

IBM Power Systems E870C and E880C Technical Overview and Introduction

Scott Vetter

Alexandre Bicas Caldeira

Volker Haug



 **Cloud**

Power Systems



International Technical Support Organization

**IBM Power Systems E870C and E880C Technical
Overview and Introduction**

October 2016

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

First Edition (October 2016)

This edition applies to the IBM Power E870C (9080-MME) and Power E880C (9080-MHE) Power Systems servers.

| This document was created or updated on November 14, 2018.

© Copyright International Business Machines Corporation 2016. All rights reserved.

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

Contents

Notices	vii
Trademarks	viii
 Preface	 ix
Authors	x
Now you can become a published author, too!	xi
Comments welcome	xi
Stay connected to IBM Redbooks	xi
 Chapter 1. General description	 1
1.1 Systems overview	3
1.1.1 Power E870C server	3
1.1.2 Power E880C server	4
1.1.3 System control unit	5
1.1.4 System nodes	6
1.1.5 I/O drawers	9
1.2 Operating environment	9
1.3 Physical package	10
1.4 System features	11
1.4.1 Power E870C system features	11
1.4.2 Power E880C system features	12
1.4.3 Minimum features	13
1.4.4 Power supply features	15
1.4.5 Processor card features	16
1.5 Disk and media features	16
1.6 I/O drawers	17
1.6.1 PCIe Gen3 I/O expansion drawer	17
1.6.2 I/O drawers and usable PCI slot	19
1.7 EXP24S SFF Gen2-bay Drawer	20
1.8 Model comparison	21
1.9 Build to order	22
1.10 Model upgrades	22
1.10.1 Power E870C	22
1.10.2 Power E880C	23
1.10.3 Upgrade considerations	23
1.11 Management consoles	24
1.12 System racks	25
1.12.1 IBM 7014 model T00 rack	26
1.12.2 IBM 7014 model T42 rack	26
1.12.3 Feature code #0551 rack	28
1.12.4 Feature code #0553 rack	28
1.12.5 AC power distribution unit and rack content	28
1.12.6 Rack-mounting rules	31
 Chapter 2. Architecture and technical overview	 35
2.1 Logical diagrams	36
2.2 IBM POWER8 processor	40
2.3 Memory subsystem	41
2.3.1 Custom DIMM	41

2.3.2	Memory placement rules	42
2.3.3	Memory activation	47
2.3.4	Memory throughput	48
2.3.5	Active Memory Mirroring	52
2.3.6	Memory Error Correction and Recovery	53
2.3.7	Special Uncorrectable Error handling	53
2.4	Capacity on Demand	53
2.4.1	Capacity Upgrade on Demand	53
2.4.2	Power enterprise pools and Mobile Capacity on Demand	55
2.4.3	Elastic Capacity on Demand	56
2.4.4	Utility Capacity on Demand	58
2.4.5	Trial Capacity on Demand	58
2.4.6	Software licensing and CoD	59
2.5	System bus	59
2.5.1	PCI Express Gen3	59
2.5.2	Service Processor Bus	61
2.6	Internal I/O subsystem	63
2.6.1	Blind-swap cassettes	64
2.6.2	System ports	64
2.7	PCI adapters	65
2.7.1	PCI Express	65
2.7.2	LAN adapters	66
2.7.3	Graphics accelerator adapters	68
2.7.4	SAS adapters	68
2.7.5	Fibre Channel adapter	69
2.7.6	Fibre Channel over Ethernet	70
2.7.7	USB adapters	71
2.7.8	InfiniBand host channel adapter	71
2.7.9	Cryptographic Coprocessor	72
2.7.10	CAPI adapters	72
2.8	Internal storage	72
2.8.1	Media features	73
2.9	External I/O subsystems	74
2.9.1	PCIe Gen3 I/O expansion drawer	74
2.9.2	PCIe Gen3 I/O expansion drawer optical cabling	75
2.9.3	PCIe Gen3 I/O expansion drawer SPCN cabling	80
2.10	External disk subsystems	81
2.10.1	EXP24S SFF Gen2-bay Drawer	81
2.10.2	EXP24S common usage scenarios	86
2.10.3	IBM System Storage	92
2.11	Hardware Management Console	93
2.11.1	Hardware appliance HMC	93
2.11.2	Virtual appliance HMC	94
2.11.3	HMC code level	94
2.11.4	HMC connectivity to the POWER8 processor-based systems	95
2.11.5	High availability HMC configuration	96
2.12	Operating system support	96
2.12.1	Virtual I/O Server	97
2.12.2	IBM AIX operating system	97
2.12.3	IBM i operating system	97
2.12.4	Linux operating systems	98
2.12.5	Supported Java versions	99
2.13	Energy management	99

2.13.1 IBM EnergyScale technology	99
2.13.2 On Chip Controller	102
2.13.3 Energy consumption estimation	102
Chapter 3. Private and Hybrid Cloud features	105
3.1 Private cloud software	106
3.1.1 IBM Cloud PowerVC Manager	106
3.1.2 Cloud-based HMC Apps as a Service.	107
3.1.3 Open source cloud automation and configuration tooling for AIX.	108
3.2 Hybrid cloud support	110
3.2.1 Hybrid infrastructure management tools	110
3.2.2 Securely connecting system of record workloads to cloud native applications.	111
3.2.3 IBM Cloud Starter Pack	111
3.2.4 Flexible capacity on-demand	111
3.3 Geographically Dispersed Resiliency for Power	112
3.4 IBM Power to Cloud Rewards Program	113
Chapter 4. Reliability, availability, serviceability, and manageability	115
4.1 RAS enhancements of POWER8 processor-based servers	116
4.1.1 POWER8 overview	117
4.2 Reliability	118
4.2.1 Designed for reliability.	118
4.2.2 Component placement	119
4.3 Processor/Memory availability	120
4.3.1 Correctable errors	120
4.3.2 Uncorrectable errors	121
4.3.3 Processor core/cache correctable error handling	121
4.3.4 Processor instruction retry and other try again techniques	121
4.3.5 Alternative processor recovery and partition availability priority	122
4.3.6 Core contained checkstops and other PowerVM error recovery	122
4.3.7 Cache uncorrectable error handling	122
4.3.8 Other processor chip functions	123
4.3.9 Other fault error handling	123
4.3.10 Memory protection	124
4.3.11 I/O subsystem availability and Enhanced Error Handling	125
4.4 Enterprise systems availability	126
4.5 Availability effects of a solution architecture	127
4.5.1 Clustering	127
4.5.2 Virtual I/O redundancy configurations	127
4.5.3 PowerVM Live Partition Mobility	128
4.6 Serviceability	129
4.6.1 Detecting errors	130
4.6.2 Error checkers, fault isolation registers, and First-Failure Data Capture	130
4.6.3 Service processor	131
4.6.4 Diagnosing	132
4.6.5 Reporting	133
4.6.6 Notifying	135
4.6.7 Locating and servicing	136
4.7 Manageability	139
4.7.1 Service user interfaces	139
4.7.2 IBM Power Systems Firmware maintenance	144
4.7.3 Concurrent firmware maintenance improvements	147
4.7.4 Electronic Services and Electronic Service Agent	147

4.8 Selected POWER8 RAS capabilities by operating system	151
Related publications	153
IBM Redbooks	153
Online resources	154
Help from IBM	155

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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, MD-NC119, Armonk, NY 10504-1785, US

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 jurisdictions 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 provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

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

The following terms are trademarks of other companies:

C3, and Phyteland device are trademarks or registered trademarks of Phytel, Inc., an IBM Company.

