result of an entire DRAM module that is faulty. This is also known as Chipkill correction. Then, it can correct at least an additional bit within the ECC word.

The additional spare DRAM modules are used so that when a DIMM experiences a Chipkill event within the DRAM modules under a port, the spare DRAM module can be substituted for a failing module, avoiding the need to replace the DIMM for a single Chipkill event.

Depending on how DRAM modules fail, it might be possible to tolerate up to four DRAM modules failing on a single DIMM without needing to replace the DIMM, and then still correct an additional DRAM module that is failing within the DIMM.

There are other DIMMs offered with these systems. A 32 GB DIMM has two ranks, where each rank is similar to the 16 GB DIMM with DRAM modules on four ports, and each port has ten x8 DRAM modules.

In addition, there is a 64 GB DIMM that is offered through x4 DRAM modules that are organized in four ranks.

In addition to the protection that is provided by the ECC and sparing capabilities, the memory subsystem also implements scrubbing of memory to identify and correct single bit soft-errors. Hypervisors are informed of incidents of single-cell persistent (hard) faults for deallocation of associated pages. However, because of the ECC and sparing capabilities that are used, such memory page deallocation is not relied upon for repair of faulty hardware,

Finally, should an uncorrectable error in data be encountered, the memory that is impacted is marked with a special uncorrectable error code and handled as described for cache uncorrectable errors.

4.3.11 I/O subsystem availability and Enhanced Error Handling

Usage of multi-path I/O and VIOS for I/O adapters and RAID for storage devices should be used to prevent application outages when I/O adapter faults occur.

To permit soft or intermittent faults to be recovered without failover to an alternative device or I/O path, Power Systems hardware supports *Enhanced Error Handling* (EEH) for I/O adapters and PCIe bus faults.

EEH allows EEH-aware device drivers to try again after certain non-fatal I/O events to avoid failover, especially in cases where a soft error is encountered. EEH also allows device drivers to terminate if there is an intermittent hard error or other unrecoverable errors, while protecting against reliance on data that cannot be corrected. This action typically is done by "freezing" access to the I/O subsystem with the fault. Freezing prevents data from flowing to and from an I/O adapter and causes the hardware/firmware to respond with a defined error signature whenever an attempt is made to access the device. If necessary, a special uncorrectable error code may be used to mark a section of data as bad when the freeze is first initiated.

In POWER8 processor-based systems, the external I/O hub and bridge adapters were eliminated in favor of a topology that integrates PCIe Host Bridges into the processor module itself. PCIe busses that are generated directly from a host bridge may drive individual I/O slots or a PCIe switch. The integrated PCIe controller supports try again (end-point error recovery) and freezing.

IBM device drivers under AIX are fully EEH-capable. For Linux under PowerVM, EEH support extends to many frequently used devices. There might be various third-party PCI devices that do not provide native EEH support.

4.4 Enterprise systems availability details

Besides all the standard RAS features described, Enterprise class systems allow for increased RAS and availability by including several features and redundant components.

Here is a list of the main features exclusive to the Enterprise class systems:

- Redundant Service Processor

  The service processor is an essential component of a system, responsible for is the initial power load (IPL), setup, monitoring, control and management. The control units, present on enterprise class systems house two redundant service processors. In case of a failure in either of the service processors, the second one allows for continued operation of the system until a replacement is scheduled. Even a system with a single system node had dual service processors in the system control unit.

- Redundant System Clock Cards

  Another component crucial to the system operations is the system clock cards. they are responsible to providing synchronized clock signals for the whole system. The control units, present on enterprise class systems house two redundant system clock cards. In case of a failure in any of the clock cards, the second one allows for continued operation of the system until a replacement is scheduled. Even a system with a single system node had dual clock cards on the system control unit.

- Dynamic Processor Sparing

  Enterprise class systems are Capacity Upgrade on Demand capable. Processor sparing helps minimize the impact to server performance caused by a failed processor. An inactive processor is activated if a failing processor reaches a predetermined error threshold, thus helping to maintain performance and improve system availability. Dynamic processor sparing happens dynamically and automatically when using dynamic logical partitioning (DLPAR) and the failing processor is detected prior to failure. Dynamic processor sparing does not require the purchase of an activation code; it requires only that the system have inactive CUoD processor cores available.

- Dynamic Memory Sparing

  Enterprise class systems are Capacity Upgrade on Demand capable. Dynamic memory sparing helps minimize the impact to server performance caused by a failed memory feature. Memory sparing occurs when on-demand inactive memory is automatically activated by the system to temporarily replace failed memory until a service action can be performed.

- Active Memory Mirroring for Hypervisor

  The hypervisor is the core part of the virtualization layer. Although minimal, its operational data must reside in memory CDIMMs. In case of a failure of CDIMM the hypervisor could become inoperative. The Active memory mirroring for hypervisor allows for the memory blocks used by the hypervisor to be written in two distinct CDIMMs. If an uncorrectable error is encountered during a read the data is retrieved from the mirrored pair and operations continue normally.

4.5 Availability impacts of a solution architecture

Any given solution should not rely only on the hardware platform. Despite IBM Power Systems being far superior RAS than other comparable systems, it is advisable to design a redundant architecture surrounding the application in order to allow for easier maintenance tasks and grater flexibility.

By working in a redundant architecture, some tasks that would require that a given application would be brought offline, can now be execute with the application running, allowing for even greater availability.

When determining a highly available architecture that fits your needs, the following topics are worth considering:

- Will I need to move my workloads off of an entire server during service or planned outages?
- If I use a clustering solution to move the workloads, how will the failover time affect my services?
- If I use a server evacuation solution to move the workloads, how long will it take to migrate all the partitions with my current server configuration?

4.5.1 Clustering

IBM Power Systems running under PowerVM and IBM i, AIX and Linux support a spectrum of clustering solutions. These solutions are designed to meet requirements not only for application availability in regard to server outages, but also data center disaster management, reliable data backups, and so forth. These offerings include distributed applications with IBM DB2® PureScale, HA solutions using clustering technology with IBM PowerHA SystemMirror®, and disaster management across geographies with PowerHA SystemMirror Enterprise Edition.

For more information, see the following references:

- *PowerHA SystemMirror for IBM i Cookbook*, SG24-7994:

  http://www.redbooks.ibm.com/abstracts/sg247994.html
- *Guide to IBM PowerHA SystemMirror for AIX Version 7.1.3, SG24-8167*

  http://www.redbooks.ibm.com/abstracts/sg248167.html
- *IBM PowerHA SystemMirror for AIX Cookbook, SG24-7739*

  http://www.redbooks.ibm.com/abstracts/sg247739.html

4.5.2 Virtual I/O redundancy configurations

Within each server, the partitions can be supported with a single VIOS. However, if a single VIOS is used and that VIOS terminates for any reason (hardware or software caused), then all the partitions using that VIOS will terminate.

Using Redundant VIOS servers would mitigate that risk. Maintaining redundancy of adapters within each VIOS, in addition to having redundant VIOS, will avoid most faults that keep a VIOS from running. Multiple paths to networks and SANs is therefore advised. Figure 4-3 shows a diagram of a partition accessing data from two distinct Virtual I/O Servers, each one with multiple network and SAN adapters to provide connectivity.

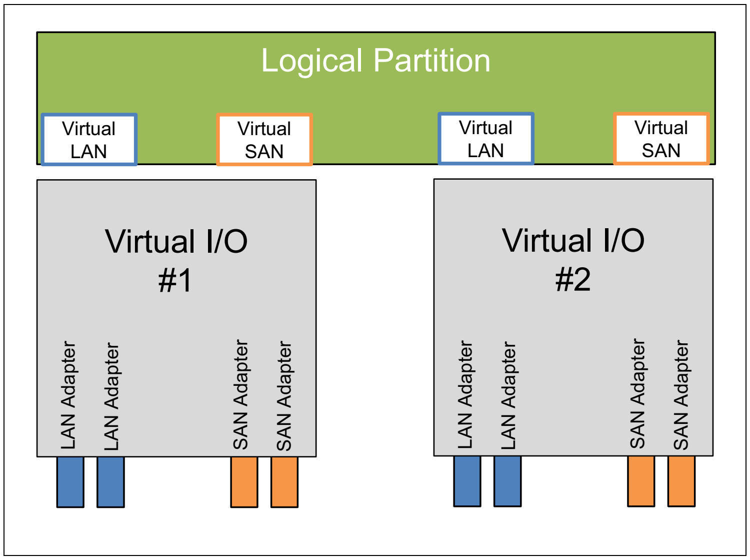


Figure 4-3 Partition with dual redundant Virtual I/O Servers

Since each VIOS can largely be considered as an AIX based partition, each VIOS also needs the ability to access a boot image, having paging space, and so forth under a root volume group or rootvg. The rootvg can be accessed through a SAN, the same as the data that partitions use. Alternatively, a VIOS can use storage locally attached to a server, either DASD devices or SSD drives. For best availability, however accessed, the rootvgs should use mirrored or RAID drives with redundant access to the devices.

4.5.3 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility allows you to move a running logical partition, including its operating system and running applications, from one system to another without any shutdown and without disrupting the operation of that logical partition. Inactive partition mobility allows you to move a powered-off logical partition from one system to another.

Live Partition Mobility provides systems management flexibility and improves system availability through the following functions:

- Avoid planned outages for hardware or firmware maintenance by moving logical partitions to another server and then performing the maintenance. Live Partition Mobility can help lead to zero downtime for maintenance because you can use it to work around scheduled maintenance activities.

- Avoid downtime for a server upgrade by moving logical partitions to another server and then performing the upgrade. This approach allows your users to continue their work without disruption.
- Avoid unplanned downtime. With preventive failure management, if a server indicates a potential failure, you can move its logical partitions to another server before the failure occurs. Partition mobility can help avoid unplanned downtime.
- Take advantage of server optimization:
  - Consolidation: You can consolidate workloads that run on several small, under utilized servers onto a single large server.
  - Deconsolidation: You can move workloads from server to server to optimize resource use and workload performance within your computing environment. With live partition mobility, you can manage workloads with minimal downtime.

Server Evacuation: This PowerVM function allows you to perform a server evacuation operation. Server Evacuation is used to move all migration-capable LPARs from one system to another if there are no active migrations in progress on the source or the target servers.

With the Server Evacuation feature, multiple migrations can occur based on the concurrency setting of the HMC. Migrations are performed as sets, with the next set of migrations starting when the previous set completes. Any upgrade or maintenance operations can be performed after all the partitions are migrated and the source system is powered off.

You can migrate all the migration-capable AIX, IBM i and Linux partitions from the source server to the destination server by running the following command from the HMC command line:

```
migrlpar -o m -m source_server -t target_server --all
```

Hardware and operating system requirements for Live Partition Mobility

Live Partition Mobility is supported by default with enterprise systems, and it is supported in compliance with all operating systems that are compatible with POWER8 technology.

The VIOS partition itself cannot be migrated.

For more information about Live Partition Mobility and how to implement it, see *IBM PowerVM Live Partition Mobility (Obsolete - See Abstract for Information)*, SG24-7460.

4.6 Serviceability

The purpose of serviceability is to repair the system while attempting to minimize or eliminate service cost (within budget objectives) and maintaining application availability and high customer satisfaction. Serviceability includes system installation, miscellaneous equipment specification (MES) (system upgrades/downgrades), and system maintenance/repair. Depending on the system and warranty contract, service might be performed by the customer, an IBM System Services Representative (SSR), or an authorized warranty service provider.

The serviceability features that are delivered in this system provide a highly efficient service environment by incorporating the following attributes:

- Design for SSR Set Up and Customer Installed Features (CIF).

- Detection and Fault Isolation (ED/FI).
- First Failure Data Capture (FFDC).
- Guiding Light service indicator architecture is used to control a system of integrated LEDs that lead the individual servicing the machine to the correct part as quickly as possible.
- Service labels, service cards, and service diagrams available on the system and delivered through the HMC.
- Step-by-step service procedures available through the HMC.

This section provides an overview of how these attributes contribute to efficient service in the progressive steps of error detection, analysis, reporting, notification, and repair found in all POWER processor-based systems.

4.6.1 Detecting errors

The first and most crucial component of a solid serviceability strategy is the ability to detect accurately and effectively errors when they occur.

Although not all errors are a guaranteed threat to system availability, those that go undetected can cause problems because the system has no opportunity to evaluate and act if necessary. POWER processor-based systems employ IBM z Systems® server-inspired error detection mechanisms, extending from processor cores and memory to power supplies and hard disk drives (HDDs).

4.6.2 Error checkers, fault isolation registers, and First-Failure Data Capture

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 that is coupled with Processor Instruction Retry and bus try again, to ECC correction on caches and system buses.

Within the processor/memory subsystem error-checker, error-checker signals are captured and stored in hardware FIRs. The associated logic circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, runtime error diagnostic tests can be deterministic so that for every check station, the unique error domain for that checker is defined and mapped to field-replaceable units (FRUs) that can be repaired when necessary.

Integral to the Power Systems design is the concept of FFDC. FFDC is a technique that involves sufficient error checking stations and co-ordination of faults so that faults are detected and the root cause of the fault is isolated. FFDC also expects that necessary fault information can be collected at the time of failure without needing re-create the problem or run an extended tracing or diagnostics program.

For the vast majority of faults, a good FFDC design means that the root cause is isolated at the time of the failure without intervention by an IBM SSR. For all faults, good FFDC design still makes failure information available to the IBM SSR, and this information can be used to confirm the automatic diagnosis. More detailed information can be collected by an IBM SSR for rare cases where the automatic diagnosis is not adequate for fault isolation.

4.6.3 Service processor

In POWER8 processor-based systems with a dedicated service processor, the dedicated service processor is primarily responsible for fault analysis of processor/memory errors.

The service processor is a microprocessor that is powered separately from the main instruction processing complex.

In addition to FFDC functions, the service processor performs many serviceability functions:

- Several remote power control options
- Reset and boot features
- Environmental monitoring

  The service processor interfaces with the OCC function, which monitors the server's built-in temperature sensors and sends instructions to the system fans to increase rotational speed when the ambient temperature is above the normal operating range. By using a designed operating system interface, the service processor notifies the operating system of potential environmentally related problems so that the system administrator can take appropriate corrective actions before a critical failure threshold is reached. The service processor can also post a warning and initiate an orderly system shutdown 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 shut down a system in the following circumstances:
    - The temperature exceeds the critical level or remains above the warning level for too long.
    - Internal component temperatures reach critical levels.
    - Non-redundant fan failures occur.
- POWER Hypervisor (system firmware) and HMC connection surveillance.

  The service processor monitors the operation of the firmware during the boot process, and also monitors the hypervisor for termination. The hypervisor monitors the service processor and can perform a reset and reload if it detects the loss of the service processor. If the reset/reload operation does not correct the problem with the service processor, the hypervisor notifies the operating system, and then the operating system can then take appropriate action, including calling for service. The FSP also monitors the connection to the HMC and can report loss of connectivity to the operating system partitions for system administrator notification.
- Uncorrectable error recovery

  The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced (AC power) failure.

  The auto-restart (reboot) option must be enabled from the ASMI or from the Control (Operator) Panel.
- Concurrent access to the service processors menus of the ASMI

  This access allows nondisruptive abilities to change system default parameters, interrogate service processor progress and error logs, and set and reset service indicators

(Light Path for low-end servers), and access all service processor functions without having to power down the system to the standby state. The administrator or IBM SSR dynamically can access the menus from any web browser-enabled console that is attached to the Ethernet service network, concurrently with normal system operation. Some options, such as changing the hypervisor type, do not take effect until the next boot.