SoftLayer, and SoftLayer device are trademarks or registered trademarks of SoftLayer, Inc., an IBM Company.

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 that covers the IBM Power® System E870C (9080-MME) and IBM Power System E880C (9080-MHE) servers that support IBM AIX®, IBM i, and Linux operating systems. The objective of this paper is to introduce the major innovative Power E870C and Power E880C offerings and their relevant functions.

The new Power E870C and Power E880C servers with OpenStack-based cloud management and open source automation enables clients to accelerate the transformation of their IT infrastructure for cloud while providing tremendous flexibility during the transition. In addition, the Power E870C and Power E880C models provide clients increased security, high availability, rapid scalability, simplified maintenance, and management, all while enabling business growth and dramatically reducing costs.

The systems management capability of the Power E870C and Power E880C servers speeds up and simplifies cloud deployment by providing fast and automated VM deployments, prebuilt image templates, and self-service capabilities, all with an intuitive interface.

Enterprise servers provide the highest levels of reliability, availability, flexibility, and performance to bring you a world-class enterprise private and hybrid cloud infrastructure. Through enterprise-class security, efficient built-in virtualization that drives industry-leading workload density, and dynamic resource allocation and management, the server consistently delivers the highest levels of service across hundreds of virtual workloads on a single system.

The Power E870C and Power E880C server includes the cloud management software and services to assist with clients' move to the cloud, both private and hybrid. The following capabilities are included:

- ▶ Private cloud management with IBM Cloud PowerVC Manager, Cloud-based HMC Apps as a service, and open source cloud automation and configuration tooling for AIX
- ▶ Hybrid cloud support
- ▶ Hybrid infrastructure management tools
- ▶ Securely connect system of record workloads and data to cloud native applications
- ▶ IBM Cloud Starter Pack
- ▶ Flexible capacity on demand
- ▶ Power to Cloud Services

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 E870C and Power E880C systems.

This paper does not replace the latest marketing materials and configuration tools. It is intended as another 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 the Product Manager for Power Systems Latin America. 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, marketing, and technical sales support. Alexandre has more than 16 years of experience working on IBM Systems Solutions and has worked as an IBM Business Partner on Power Systems hardware, AIX, and IBM PowerVM® virtualization products.

Volker Haug is an Executive IT Specialist & Open Group Distinguished IT Specialist within IBM Systems in Germany supporting Power Systems clients and Business Partners. He holds a degree in Business Management from the University of Applied Studies in Stuttgart. His career includes more than 29 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.

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:

George Ahrens
Joanna Bartz
Thomas Bosworth
Petra Buehrer
David Bruce
Sertak Cakici
Joseph W Cropper
Timothy Damron
Daniel Henderson
Dan Hurliman
Roxette Johnson
Ann Lund
Edgar Michel Garcia
Hans Mozes
Michael J Mueller
Mark Olson
Ravi A Shankar
Deepak C Shetty
Bill Starke
Dawn C Stokes
Jeff Stuecheli
Doug Szerdi
Joel Sebastian Torres De la Torre
IBM

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>



General description

The IBM Power System E870C (9080-MME) and IBM Power System E880C (9080-MHE) servers use POWER8 processor technology that is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads. The servers are optimized for running AIX, IBM i, and Linux workloads.

The Power E870C is a modular-built system and uses one or two system nodes together. The Power E880C system is built of one, two, three, or four system nodes together.

Each system requires a system control unit (SCU). Each system node is five EIA-units tall (5U) and the SCU is two EIA-units (2U) tall. The servers are housed in a 19-inch rack.

The servers with OpenStack-based cloud management and open source automation enable clients to accelerate the transformation of their IT infrastructure for cloud while providing tremendous flexibility during the transition. In addition, both models provide clients increased security, high availability, rapid scalability, simplified maintenance, and management, all while enabling business growth and dramatically reducing costs.

The systems management capability speeds up and simplifies cloud deployment by providing fast and automated VM deployments, prebuilt image templates, and self-service capabilities all with an intuitive interface.

Both servers include the cloud management software and services to assist with clients' move to the cloud, both private and hybrid. The following capabilities are included:

- ▶ Private cloud management with IBM Cloud PowerVC Manager, Cloud-based HMC Apps as a service, and open source cloud automation, and configuration tooling for AIX
- ▶ Hybrid cloud support
- ▶ Hybrid infrastructure management tools
- ▶ Securely connect system of record workloads and data to cloud native applications
- ▶ IBM Cloud Starter Pack
- ▶ Flexible capacity on demand
- ▶ Power to Cloud Services

For more information about the cloud management software and services, see Chapter 3, “Private and Hybrid Cloud features” on page 105.

This chapter includes the following topics:

- ▶ 1.1, “Systems overview”
- ▶ 1.2, “Operating environment” on page 9
- ▶ 1.3, “Physical package” on page 10
- ▶ 1.4, “System features” on page 11
- ▶ 1.5, “Disk and media features” on page 16
- ▶ 1.6, “I/O drawers” on page 17
- ▶ 1.7, “EXP24S SFF Gen2-bay Drawer” on page 20
- ▶ 1.8, “Model comparison” on page 21
- ▶ 1.9, “Build to order” on page 22
- ▶ 1.10, “Model upgrades” on page 22
- ▶ 1.11, “Management consoles” on page 24
- ▶ 1.12, “System racks” on page 25

1.1 Systems overview

The following sections provide more information about the Power E870C and Power E880C systems.

1.1.1 Power E870C server

The Power E870C (9080-MME) server is a powerful POWER8 based system that scales up to eight sockets. Each socket contains a single 8-core POWER8 processor. Thus, a fully configured Power E870C can scale up to 64 cores.

The Power E870C is a modular system that is built from a combination of the following four building blocks:

- ▶ SCU
- ▶ System node
- ▶ PCIe Gen3 I/O expansion drawer
- ▶ EXP24S SFF Gen2-bay drawer

The SCU 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 SCU provides clock signals to the system nodes with semi-flexible connectors. A SCU is required for every E870C system.

Each system node provides four processor sockets for POWER8 processors and 32 CDIMM slots, which supports a maximum of eight memory features. By using the 1024 GB memory features, the system node can support a maximum of 8 TB of RAM. A fully configured Power E870C can support up to 16 TB of RAM.

Each system node provides eight PCIe Gen3 x16 low profile slots. One or two system nodes can be included in an E870C 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 Gen3 slots. The I/O expansion drawer connects to the system node with a pair of PCIe x16 to optical CXP converter adapters that are housed in the system node. Each system node can support up to four I/O expansion drawers. A fully configured Power E870C 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 E870C server by using a SAS adapter in a PCIe slot in a system node or in an I/O expansion drawer.

Figure 1-1 shows a single system node Power E870C in a T42 rack with two PCIe I/O drawers and an EXP24S disk drawer.