- Management of the interfaces for connecting uninterruptible power source systems to the POWER processor-based systems and performing timed power-on (TPO) sequences.

4.6.4 Diagnosing

General diagnostic objectives are to detect and identify problems so that they can be resolved quickly. The IBM diagnostic strategy includes the following elements:

- Provide a common error code format that is equivalent to a system reference code, system reference number, checkpoint, or firmware error code.
- Provide fault detection and problem isolation procedures. Support a remote connection ability that is used by the IBM Remote Support Center or IBM Designated Service.
- Provide interactive intelligence within the diagnostic tests with detailed online failure information while connected to IBM back-end system.

Using the extensive network of advanced and complementary error detection logic that is built directly into hardware, firmware, and operating systems, the IBM Power Systems servers can perform considerable self-diagnosis.

Because of the FFDC technology that is designed in to IBM servers, re-creating diagnostic tests for failures or requiring user intervention is not necessary. Solid and intermittent errors are designed to be correctly detected and isolated at the time that the failure occurs. Runtime and boot time diagnostic tests fall into this category.

Boot time

When an IBM Power Systems server powers up, the service processor initializes the system hardware. Boot-time diagnostic testing uses a multi-tier approach for system validation, starting with managed low-level diagnostic tests that are supplemented with system firmware initialization and configuration of I/O hardware, followed by OS-initiated software test routines.

To minimize boot time, the system determines which of the diagnostic tests are required to be started to ensure correct operation, which is based on the way that the system was powered off, or on the boot-time selection menu.

Host Boot IPL

In POWER8, the initialization process during IPL changed. The Flexible Service Processor (FSP) is no longer the only instance that initializes and runs the boot process. With POWER8, the FSP initializes the boot processes, but on the POWER8 processor itself, one part of the firmware is running and performing the Central Electronics Complex chip initialization. A new component that is called the PNOR chip stores the Host Boot firmware and the Self Boot Engine (SBE) is an internal part of the POWER8 chip itself and is used to boot the chip.

With this Host Boot initialization, new progress codes are available. An example of an FSP progress code is C1009003. During the Host Boot IPL, progress codes, such as CC009344, appear.

If there is a failure during the Host Boot process, a new Host Boot System Dump is collected and stored. This type of memory dump includes Host Boot memory and is off-loaded to the HMC when it is available.

Run time

All Power Systems servers can monitor critical system components during run time, and they can take corrective actions when recoverable faults occur. The IBM hardware error-check architecture can report non-critical errors in the Central Electronics Complex in an *out-of-band* communications path to the service processor without affecting system performance.

A significant part of IBM runtime diagnostic capabilities originate with the service processor. Extensive diagnostic and fault analysis routines were developed and improved over many generations of POWER processor-based servers, and enable quick and accurate predefined responses to both actual and potential system problems.

The service processor correlates and processes runtime error information by using logic that is derived from IBM engineering expertise to count recoverable errors (called *thresholding*) and predict when corrective actions must be automatically initiated by the system. These actions can include the following items:

- 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, diagnostic tests are best performed by operating system-specific drivers, most notably adapters or I/O devices that are owned directly by a logical partition. In these cases, the operating system device driver often works 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. In non-HMC managed servers, the OS can start the Call Home application to report the service event to IBM. For optional HMC managed servers, the event is reported to the HMC, which can initiate the Call Home request to IBM. I/O devices can also include specific exercisers that can be started by the diagnostic facilities for problem recreation (if required by service procedures).

4.6.5 Reporting

In the unlikely event that a system hardware or environmentally induced failure is diagnosed, IBM Power Systems servers report the error through various 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.

Using the Call Home infrastructure, the system automatically can send an alert through a phone line to a pager, or call for service if there is a critical system failure. A hardware fault also illuminates the amber system fault LED, which is on the system unit, to alert the user of an internal hardware problem.

On POWER8 processor-based servers, hardware and software failures are recorded in the system log. When a management console is attached, an ELA routine analyzes the error, forwards the event to the Service Focal Point™ (SFP) application running on the management console, and can notify the system administrator that it isolated a likely cause of the system problem. The service processor event log also records unrecoverable checkstop conditions, forwards them to the SFP application, and notifies the system administrator.

After the information is logged in the SFP application, if the system is correctly configured, a Call Home service request is initiated and the pertinent failure data with service parts information and part locations is sent to the IBM service organization. This information also contains the client contact information as defined in the IBM Electronic Service Agent (ESA) guided setup wizard. With the new HMC V8R8.1.0 a Serviceable Event Manager is available to block problems from being automatically transferred to IBM. For more information about this topic, see “Service Event Manager” on page 193 for more details.

Error logging and analysis

When the root cause of an error is identified by a fault isolation component, an error log entry is created with basic data, such as the following examples:

- An error code that uniquely describes the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data such as engineering and manufacturing levels
- Return codes
- Resource identifiers
- FFDC data

Data that contains information about the effect that the repair has on the system is also included. Error log routines in the operating system and FSP can then use this information and decide whether the fault is a Call Home candidate. If the fault requires support intervention, a call is placed with service and support, and a notification is sent to the contact that is defined in the ESA-guided setup wizard.

Remote support

The Remote Management and Control (RMC) subsystem is delivered as part of the base operating system, including the operating system that runs on the HMC. RMC provides a secure transport mechanism across the LAN interface between the operating system and the optional HMC and is used by the operating system diagnostic application for transmitting error information. It performs several other functions, but they are not used for the service infrastructure.

Service Focal Point application for partitioned systems

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 be reported only once, regardless of how many logical partitions experience the potential effect of the error. The SFP application on the management console or in the Integrated Virtualization Manager (IVM) is responsible for aggregating duplicate error reports, and ensures that all errors are recorded for review and management. The SFP application provides other service-related functions, such as controlling service indicators, setting up Call Home, and providing guided maintenance.

When a local or globally reported service request is made to the operating system, the operating system diagnostic subsystem uses the RMC subsystem to relay error information to the optional HMC. For global events (platform unrecoverable errors, for example), the service processor also forwards error notification of these events to the HMC, providing a redundant error-reporting path in case there are errors in the RMC subsystem network.

The first occurrence of each failure type is recorded in the Manage Serviceable Events task on the management console. This task then filters and maintains a history of duplicate reports from other logical partitions or from the service processor. It then looks at all active service event requests within a predefined timespan, analyzes the failure to ascertain the root cause and, if enabled, initiates a Call Home for service. This methodology ensures that all platform errors are reported through at least one functional path, ultimately resulting in a single notification for a single problem. Similar service functionality is provided through the

SFP application on the IVM for providing service functions and interfaces on non-HMC partitioned servers.

Extended error data

Extended error data (EED) is additional data that is collected either automatically at the time of a failure or manually at a later time. The data that is collected depends 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 either to assist the service support organization with preparing a service action plan for the IBM SSR or for additional analysis.

System dump handling

In certain circumstances, an error might require a memory dump to be automatically or manually created. In this event, the memory dump may be off-loaded to the optional HMC. Specific management console information is included as part of the information that optionally can be sent to IBM Support for analysis. If additional information that relates to the memory dump is required, or if viewing the memory dump remotely becomes necessary, the management console memory dump record notifies the IBM Support center regarding on which managements console the memory dump is located. If no management console is present, the memory dump might be either on the FSP or in the operating system, depending on the type of memory dump that was initiated and whether the operating system is operational.

4.6.6 Notifying

After a Power Systems server detects, diagnoses, and reports an error to an appropriate aggregation point, it then takes steps to notify the client and, if necessary, the IBM Support organization. Depending on the assessed severity of the error and support agreement, this client notification might range from a simple notification to having field service personnel automatically dispatched to the client site with the correct replacement part.

Client Notify

When an event is important enough to report, but does not indicate the need for a repair action or the need to call home to IBM 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. These events include 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 (although these events do not necessarily require a part replacement or repair action)

Client Notify events are serviceable events because they indicate that something happened that requires client awareness if the client wants to take further action. These events can be reported to IBM at the discretion of the client.

Call Home

Call Home refers to an automatic or manual call from a customer location to an IBM Support structure with error log data, server status, or other service-related information. The Call

Home feature starts the service organization so that the appropriate service action can begin. Call Home can be done through HMC or most non-HMC managed systems.

Although configuring a Call Home function is optional, clients are encouraged to implement this feature to obtain service enhancements, such as reduced problem determination and faster and potentially more accurate transmission of error information. In general, using the Call Home feature can result in increased system availability. The ESA application can be configured for automated Call Home. For more information, see 4.7.4, "Electronic Services and Electronic Service Agent" on page 191.

Vital product data and inventory management

Power Systems store vital product data (VPD) internally, which keeps a record of how much memory is installed, how many processors are installed, the manufacturing level of the parts, and so on. These records provide valuable information that can be used by remote support and IBM SSRs, enabling the IBM SSRs to assist in keeping the firmware and software current on the server.

IBM Service and Support Problem Management database

At the IBM Support center, historical problem data is entered into the IBM Service and Support Problem Management database. All of the information that is related to the error, along with any service actions that are taken by the IBM SSR, is recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

4.6.7 Locating and servicing

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Packaging for service

The following service enhancements are included in the physical packaging of the systems to facilitate service:

- Color coding (touch points)

  Terracotta-colored touch points indicate that a component (FRU or CRU) can be concurrently maintained.

  Blue-colored touch points delineate components that may not be concurrently maintained (they might require that the system is 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 require basic tools such as flathead screw drivers, to service the hardware components.

- Positive retention

  Positive retention mechanisms help ensure proper connections between hardware components, such as from cables to connectors, and between two cards that attach to each other. Without positive retention, hardware components risk become loose during shipping or installation, which prevents 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 function is for scale-out systems that can be repaired by clients. In the Light Path LED implementation, when a fault condition is detected on the POWER8 processor-based system, an amber FRU fault LED is illuminated (turned on solid), which is then rolled up to the system fault LED. The Light Path system pinpoints the exact part by lighting the amber FRU fault LED that is associated with the part that must be replaced.

The service person can clearly identify components for replacement by using specific component level identify LEDs, and can also guide the IBM SSR directly to the component by signaling (flashing) the FRU component identify LED, and rolling up to the blue enclosure Locate LED.

After the repair, the LEDs shut off automatically when the problem is fixed. The Light Path LEDs are only visible while system is in standby power. There are two gold caps implemented. The gold cap is used to illuminate the amber LEDs once power is removed from the system. One is inside the drawer, to identify DIMMs, processors and VRMs and one is in the RAID assembly.

IBM KnowledgeCenter

IBM Knowledge Center provides you with a single information center where you can access product documentation for IBM systems hardware, operating systems, and server software.

The latest version of the documentation is accessible through the Internet; however, a CD-ROM based version is also available.

The purpose of KnowledgeCenter, in addition to providing client related product information, is to provide softcopy information to diagnose and fix any problems that might occur with the system. Because the information is electronically maintained, changes due to updates or addition of new capabilities can be used by service representatives immediately.

The IBM KnowledgeCenter can be found online at:

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

Service labels

Service providers use these labels to assist with maintenance actions. Service labels are in various formats and positions, and are intended to transmit readily available information to the service representative during the repair process.

Several of these service labels and their purposes are described in the following list:

- *Location diagrams* are strategically positioned on the system hardware and relate information about the placement of hardware components. Location diagrams can include location codes, drawings of physical locations, concurrent maintenance status, or other data that is pertinent to a repair. Location diagrams are especially useful when multiple components are installed, such as DIMMs, sockets, processor cards, fans, adapter, LEDs, and power supplies.
- *Remove or replace procedure labels* contain procedures that are often found on a cover of the system or in other locations that are accessible to the service representative. These labels provide systematic procedures, including diagrams, detailing how to remove and replace certain serviceable hardware components.
- *Numbered arrows* are used to indicate the order of operation and serviceability direction of components. Various serviceable parts, such as latches, levers, and touch points, must be pulled or pushed in a certain direction and order so that the mechanical mechanisms can engage or disengage. Arrows generally improve the ease of serviceability.

The operator panel

The operator panel on a POWER processor-based system is an LCD display (two rows by 16 elements) that is used to present boot progress codes, indicating advancement through the system power-on and initialization processes. The operator panel is also used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, enabling an IBM SSR or client to change various boot-time options and for other limited service functions.

Concurrent maintenance

The IBM POWER8 processor-based systems are designed with the understanding that certain components have higher intrinsic failure rates than others. These components can include fans, power supplies, and physical storage devices. Other devices, such as I/O adapters, can begin to wear from repeated plugging and unplugging. For these reasons, these devices are designed to be concurrently maintainable when properly configured. Concurrent maintenance is facilitated by the redundant design for the power supplies, fans, and physical storage.

In addition to the previously mentioned components, the operator panel can be replaced concurrently by using service functions of the ASMI menu.

Repair and verify services

Repair and verify (R&V) services are automated service procedures that are used to guide a service provider, step-by-step, through the process of repairing a system and verifying that the problem was repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being serviced. The following scenarios are covered by R&V services:

- Replacing a defective FRU or a CRU
- 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 can be used by both end-user engineers and IBM SSR providers who are familiar with the task and those who are not. Education-on-demand content is placed in the procedure at the appropriate locations. Throughout the R&V procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event to ensure that the guided maintenance procedures are operating correctly.

Clients can subscribe through the subscription services on the IBM Support Portal to obtain notifications about the latest updates that are available for service-related documentation.

4.7 Manageability

Several functions and tools help manageability so you can efficiently and effectively manage your system.

4.7.1 Service user interfaces

The service interface allows support personnel or the client to communicate with the service support applications in a server by using a console, interface, or terminal. Delivering a clear, concise view of available service applications, the service interface allows the support team to manage system resources and service information in an efficient and effective way. Applications that are available through the service interface are carefully configured and placed to give service providers access to important service functions.

Various service interfaces are used, depending on the state of the system and its operating environment. Here are the primary service interfaces:

- Light Path (See “Light Path” on page 181 and “Service labels” on page 181.)
- Service processor and ASMI
- Operator panel
- Operating system service menu
- SFP on the HMC
- SFP Lite on IVM

Service processor

The service processor is a controller that is running its own operating system. It is a component of the service interface card.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface that is connected to the POWER processor. The service processor is always working, regardless of the main system unit’s state. The system unit can be in the following states:

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

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring that the connection to the management console for manageability purposes and accepting ASMI Secure Sockets Layer (SSL) network connections. The service processor can view and manage the machine-wide settings by using the ASMI, and enables complete system and partition management from the HMC.

> **Analyzing a system that does not boot:** The FSP can analyze a system that does not boot. Reference codes and detailed data is available in the ASMI and are transferred to the HMC.

The service processor uses two Ethernet ports that run at 1 Gbps speed. Consider the following information:

- Both Ethernet ports are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI. The ASMI options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- Both Ethernet ports support only auto-negotiation. Customer-selectable media speed and duplex settings are not available.
- Both Ethernet ports have a default IP address, as follows:
  - Service processor eth0 (HMC1 port) is configured as 169.254.2.147.
  - Service processor eth1 (HMC2 port) is configured as 169.254.3.147.

The following functions are available through the service processor:

- Call Home
- ASMI
- Error information (error code, part number, and location codes) menu
- View of guarded components
- Limited repair procedures
- Generate dump
- LED Management menu
- Remote view of ASMI menus
- Firmware update through a USB key

Advanced System Management Interface

ASMI is the interface to the service processor that enables you to manage the operation of the server, such as auto-power restart, and to view information about the server, such as the error log and VPD. Various repair procedures require connection to the ASMI.