Figure 1-1 Power E870C in a T42 rack

1.1.2 Power E880C server

The Power E880C (9080-MHE) server is a powerful POWER8 based system that scales up to 16 sockets. Each socket contains a single 8-core, 10-core, or 12-core POWER8 processor. A fully configured Power E880C can scale up to 128, 160, or 192 cores.

The Power E880C is a modular system that is built from a combination of the following four building blocks:

- ▶ SCU
- ▶ System node
- ▶ PCIe Gen3 I/O expansion drawer
- ▶ EXP24S SFF Gen2-bay drawer

The SCU provides redundant system master clock and redundant system master Flexible Service Processor (FSP) support for the Operator Panel, the system VPD, and the base DVD. The SCU provides clock signals to the system nodes with semi-flexible connectors. A SCU is required for every E880C system.

Each system node provides four processor sockets for POWER8 processors and 32 CDIMM slots, which support a maximum of eight memory features. By using the 1024 GB memory features, each system node can support a maximum of 8 TB of RAM. A fully configured four-node Power E880C 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 12 PCIe Gen 3 slots. The I/O expansion drawer connects to the system node with a pair of PCIe x16 to Optical CXP converter adapters that are housed in the system node. 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 E880C can support a maximum of 16 I/O expansion drawers, which provides 192 PCIe Gen3 slots.

Each EXP24S SFF Gen2-bay Drawer provides 24 x 2.5-inch form-factor (SFF) SAS bays. The EXP24S is connected to the Power E880C server by using a SAS adapter in a PCIe slot in a system node or in an I/O expansion drawer.

Figure 1-2 shows a four system node Power E880C with two PCIe I/O drawers and an EXP24S disk drawer.



Figure 1-2 Power E880C in a T42 rack

1.1.3 System control unit

The SCU in a Power E870C and E880C 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 E870C and E880C systems, even if only one system node is installed. The SCU also provides a DVD option that is connected to a PCIe adapter by using a USB cable.

The SCU is powered from the system nodes by using cables that are 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 cable.

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

The SCU provides redundant service processor function to the system nodes by using FSP cables. Each system node has two FSP connections to the SCU. The SCU provides redundant connections to one or two HMCs by using 1 Gbps RJ45 Ethernet connections.

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

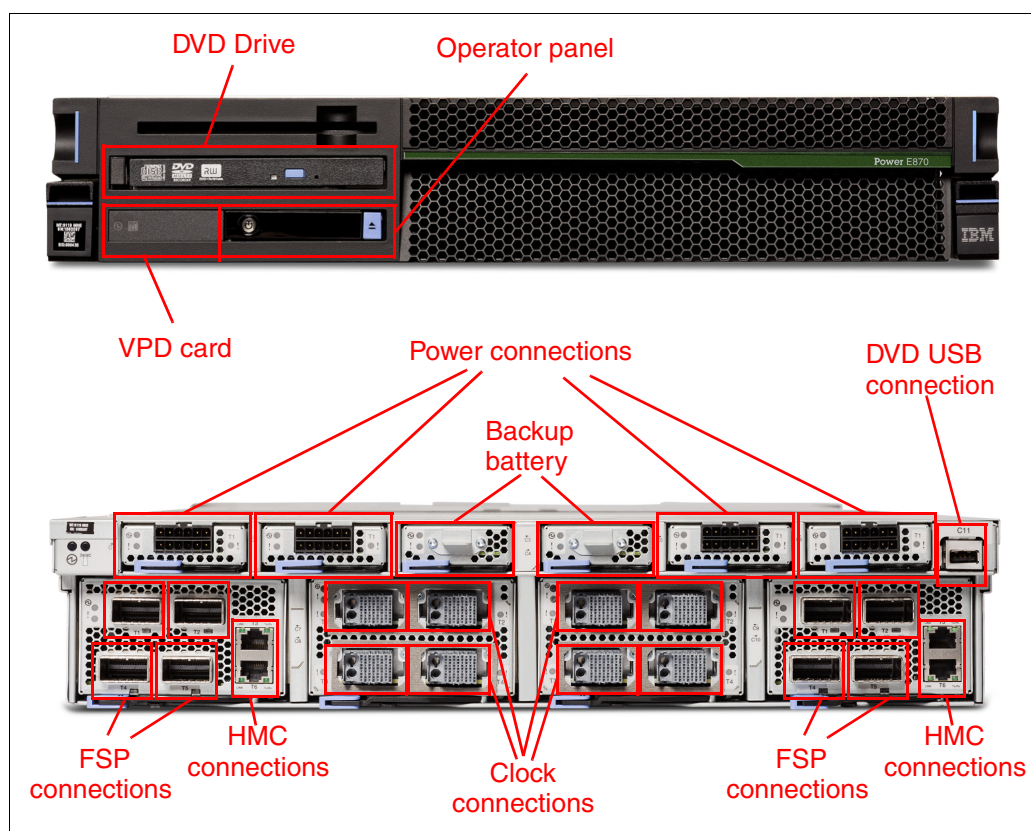


Figure 1-3 Front and rear view of the system control unit

1.1.4 System nodes

The system nodes in the Power E870C and E880C 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 SCU with FSP, clock, and power connectors. The system node connects to other system nodes by using SMP connectors.

Each system node in a Power E870C or E880C server provides four POWER8 processors, 32 CDIMM slots, and eight PCIe Gen3 x16 slots.

The front view of the system node is shown in Figure 1-4.

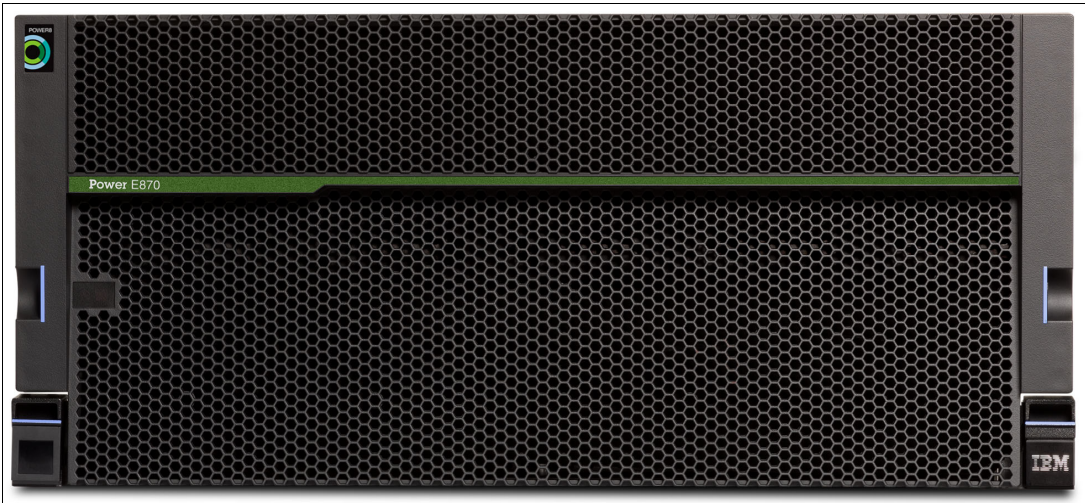


Figure 1-4 Front view of the system node

The rear view of the system node is shown in Figure 1-5 and notable features are highlighted.

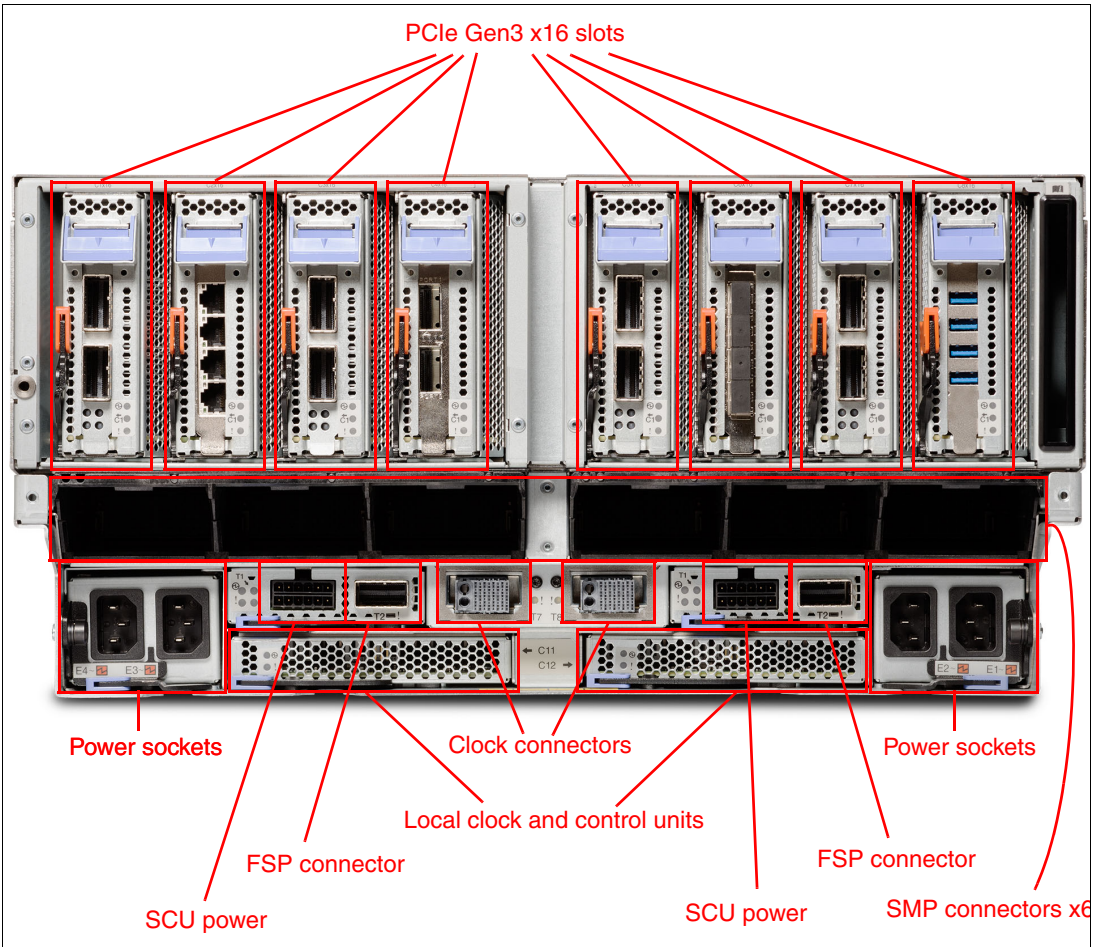


Figure 1-5 Rear view of the system node

The SMP cables on the Power E870C and E880C servers are fully flexible and do not restrict allowed rack specifications. How SMP cables can be routed within a rack on a Power E880C with all four system nodes installed is shown in Figure 1-6.

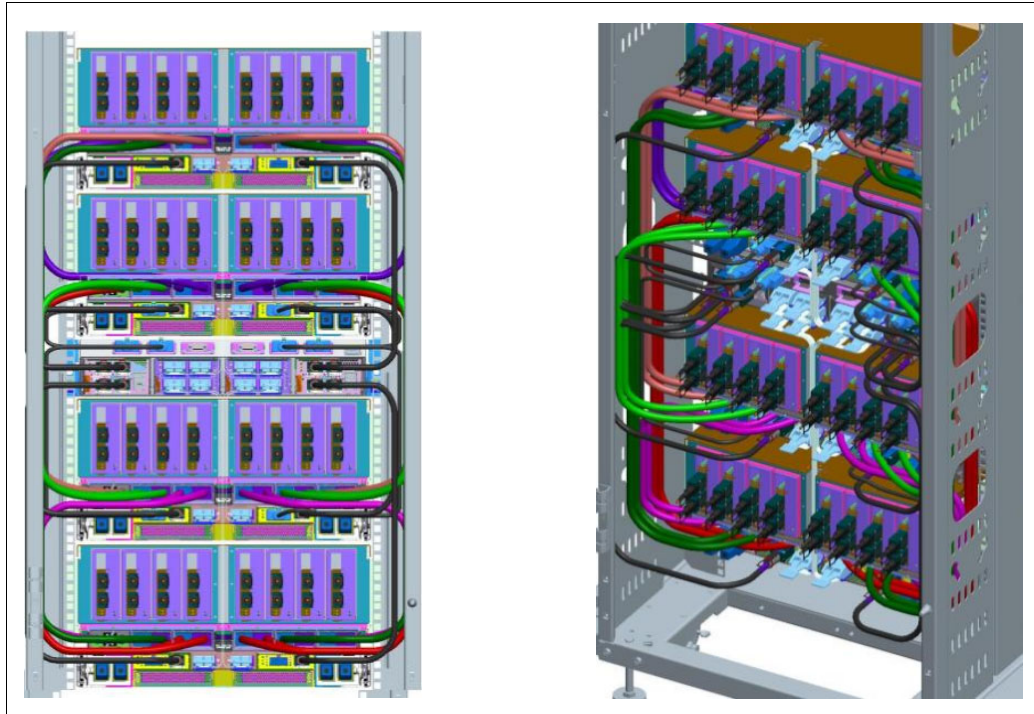


Figure 1-6 SMP cable routing in a four-drawer system

Figure 1-7 shows SMP cable routing in a two-drawer Power E870C or E880C.