The ASMI is accessible through the management console. It is also accessible by using a web browser on a system that is connected directly to the service processor (in this case, either a standard Ethernet cable or a crossed cable) or through an Ethernet network. ASMI can also be accessed from an ASCII terminal, but this is available only while the system is in the platform powered-off mode.

Use the ASMI to change the service processor IP addresses or to apply certain security policies and prevent access from unwanted IP addresses or ranges.

You might be able to use the service processor's default settings. In that case, accessing the ASMI is not necessary. To access ASMI, use one of the following methods:

- Use a management console.

  If configured to do so, the management console connects directly to the ASMI for a selected system from this task.

  To connect to the ASMI from a management console, complete the following steps:

  a. Open **Systems Management** from the navigation pane.
  b. From the work window, select one of the managed systems.
  c. From the System Management tasks list, click **Operations** → **Launch Advanced System Management (ASM)**.

- Use a web browser.

  At the time of writing, supported web browsers are Microsoft Internet Explorer (Version 10.0.9200.16439), Mozilla Firefox ESR (Version 24), and Chrome (Version 30). Later versions of these browsers might work, but are not officially supported. The JavaScript language and cookies must be enabled and TLS 1.2 might need to be enabled.

  The web interface is available during all phases of system operation, including the initial program load (IPL) and run time. However, several of the menu options in the web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides an SSL web connection to the service processor. To establish an SSL connection, open your browser by using the following address:

  ```
  https://<ip_address_of_service_processor>
  ```

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

- Use an ASCII terminal.

  The ASMI on an ASCII terminal supports a subset of the functions that are provided by the web interface and is available only when the system is in the platform powered-off mode. The ASMI on an ASCII console is not available during several phases of system operation, such as the IPL and run time.

- Command-line start of the ASMI

  Either on the HMC itself or when properly configured on a remote system, it is possible to start ASMI web interface from the HMC command line. Open a terminal window on the HMC or access the HMC with a terminal emulation and run the following command:

  ```
  asmmenu --ip <ip address>
  ```

  On the HMC itself, a browser window opens automatically with the ASMI window and, when configured properly, a browser window opens on a remote system when issued from there.

The operator panel

The service processor provides an interface to the operator panel, which is used to display system status and diagnostic information. The operator panel can be accessed in two ways:

- By using the normal operational front view
- By pulling it out to access the switches and viewing the LCD display

Here are several of the operator panel features:

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

The following functions are available through the operator panel:

- Error information
- Generate dump
- View machine type, model, and serial number
- Limited set of repair functions

Operating system service menu

The system diagnostic tests consist of IBM i service tools, stand-alone diagnostic tests that are loaded from the DVD drive, and online diagnostic tests (available in AIX).

Online diagnostic tests, when installed, are a part of the AIX or IBM i operating system on the disk or server. They can be booted in single-user mode (service mode), run in maintenance mode, or run concurrently (concurrent mode) with other applications. They have access to the AIX error log and the AIX configuration data. IBM i has a service tools problem log, IBM i history log (QHST), and IBM i problem log.

The modes are as follows:

- Service mode

  This mode requires a service mode boot of the system and enables the checking of system devices and features. Service mode provides the most complete self-check of the system resources. All system resources, except the SCSI adapter and the disk drives that are used for paging, can be tested.

- Concurrent mode

  This mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, certain devices might require additional actions by the user or a 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 that they are started. Maintenance mode requires that all activity on the operating system is stopped. Run `shutdown -m` to stop all activity on the operating system and put the operating system into maintenance mode.

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

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

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

> **Alternative method:** When you order a Power System, a DVD-ROM or DVD-RAM might be an option. An alternative method for maintaining and servicing the system must be available if you do not order the DVD-ROM or DVD-RAM.

IBM i and its associated machine code provide dedicated service tools (DSTs) as part of the IBM i licensed machine code (Licensed Internal Code) and System Service Tools (SSTs) as part of IBM i. DSTs can be run in dedicated mode (no operating system is loaded). DSTs and diagnostic tests are a superset of those available under SSTs.

The IBM i End Subsystem (`ENDSBS *ALL`) command can shut down all IBM and customer applications subsystems except for the controlling subsystem QTCL. The Power Down System (`PWRDWNSYS`) command can be set to power down the IBM i partition and restart the partition in DST mode.

You can start SST during normal operations, which keeps all applications running, by using the IBM i Start Service Tools (`STRSST`) command (when signed onto IBM i with the appropriately secured user ID).

With DSTs and SSTs, you can look at various logs, run various diagnostic tests, or take several kinds of system memory dumps or other options.

Depending on the operating system, the following service-level functions are what you typically see when you use the operating system service menus:

- Product activity log
- Trace Licensed Internal Code
- Work with communications trace
- Display/Alter/Dump
- Licensed Internal Code log
- Main storage memory dump manager
- Hardware service manager
- Call Home/Customer Notification
- Error information menu
- LED management menu

- Concurrent/Non-concurrent maintenance (within scope of the OS)
- Managing firmware levels
  - Server
  - Adapter
- Remote support (access varies by OS)

Service Focal Point on the Hardware Management Console

Service strategies become more complicated in a partitioned environment. The Manage Serviceable Events task in the management console can help streamline this process.

Each logical partition reports errors that it detects and forwards the event to the SFP application that is running on the management console, without determining whether other logical partitions also detect and report the errors. For example, if one logical partition reports an error for a shared resource, such as a managed system power supply, other active logical partitions might report the same error.

By using the Manage Serviceable Events task in the management console, you can avoid long lists of repetitive Call Home information by recognizing that these are repeated errors and consolidating them into one error.

In addition, you can use the Manage Serviceable Events task to initiate service functions on systems and logical partitions, including the exchanging of parts, configuring connectivity, and managing memory dumps.

4.7.2 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 its associated I/O adapters.

Firmware entitlement

With the new HMC Version V8R8.1.0.0 and Power Systems servers, the firmware installations are restricted to entitled servers. The customer must be registered with IBM and entitled with a service contract. During the initial machine warranty period, the access key is already installed in the machine by manufacturing. The key is valid for the regular warranty period plus some additional time. The Power Systems Firmware is relocated from the public repository to the access control repository. The I/O firmware remains on the public repository, but the server must be entitled for installation. When the **`lslic`** command is run to display the firmware levels, a new value, **`update_access_key_exp_date`**, is added. The HMC GUI and the ASMI menu show the Update access key expiration date.

When the system is no longer entitled, the firmware updates fail. Some new System Reference Code (SRC) packages are available:

- E302FA06: Acquisition entitlement check failed
- E302FA08: Installation entitlement check failed

Any firmware release that was made available during the entitled time frame can still be installed. For example, if the entitlement period ends on 31 December 2014, and a new firmware release is release before the end of that entitlement period, then it can still be installed. If that firmware is downloaded after 31 December 2014, but it was made available before the end of the entitlement period, it can still be installed. Any newer release requires a new update access key.

Note: The update access key expiration date requires a valid entitlement of the system to perform firmware updates.

You can find an update access key at the IBM CoD Home website:

`http://www-912.ibm.com/pod/pod`

To access the IBM entitled Software Support page for further details, go to the following website:

`http://www.ibm.com/servers/eserver/ess`

Firmware updates

System firmware is delivered as a release level or a service pack. Release levels support the general availability (GA) of new functions or features, and new machine types or models. Upgrading to a higher release level is disruptive to customer operations. IBM intends to introduce no more than two new release levels per year. These release levels will be supported by service packs. Service packs are intended to contain only firmware fixes and not introduce new functions. A *service pack* is an update to an existing release level.

The management console is used for system firmware updates. By using the management console, you can take advantage of the CFM option when concurrent service packs are available. CFM is the IBM Power Systems Firmware updates that can be partially or wholly concurrent or nondisruptive. With the introduction of CFM, IBM is increasing its clients' opportunity to stay on a given release level for longer periods. Clients that want maximum stability can defer until there is a compelling reason to upgrade, such as the following reasons:

- A release level is approaching its end-of-service date (that is, it has been available for about a year, and soon service will not be supported).
- They want to move a system to a more standardized release level when there are multiple systems in an environment with similar hardware.
- A new release has a new function that is needed in the environment.
- A scheduled maintenance action causes a platform reboot, which provides an opportunity to also upgrade to a new firmware release.

The updating and upgrading of system firmware depends on several factors, including the current firmware that is installed, and what operating systems are running on the system. These scenarios and the associated installation instructions are comprehensively outlined in the firmware section of Fix Central, found at the following website:

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

You might also want to review the preferred practice white papers that are found at the following website:

`http://www14.software.ibm.com/webapp/set2/sas/f/best/home.html`

Firmware update steps

The system firmware consists of service processor microcode, Open Firmware microcode, and Systems Power Control Network (SPCN) microcode.

The firmware and microcode can be downloaded and installed either from the HMC, or from a running partition.

Power Systems has a permanent firmware boot side (A side) and a temporary firmware boot side (B side). New levels of firmware must be installed first on the temporary side to test the update's compatibility with existing applications. When the new level of firmware is approved, it can be copied to the permanent side.

For access to the initial websites that address this capability, see the POWER8 section on the IBM Support Portal:

https://www.ibm.com/support/entry/portal/product/power/

For POWER8 based Power Systems, select the **POWER8 systems** link.

Within this section, search for **Firmware and HMC updates** to find the resources for keeping your system's firmware current.

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

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

- Installed level

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

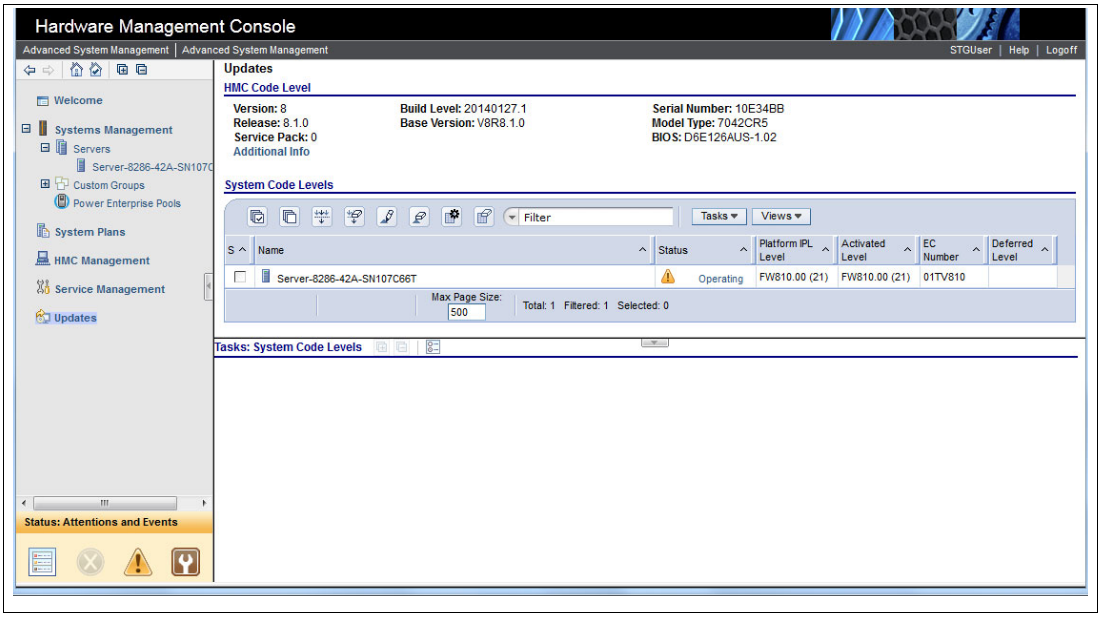
- Activated level

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

- Accepted level

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

Figure 4-4 shows the different levels in the HMC.



Updates

HMC Code Level

Version: 8
Release: 8.1.0
Service Pack: 0
Additional Info

Build Level: 20140127.1
Base Version: V8R8.1.0

Serial Number: 10E34BB
Model Type: 7042CR5
BIOS: D6E126AUS-1.02

System Code Levels

S	Name	Status	Platform IPL Level	Activated Level	EC Number	Deferred Level
□	Server-8286-42A-SN107C66T	Operating	FW810.00 (21)	FW810.00 (21)	01TV810	

Max Page Size: 500 Total: 1 Filtered: 1 Selected: 0

Figure 4-4 HMC System Firmware window

IBM provides the CFM function on the Power E870 and Power E880 models. This function supports applying nondisruptive system firmware service packs to the system concurrently (without requiring a reboot operation to activate changes).

The concurrent levels of system firmware can, on occasion, contain fixes that are known as *deferred*. These deferred fixes can be installed concurrently but are not activated until the next IPL. Deferred fixes, if any, are identified in the Firmware Update Descriptions table of the firmware document. For deferred fixes within a service pack, only the fixes in the service pack that cannot be concurrently activated are deferred.

Table 4-1 shows the file-naming convention for system firmware.

Table 4-1 Firmware naming convention

PPNNSSS_FFF_DDD			
PP	Package identifier	01	-
NN	Platform and class	SV	Low end
SSS	Release indicator		
FFF	Current fix pack		
DDD	Last disruptive fix pack		

The following example uses the convention:

01SV810_030_030 = POWER8 Entry Systems Firmware for 8286-41A and 8286-42A

An installation is disruptive if the following statements are true:

- The release levels (SSS) of the currently installed and the new firmware differ.
- The service pack level (FFF) and the last disruptive service pack level (DDD) are equal in the new firmware.

Otherwise, an installation is concurrent if the service pack level (FFF) of the new firmware is higher than the service pack level that is installed on the system and the conditions for disruptive installation are not met.

4.7.3 Concurrent firmware maintenance improvements

Since POWER6, firmware service packs are concurrently applied and take effect immediately. Occasionally, a service pack is shipped where most of the features can be concurrently applied, but because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during operation, a patch in this area required a system reboot for activation.

With the Power-On Reset Engine (PORE), the firmware can now dynamically power off processor components, change the registers, and reinitialize while the system is running, without discernible impact to any applications running on a processor. This potentially allows concurrent firmware changes in POWER8, which in earlier designs required a reboot to take effect.

Activating new firmware functions requires installation of a firmware release level. This process is disruptive to server operations and requires a scheduled outage and full server reboot.

4.7.4 Electronic Services and Electronic Service Agent

IBM transformed its delivery of hardware and software support services to help you achieve higher system availability. Electronic Services is a web-enabled solution that offers an exclusive, no additional charge enhancement to the service and support that is available for IBM servers. These services provide the opportunity for greater system availability with faster problem resolution and preemptive monitoring. The Electronic Services solution consists of two separate, but complementary, elements:

- Electronic Services news page
- Electronic Service Agent

Electronic Services news page

The Electronic Services news page is a single Internet entry point that replaces the multiple entry points that traditionally are used to access IBM Internet services and support. With the news page, you can gain easier access to IBM resources for assistance in resolving technical problems.

Electronic Service Agent

The ESA is software that is on your server. It monitors events and transmits system inventory information to IBM on a periodic, client-defined timetable. The ESA automatically reports hardware problems to IBM.

Early knowledge about potential problems enables IBM to deliver proactive service that can result in higher system availability and performance. In addition, information that is collected through the Service Agent is made available to IBM SSRs when they help answer your questions or diagnose problems. Installation and use of ESA for problem reporting enables IBM to provide better support and service for your IBM server.