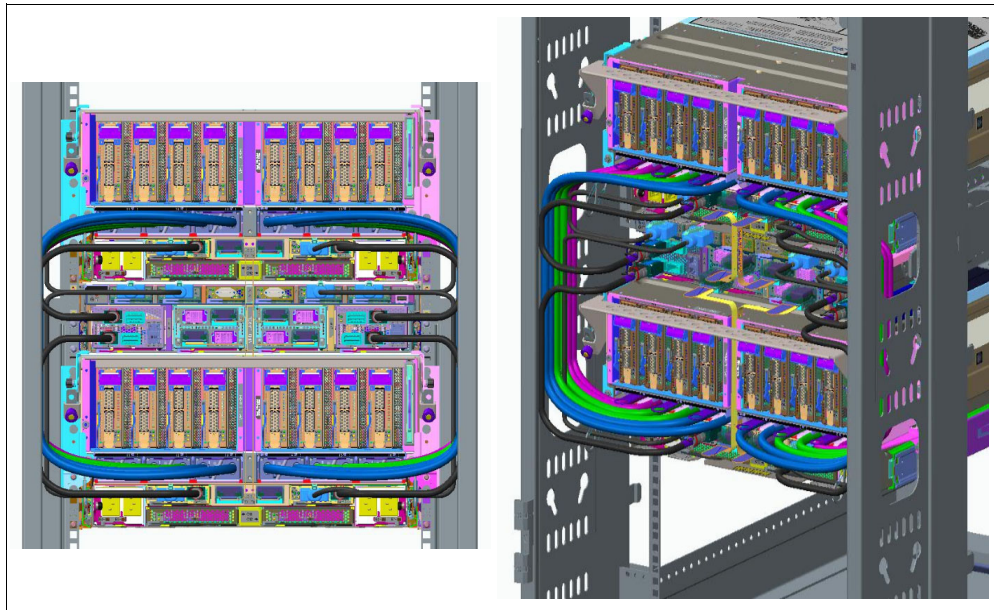


Figure 1-7 SMP cable routing in a two-drawer Power E870C or E880C

1.1.5 I/O drawers

I/O drawers provide more PCIe adapter slots and SAS disk slots for the Power E870C and E880C. For more information, see 1.6, “I/O drawers” on page 17.

1.2 Operating environment

The operating environment for the Power E870C and E880C servers is listed in Table 1-1.

Table 1-1 Operating environment for Power E870C and Power E880C

Power E870C and Power E880C operating environment				
System	Power E870C	Power E880C	Power E870C	Power E880C
	Operating		Non-operating	
Temperature	5 - 40 °C (41 - 104 °F) ⁷		5 - 45 °C (41 - 113 °F)	
Relative humidity	20 - 80%		8 - 80%	
Maximum dew point	29 °C (84 °F)		28 °C (82 °F)	
Operating voltage	200 - 240 V AC		N/A	
Operating frequency	47 - 63 Hz AC		N/A	
Maximum power consumption	4150 Watts per system node		N/A	
Maximum power source loading	4.20 kVA per system node		N/A	
Maximum thermal output	14,164 BTU/hour per system node		N/A	
Maximum altitude	3,048 m (10,000 ft.)		N/A	

For more information about operating noise levels, see *IBM Power Systems E870 and E880 Technical Overview and Introduction*, REDP-5137, which is available at this website:

<http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp5137.html>

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 IBM Systems Energy Estimator tool is available at this website:

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

1.3 Physical package

The physical dimensions of the SCU and of individual system nodes are listed in Table 1-2. Both servers are available only in a rack-mounted form factor.

Table 1-2 Physical dimensions of the Power E870C and Power E880C

Dimension	Power E870C or Power E880C system node	System control unit	PCIe Gen3 I/O expansion drawer
Width	445 mm (17.5 in.)	434 mm (17.1 in.)	482 mm (19 in.)
Depth	902 mm (35.5 in.)	813 mm (32.0 in.)	802 mm (31.6 in.)
Height	219 mm (8.6 in.) 5 EIA units	86 mm (3.4 in.) 2 EIA units	173 mm 6.8 in.) 4 EIA units
Weight	75.7 kg (167 lb)	23.6 kg (52.0 lb)	54.4 kg (120 lb)

The Power E870C is a modular system that can be constructed from a single SCU and one or two system nodes.

The Power E880C is a modular system that can be constructed from a single SCU and one, two, three, or four system nodes.

The SCU requires 2U and each system node requires 5U. 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.

The front view of a Power E870C SCU and system node is shown in Figure 1-8.



Figure 1-8 Power 870C SCU and system node

1.4 System features

The Power E870C and Power E880C system nodes contain four processor modules with 512 KB L2 cache and 8 MB L3 cache per core.

1.4.1 Power E870C system features

The following features are available on the Power E870C:

- ▶ One or two 5U 19-inch rack mount system node drawers
- ▶ One 2U 19-inch rack-mount SCU drawer
- ▶ 7U for a system with one system node drawer
- ▶ 12U for a system with two system node drawer
- ▶ One processor feature per system node: 4.02 GHz, (4 X 0/8-core) 32-core POWER8 processor (#EPBA)
- ▶ Static or mobile processor activation features available on a per-core basis
- ▶ Power IFL (Integrated Facility for Linux)
- ▶ POWER8 DDR3 or DDR4 memory CDIMMs (32 CDIMM slots per system node, 16 sites populated per system node minimum. For more information about #EM8Y restrictions, see 2.3.2, “Memory placement rules” on page 42):
 - 0/64 GB (4 X 16 GB), 1600 MHz (#EM8J) DDR3
 - 0/64 GB (4 X 16 GB), 1600 MHz (#EM8U) DDR4
 - 0/128 GB (4 X 32 GB), 1600 MHz (#EM8K) DDR3
 - 0/128 GB (4 X 32 GB), 1600 MHz (#EM8V) DDR4
 - 0/256 GB (4 X 64 GB), 1600 MHz (#EM8L) DDR3
 - 0/256 GB (4 X 64 GB), 1600 MHz (#EM8W) DDR4
 - 0/512 GB (4 x 128 GB), 1600MHz (#EM8M) DDR3
 - 0/512 GB (4 x 128 GB), 1600MHz (#EM8X) DDR4
 - 0/1024 GB (4 x 256 GB), 1600 MHz (#EM8Y) DDR4
- ▶ CoD memory activation features include:
 - 1 GB (static) Memory Activation (#EMA5)
 - 100 GB (static) Memory Activations (#EMA6)
 - 100 GB Mobile Memory Activations (#EMA7)
 - 100 GB Mobile Enabled Memory Activations (#EMA9)
 - 512 GB Memory Activations for IFL (#EMB8)
 - Plus activations for a few specific bundle scenarios
- ▶ IBM Active Memory™ Expansion, which is optimized onto the processor chip (#EM82)
- ▶ Eight PCIe Gen3 x16 I/O expansion slots per system node drawer (maximum 16 with 2-drawer system)
- ▶ One optional slim-line, SATA media bay per SCU enclosure (DVD drive selected as default with the option to clear)
- ▶ Redundant hot-swap ac power supplies in each system node drawer
- ▶ Two HMC 1 Gbps ports per Flexible Service Processor (FSP) in the SCU
- ▶ Dynamic logical partition (LPAR) support
- ▶ Processor and memory Capacity Upgrade on Demand (CUoD)
- ▶ 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 by way of two CXP Active Optical Cables (#ECC6, #ECC8, or #ECC9) to one PCIe3 Optical CXP Adapter (#EJ07)

1.4.2 Power E880C system features

The following features are available on the Power E880C:

- ▶ One, two, three, or four 5U 19-inch rack-mount system node drawers
- ▶ One 2U 19-inch rack-mount SCU drawer
- ▶ 7U for a system with one system node drawer plus one SCU
- ▶ 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. For more information about #EM8Y restrictions, see 2.3.2, “Memory placement rules” on page 42):
 - 0/64 GB (4 X 16 GB), 1600 MHz (#EM8J) DDR3
 - 0/64 GB (4 X 16 GB), 1600 MHz (#EM8U) DDR4
 - 0/128 GB (4 X 32 GB), 1600 MHz (#EM8K) DDR3
 - 0/128 GB (4 X 32 GB), 1600 MHz (#EM8V) DDR4
 - 0/256 GB (4 X 64 GB), 1600 MHz (#EM8L) DDR3
 - 0/256 GB (4 X 64 GB), 1600 MHz (#EM8W) DDR4
 - 0/512 GB (4 x 128 GB), 1600MHz (#EM8M) DDR3
 - 0/512 GB (4 x 128 GB), 1600MHz (#EM8X) DDR4
 - 0/1024 GB (4 x 256 GB), 1600 MHz (#EM8Y) DDR4
- ▶ Active Memory Expansion, which is 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 optional slim-line, SATA media bay per SCU enclosure (DVD drive defaulted on order, option to clear)
- ▶ Redundant hot-swap AC power supplies in each system node drawer

- ▶ Two HMC 1 Gbps ports per FSP in the SCU
- ▶ Dynamic logical partition (LPAR) support
- ▶ Processor and memory CUoD
- ▶ 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 by way of two CXP Active Optical Cables (#ECC6, #ECC8, or #ECC9) to one PCIe3 Optical CXP Adapter (#EJ07)

1.4.3 Minimum features

Each Power E870C or Power E880C initial order must include a minimum of the following items:

- ▶ 1 x system node with a choice of the following processors:
 - 4.02 GHz, 32-core POWER8 processor module (#EPBA) for Power E870C
 - 4.35 GHz, 32-core POWER8 processor module (#EPBB) for Power E880C
 - 4.19 GHz, 40-core POWER8 processor module (#EPBS) for Power E880C
 - 4.02 GHz, 48-core POWER8 processor module (#EPBD) for Power E880C
- ▶ 8 x 1 core processor activation:
 - 1 core permanent Processor Activation for #EPBA (#EPBJ) for Power E870C
 - 1 core permanent Processor Activation for #EPBB (#EPBK) for Power E880C
 - 1 core permanent Processor Activation for #EPBS (#EPBU) for Power E880C
 - 1 core permanent Processor Activation for #EPBD (#EPBM) for Power E880C
- ▶ 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 that uses the following 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 E870C
 - IBM Rack-mount Drawer Bezel and Hardware (#EBA3) for Power E880C
- ▶ 1 x System Node to SCU 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 AIX or Linux are the primary operating systems, the order also must include a minimum of the following 1 x Primary Operating System Indicators:

- ▶ 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)

Note: Consider the following points:

- ▶ More optional features can be added, as needed.
- ▶ IBM i systems require a DVD to be available to the system when required. This DVD can be in the SCU (DVD feature #EU13) or it can be external in an enclosure, such as 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 can be used to substitute for a DVD device.
- ▶ Feature-coded racks are allowed for I/O expansion only.
- ▶ A machine type or model rack (if wanted) must 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. A total of 25% of the available memory must be permanently activated by using #EMA5 and #EMA6. The remaining 75% of the active memory can be enabled by using permanent or mobile options.
- ▶ Memory sizes can differ across CPU modules, but the eight CDIMM slots that are connected to the same processor module must be filled with identical CDIMMs.
- ▶ All memory operates at 1600 MHz.
- ▶ 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 SCU on each system. The SCU is considered the system with the serial number.
- ▶ An HMC is required for management of every Power E870C or Power E880C; however, a shared HMC is acceptable. HMCs supported on POWER8 hardware are 7042-C08 and 7042-CR5 through 7042-CR9 and the virtual appliance HMC. For more information, see 1.11, “Management consoles” on page 24.

1.4.4 Power supply features

This section describes how the system nodes and SCUs are powered.

System node power

Four AC 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 power cords are used for each system node (one per power supply) and are ordered by 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 that are in the front of the system node where they are more accessible for service.

SCU power

The SCU is powered from the system nodes. UPIC cables provide redundant power to the SCU. 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 provide N+1 redundant power to the SCU.

1.4.5 Processor card features

Each system node delivers a set of four identical processors. All processors in the system must be identical. Cable features are required to connect system nodes to the SCU and to other system nodes. Consider the following points:

- ▶ #ECCA is required for a single system node configuration. #ECCA provides cables to connect the system node with the SCU.
- ▶ #ECCB is required for a dual system node configuration. #ECCB provides cables to connect the system nodes with the SCU and cables to connect the two system nodes.
- ▶ For three or four system node systems, no other cables are required because redundant power is provided through system nodes 1 and 2.

Each system must have a minimum of eight active cores.

The Power E870C has one type of processor, which offers the 4.02 GHz, (4 X 0/8-core) 32-core POWER8 processor (#EPBA) feature.

The Power E880C has three types of processors, which offer 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 E870C and Power 880 servers to help meet changing resource requirements in an on-demand environment by using resources that are installed on the system but not activated. CoD allows you to purchase extra permanent processor or memory capacity and dynamically activate it when needed. The #EPJ3 provides no-charge elastic processor days for Power E880C. The #EMJ8 feature provides no-charge elastic memory days for Power E880C.

1.5 Disk and media features

The Power E870C and Power E880C SCU and system nodes do not support internal disks. Any required disk must be 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 E870C or Power E880C.

Each SCU enclosure has one slim-line bay, which can support one DVD drive (#EU13). The #EU13 DVD is cabled to a USB PCIe adapter that is in 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 in the SCU (DVD feature #EU13) or it can be external in an enclosure, such as 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 can be used to substitute for a DVD device if VIOS is used.

1.6 I/O drawers

If more Gen3 PCIe slots that are 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) also are supported, which provides 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 12 PCIe I/O full-length, full-height slots. One FanOut Module provides six PCIe slots (labeled C1 - 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 fit into these slots. The BSC is the same BSC that was used with the previous generation server's 12X attached I/O drawers (#5802, #5803, #5877, and #5873). The drawer is shipped with a full set of BSC, even if the BSC is empty.

Concurrent repairing, adding, and removing 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 includes two AOC cables. Each PCIe Gen3 I/O expansion drawer features two power supplies.

Each system node supports zero, one, two, three, or four PCIe Gen3 I/O expansion drawers. A half drawer, which consists of one PCIe fan-out module in the I/O drawer, also is supported. This drawer allows a lower-cost configuration if fewer PCIe slots are required.

A system node supports the following half drawer options:

- ▶ One half drawer
- ▶ Two half drawers
- ▶ Three half drawers
- ▶ 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 also can 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 is used.

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

A PCIe Gen3 I/O expansion drawer is shown in Figure 1-9.



Figure 1-9 PCIe Gen3 I/O expansion drawer

For more information about connecting PCIe Gen3 I/O expansion drawers to the Power E870C and Power E880C servers, see 2.9.1, “PCIe Gen3 I/O expansion drawer” on page 74.

1.6.2 I/O drawers and usable PCI slot

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 is shown in Figure 1-10.

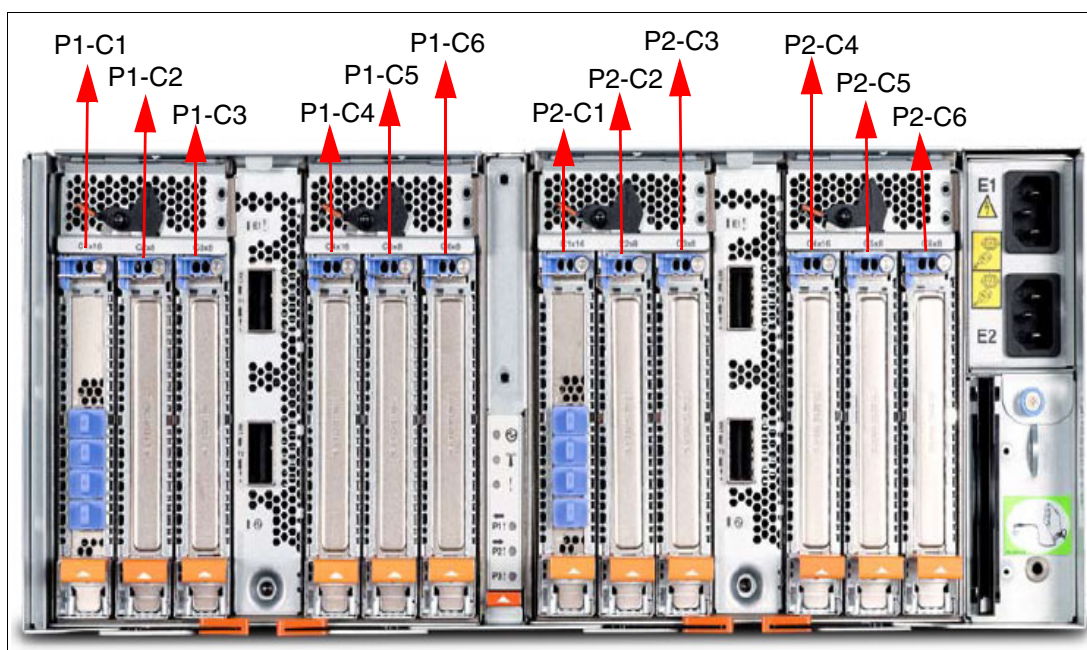


Figure 1-10 Rear view of a PCIe Gen3 I/O expansion drawer with PCIe slots location codes

The PCI slots in the PCIe Gen3 I/O expansion drawer are listed in Table 1-3.

Table 1-3 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

Consider the following points:

- ▶ 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.

The maximum number of I/O drawers that is supported and the total number of PCI slots that are available when expansion consists of a single drawer type is listed in Table 1-4.

Table 1-4 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 must change the mode after installation, ask your IBM support representative to refer to the following website:

<http://w3.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5121>

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

The EXP24S SFF Gen2-bay drawer is shown in Figure 1-11.

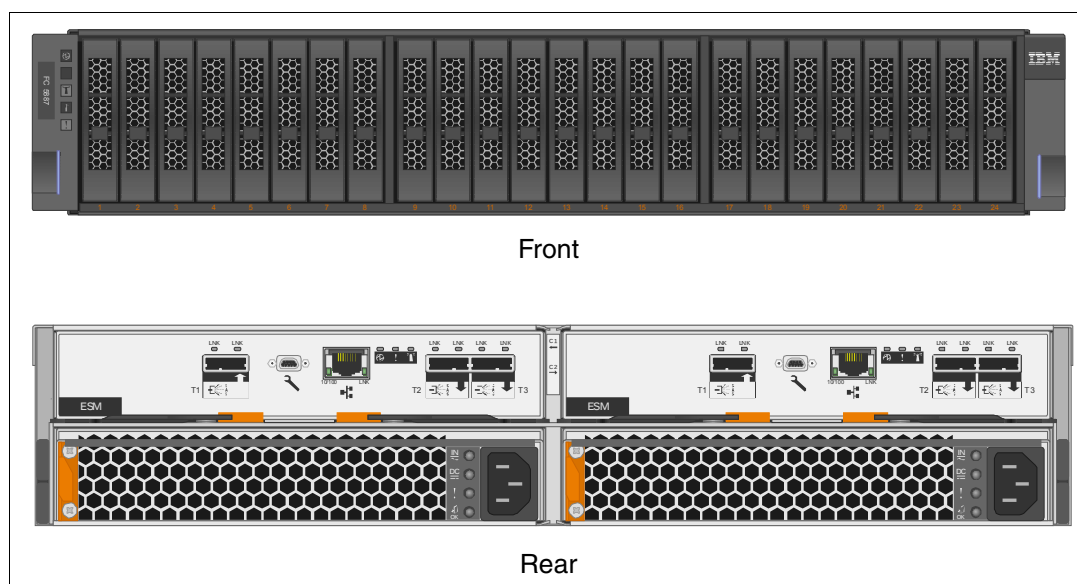


Figure 1-11 EXP24S SFF Gen2-bay drawer

For more information about connecting the EXP24S Gen2-bay drawer to the Power E870C and Power E880C servers, see 2.10.1, “EXP24S SFF Gen2-bay Drawer” on page 81.

1.8 Model comparison

The Power E870C offers configuration options in which the POWER8 processor can have one of two processor speeds. Each system node includes four single chip modules (SCMs). Each system node contains the four 4.02 GHz 8-core SCMs processor configuration.

A Power E870C with either of the processor configurations can have as few as eight cores that are 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 E870C can be installed with one or two system nodes that are connected to the SCU. Each system node can have up to 8 TB of memory installed.

The Power E880C also offers system nodes that include four SCMs. The Power E880C system node has the following two processor configurations:

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

A Power E880C with either of the processor configurations can have as few as eight cores that are activated or up to 100% of the cores can be activated. Incrementing one core at a time is available through built-in CoD functions to the full capacity of the system. The Power E880C can be installed with one, two, three, or four system nodes that are connected to the SCU. Each system node can have up to 8 TB of memory installed.

The processor and memory maximums for the Power E870C and Power E880C are listed in Table 1-5.

Table 1-5 Summary of processor and memory maximums for the Power E870C and Power E880C

System	Cores per SCM	Core speed	System node core maximum	System core maximum	System node memory maximum	System memory maximum
Power E870C	8	4.02 GHz	32	64	8 TB	16 TB
Power E880C	8	4.35 GHz	32	128	8 TB	32 TB
Power E880C	10	4.19 GHz	40	160	8 TB	32 TB
Power E880C	12	4.02 GHz	48	192	8 TB	32 TB

1.9 Build to order

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

This method is the only configuration method that is available for the Power E870C and Power E880C servers.

IBM editions: IBM edition offerings are not available for the Power E870C and Power E880C servers.

1.10 Model upgrades

The following sections describe the various upgrades that are available.

1.10.1 Power E870C

A model conversion from a Power 770 (9117-MMD) to a Power E870C (9080-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 components that are being replaced during a model or feature conversion become the property of IBM and must be returned.

Note: There is no model conversion available from a 9119-MME to a 9080-MME.

1.10.2 Power E880C

A model conversion from a Power 780 (9179-MHD) to a Power E880C (9080-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 E880C (9080-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 E880C from a Power 795 (9119-FHB) are not available.