To learn how Electronic Services can work for you, see the following website (an IBM ID is required):

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

Here are some of the benefits of Electronic Services:

- Increased uptime

  The ESA tool enhances the warranty or maintenance agreement by providing faster hardware error reporting and uploading system information to IBM Support. This can translate to less time that is wasted monitoring the symptoms, diagnosing the error, and manually calling IBM Support to open a problem record.

  Its 24x7 monitoring and reporting mean no more dependence on human intervention or off-hours customer personnel when errors are encountered in the middle of the night.

- Security

  The ESA tool is designed to be secure in monitoring, reporting, and storing the data at IBM. The ESA tool securely transmits either through the Internet (HTTPS or VPN) or modem, and can be configured to communicate securely through gateways to provide customers a single point of exit from their site.

  Communication is one way. Activating ESA does not enable IBM to call into a customer's system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. It is viewable only by the customer and IBM. The customer's business applications or business data is never transmitted to IBM.

- More accurate reporting

  Because system information and error logs are automatically uploaded to the IBM Support center with the service request, customers are not required to find and send system information, decreasing the risk of misreported or misdiagnosed errors.

  When inside IBM, problem error data is run through a data knowledge management system and knowledge articles are appended to the problem record.

- Customized support

  By using the IBM ID that you enter during activation, you can view system and support information by selecting **My Systems** at the Electronic Support website:

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

  My Systems provides valuable reports of installed hardware and software, using information that is collected from the systems by ESA. Reports are available for any system that is associated with the customers IBM ID. Premium Search combines the function of search and the value of ESA information, providing advanced search of the technical support knowledge base. Using Premium Search and the ESA information that was collected from your system, your clients can see search results that apply specifically to their systems.

  For more information about how to use the power of IBM Electronic Services, contact your IBM SSR, or see the following website:

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

Service Event Manager

The Service Event Manager allows the user to decide which of the Serviceable Events are called home with the ESA. It is possible to lock certain events. Some customers might not allow data to be transferred outside their company. After the SEM is enabled, the analysis of the possible problems might take longer.

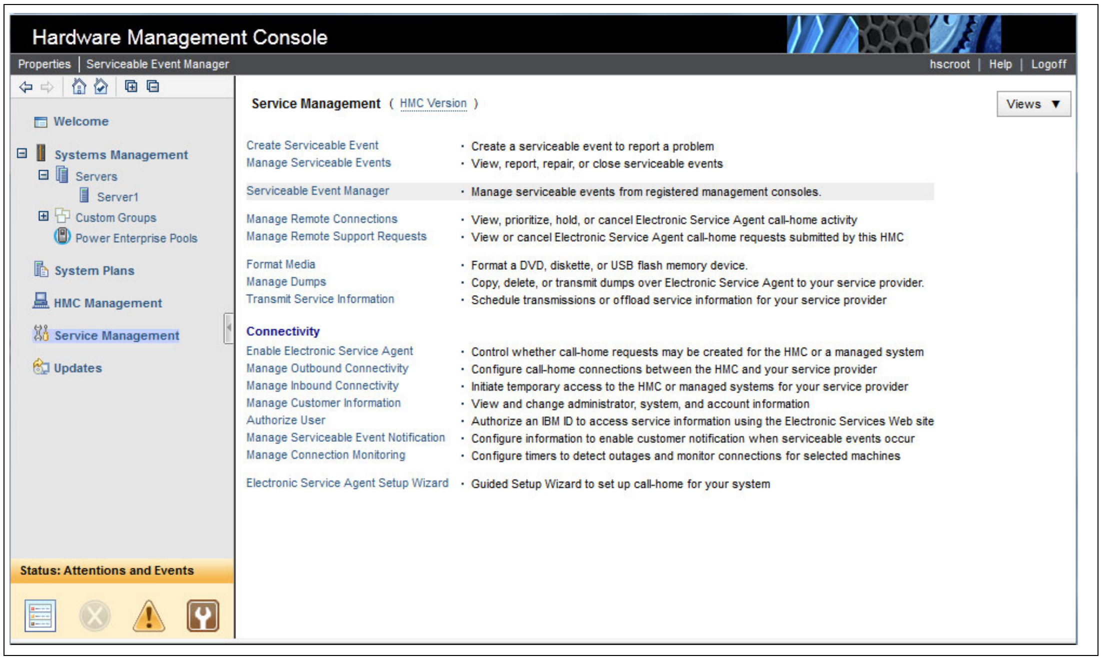
- The SEM can be enabled by running the following command:

```
chhmc -c sem -s enable
```

- You can disable SEM mode and specify what state in which to leave the Call Home feature by running the following commands:

```
chhmc -c sem -s disable --callhome disable
chhmc -c sem -s disable --callhome enable
```

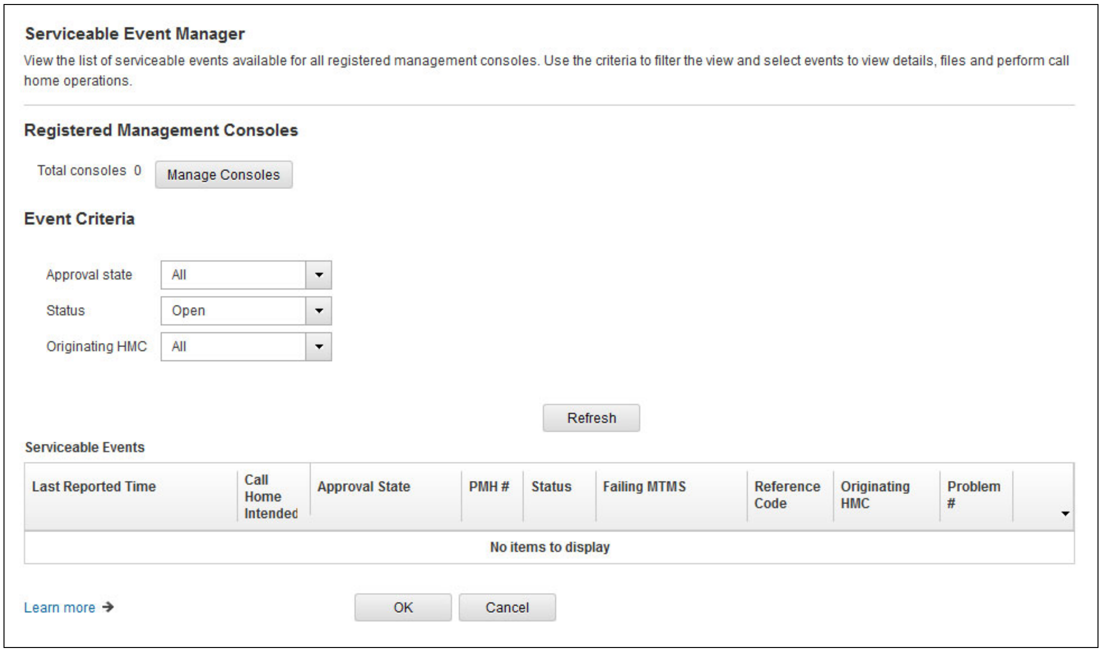
You can do the basic configuration of the SEM from the HMC GUI. After you select the Service Event Manager, as shown in Figure 4-5, you must add the HMC console.



Create Serviceable Event	· Create a serviceable event to report a problem
Manage Serviceable Events	· View, report, repair, or close serviceable events
Serviceable Event Manager	· Manage serviceable events from registered management consoles.
Manage Remote Connections	· View, prioritize, hold, or cancel Electronic Service Agent call-home activity
Manage Remote Support Requests	· View or cancel Electronic Service Agent call-home requests submitted by this HMC
Format Media	· Format a DVD, diskette, or USB flash memory device.
Manage Dumps	· Copy, delete, or transmit dumps over Electronic Service Agent to your service provider.
Transmit Service Information	· Schedule transmissions or offload service information for your service provider
Connectivity	
Enable Electronic Service Agent	· Control whether call-home requests may be created for the HMC or a managed system
Manage Outbound Connectivity	· Configure call-home connections between the HMC and your service provider
Manage Inbound Connectivity	· Initiate temporary access to the HMC or managed systems for your service provider
Manage Customer Information	· View and change administrator, system, and account information
Authorize User	· Authorize an IBM ID to access service information using the Electronic Services Web site
Manage Serviceable Event Notification	· Configure information to enable customer notification when serviceable events occur
Manage Connection Monitoring	· Configure timers to detect outages and monitor connections for selected machines
Electronic Service Agent Setup Wizard	· Guided Setup Wizard to set up call-home for your system

Figure 4-5 HMC selection for Service Event Manager

In the next window, you can configure the HMC that is used to manage the Serviceable Events and proceed with further configuration steps, as shown in Figure 4-6.



Serviceable Event Manager

View the list of serviceable events available for all registered management consoles. Use the criteria to filter the view and select events to view details, files and perform call home operations.

Registered Management Consoles

Total consoles 0

Event Criteria

Approval state: All

Status: Open

Originating HMC: All

Serviceable Events

Last Reported Time	Call Home Intended	Approval State	PMH #	Status	Failing MTMS	Reference Code	Originating HMC	Problem #
No items to display								

Figure 4-6 Initial SEM window

Here are detailed descriptions of the different configurable options:

- Registered Management Consoles

  “Total consoles” lists the number of consoles that are registered. Select **Manage Consoles** to manage the list of RMCs.

- Event Criteria

  Select the filters for filtering the list of serviceable events that are shown. After the selections are made, click **Refresh** to refresh the list based on the filter values.

- Approval state

  Select the value for approval state to filter the list.

- Status

  Select the value for the status to filter the list.

- Originating HMC

  Select a single registered console or **All consoles** to filter the list.

- Serviceable Events

  The Serviceable Events table shows the list of events based on the filters that are selected. To refresh the list, click **Refresh**.

The following menu options are available when you select an event in the table:

- View Details...

  Shows the details of this event.

- View Files...

  Shows the files that are associated with this event.

- Approve Call Home

  Approves the Call Home of this event. This option is available only if the event is not approved already.

The Help / Learn more function can be used to get more information about the other available windows for the Serviceable Event Manager.

4.8 Selected POWER8 RAS capabilities by operating system

Table 4-2 provides a list of the Power Systems RAS capabilities by operating system. The HMC is an optional feature on scale-out Power Systems servers.

Table 4-2 Selected RAS features by operating system

RAS feature	AIX V7.1 TL3 SP3 V6.1 TL9 SP3	IBM i V7R1M0 TR8 V7R2M0	Linux RHEL6.5 RHEL7.1 SLES11SP3 SLES12 Ubuntu 15.04
Processor			
FFDC for fault detection/error isolation	X	X	X
Dynamic Processor Deallocation	X	X	X[a]
Dynamic Processor Sparing using capacity from spare pool	X	X	X[a]
Core Error Recovery			
- Alternative processor recovery	X	X	X[a]
- Partition Core Contained Checkstop	X	X	X[a]
I/O subsystem			
PCI Express bus enhanced error detection	X	X	X
PCI Express bus enhanced error recovery	X	X	X[b]
PCI Express card hot-swap	X	X	X[a]
Memory availability			
Memory Page Deallocation	X	X	X
Special Uncorrectable Error Handling	X	X	X
Fault detection and isolation			
Storage Protection Keys	X	Not used by OS	Not used by OS
Error log analysis	X	X	X[b]
Serviceability			
Boot-time progress indicators	X	X	X
Firmware error codes	X	X	X

RAS feature	AIX V7.1 TL3 SP3 V6.1 TL9 SP3	IBM i V7R1M0 TR8 V7R2M0	Linux RHEL6.5 RHEL7.1 SLES11SP3 SLES12 Ubuntu 15.04
Operating system error codes	X	X	X[b]
Inventory collection	X	X	X
Environmental and power warnings	X	X	X
Hot-swap DASD / media	X	X	X
Dual disk controllers / Split backplane	X	X	X
EED collection	X	X	X
SP “Call Home” on non-HMC configurations	X	X	X[a]
IO adapter/device stand-alone diagnostic tests with PowerVM	X	X	X
SP mutual surveillance with POWER Hypervisor	X	X	X
Dynamic firmware update with HMC	X	X	X
Service Agent Call Home Application	X	X	X[a]
Service Indicator LED support	X	X	X
System dump for memory, POWER Hypervisor, and SP	X	X	X
Information center / IBM Systems Support Site service publications	X	X	X
System Support Site education	X	X	X
Operating system error reporting to HMC SFP application	X	X	X
RMC secure error transmission subsystem	X	X	X
Healthcheck scheduled operations with HMC	X	X	X
Operator panel (real or virtual)	X	X	X
Concurrent Operator Panel Maintenance	X	X	X
Redundant HMCs	X	X	X
Automated server recovery/restart	X	X	X
High availability clustering support	X	X	X
Repair and Verify Guided Maintenance with HMC	X	X	X
PowerVM Live Partition / Live Application Mobility With PowerVM Enterprise Edition	X	X[c]	X
EPOW			
EPOW errors handling	X	X	X[a]

a. Supported in POWER Hypervisor, but not supported in a PowerKVM environment
b. Supported in POWER Hypervisor
c. For POWER8 systems, IBM i requires IBM i 7.1 TR9 and IBM i 7.2 TR1.

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

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491
- *IBM Power Systems S812L and S822L Technical Overview and Introduction*, REDP-5098
- *IBM Power System S822 Technical Overview and Introduction*, REDP-5102
- *IBM Power Systems S814 and S824 Technical Overview and Introduction*, REDP-5097
- *IBM Power System E870 and E880 Technical Overview and Introduction,* REDP-5137
- *IBM Power Systems SR-IOV: Technical Overview and Introduction*, REDP-5065
- *IBM PowerVM Best Practices*, SG24-8062
- *IBM PowerVM Enhancements What is New in 2013*, SG24-8198
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *Performance Optimization and Tuning Techniques for IBM Processors, including IBM POWER8*, SG24-8171

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- *Active Memory Expansion: Overview and Usage Guide* documentation:

  http://public.dhe.ibm.com/common/ssi/ecm/en/pow03037usen/POW03037USEN.PDF

- *IBM EnergyScale for POWER8 Processor-Based Systems* white paper:

  http://public.dhe.ibm.com/common/ssi/ecm/po/en/pow03125usen/POW03125USEN.PDF

- IBM Power Facts and Features - IBM Power Systems, IBM PureFlex System, and Power Blades:

  http://www.ibm.com/systems/power/hardware/reports/factsfeatures.html

- IBM Power System E870 server specifications:

  http://www.ibm.com/systems/power/hardware/e870/specs.html

- IBM Power System E880 server specifications:

  http://www.ibm.com/systems/power/hardware/e880/specs.html

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM Power Systems E870 and E880 Technical Overview and Introduction

New modular architecture for increased reliability

Enterprise POWER8 processor-based servers

Exceptional memory and I/O bandwidth

This IBM Redpaper publication is a comprehensive guide covering the IBM Power System E870 (9119-MME) and IBM Power System E880 (9119-MHE) servers that support IBM AIX, IBM i, and Linux operating systems. The objective of this paper is to introduce the major innovative Power E870 and Power E880 offerings and their relevant functions:

- The new IBM POWER8 processor, available at frequencies of 4.024 GHz, 4.190 GHz, and 4.350 GHz.
- Significantly strengthened cores and larger caches
- Two integrated memory controllers with improved latency and bandwidth
- Integrated I/O subsystem and hot-pluggable PCIe Gen3 I/O slots
- Improved reliability, serviceability, and availability (RAS) functions
- IBM EnergyScale technology that provides features such as power trending, power-saving, capping of power, and thermal measurement

This publication is for professionals who want to acquire a better understanding of IBM Power Systems products.

This paper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power E870 and Power E880 systems.

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

REDP-5137-00