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

Note: There is no model conversion available from a 9119-MHE to a 9080-MHE.

1.10.3 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
- ▶ IBM 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 E870C and E880C:

- ▶ 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 you place the upgrade MES order for the new Power E870C or E880C system to keep the elastic CoD function active. To start 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 that is owed to IBM on the system must be paid before installing the Power E870C or E880C.

1.11 Management consoles

This section describes the supported management interfaces for the servers.

The Hardware Management Console (HMC) is required for managing the IBM Power E870C and Power E880C. It features the following 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

The HMC can be a hardware or a virtual appliance. The available configurations for a Power System HMC infrastructure are shown in Figure 1-12.

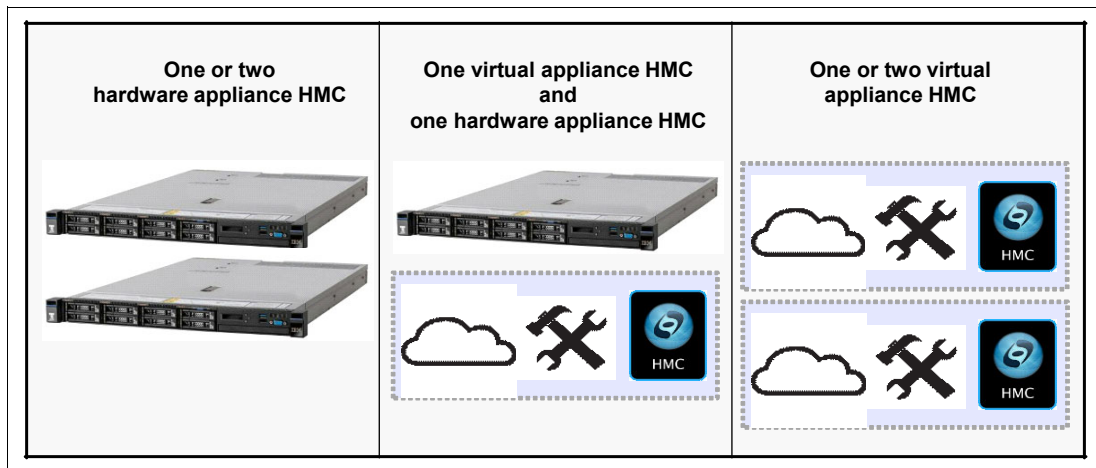


Figure 1-12 HMC configurations

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.

Note: The IBM Power E870C and Power E880C are not supported by the Integrated Virtualization Manager (IVM).

Several hardware appliance HMC models are supported to manage POWER8 based systems. The 7042-CR9 is available for ordering at the time of this writing, but you also can use one of the withdrawn models that are listed in Table 1-6 on page 25.

Table 1-6 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 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, the HMC requires a memory upgrade to a minimum of 4 GB.

Requirements: Consider the following HMC code and system firmware requirements:

- ▶ HMC base Licensed Machine Code Version 850.10, or later is required to support the Power E870C (9080-MME) and Power E880C (9080-MHE).
- ▶ System firmware level 840.30, or later.

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

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

For more information about managing the Power E870C and Power E880C servers from an HMC, see 2.11, “Hardware Management Console” on page 93.

1.12 System racks

The Power E870C and E880C and its I/O drawers are mounted in the following IBM racks:

- ▶ 7014-T00
- ▶ 7014-T42
- ▶ #0551
- ▶ #0553

Order the Power E870C and E880C server with an IBM 42U enterprise rack (7014-T42 or #0553). An initial system order is placed in a 7014-T42 rack. A same serial number model upgrade MES is placed in an equivalent #0553 rack. This placement is done to ease and speed client installation, which provides a complete and higher-quality environment for IBM Manufacturing system assembly and testing and a complete shipping package.

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 then ships as separate packages for installation into a rack.

The Power E870C and Power E880C use a new type of connector between system drawers. Therefore, the systems do not require wider racks and an OEM rack or cabinet that meets the requirements can be used.

Installing in non-IBM racks: You can choose to place the server in other racks if you are confident those racks have the strength, rigidity, depth, and hole pattern characteristics that are needed. Clients should work with IBM Service to determine other racks' appropriateness.

Compared to the existing Power E850, the E850C server draws more power. Using previously provided IBM power distribution unit (PDU) features #7188, #7109, and #7196 reduces the number of E850C servers and other equipment that can be held most efficiently in a rack. The new high-function PDUs provide more electrical power per PDU and thus offer better "PDU footprint" efficiency. In addition, they are intelligent PDUs that provide insight to actual power usage by receptacle and also provide remote power on/off capability for easier support by individual receptacle.

The new PDUs are features #EPTJ, #EPTL, #EPTN, and #EPTQ available on the 7014-T00, 7014-T42, and 7965-94Y racks. For more information, see 1.12.5, "AC power distribution unit and rack content" on page 28.

1.12.1 IBM 7014 model T00 rack

The 1.8-meter (71-inch) model T00 is compatible with past and present IBM Power systems servers. The T00 rack includes the following features:

- ▶ 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 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:
 - 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.), which equates to 13.6 kg (30 lb.)/EIA.

1.12.2 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.
- ▶ The model T42 supports AC power only.

- ▶ **Weights:**
 - T42 base empty rack: 261 kg (575 lb.)
 - T42 full rack: 930 kg (2045 lb.)
- ▶ The following five rack front door options are supported for the 42U enterprise rack (7014-T42 or #0553):
 - The original acoustic door (#6249 [front and back] or #ERGB [front only]).
 - The newer thinner acoustic door (#EC07/#EC08).
 - The ruggedized door (#ERGD).
 - The attractive geometrically accented black full-height rack door (#ERG7). 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 cost-effective plain black front door (#6069).
 - The front trim kit is also supported (#6272).

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

When considering an acoustic door, note that the majority of the acoustic value is provided by the front door because the servers' fans are mostly in the front of the rack. The absence of a rear acoustic door saves some floor space, which might make it easier to use.

- ▶ A minimum of one of the following features is required:
 - Feature #ER2B reserves 2U of space at the bottom of the rack.
 - Feature #ER2T reserves 2U of space at the top of the rack.

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 recommended where large numbers of thicker I/O cables are present or might be added in the future. The definition of a large number depends on the type of I/O cables that are used; approximately 64 short-length SAS cables per side of a rack or around 50 longer-length (thicker) SAS cables per side of a rack is recommended.

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) that are driven by multiple PCIe SAS adapters. For this reason, it can be a 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 by using shorter, thinner SAS cables. The feature ERG0 extension can be good to use even with a smaller number of cables as it enhances the ease of cable management with the extra space it provides.

Recommended Rack: The 7014-T42 System rack with extra rear extension (#ERG0) is recommended for all initial Power E870C and Power E880C system orders. These options are the default options for new orders.

1.12.3 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 that is 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 E870C and E880C servers.

1.12.4 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 that is 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.

1.12.5 AC power distribution unit and rack content

For rack models T00 and T42, 12-outlet PDUs are available. The following PDUs are available:

- ▶ PDUs Universal UTG0247 Connector (#7188)
- ▶ Intelligent PDU+ Universal UTG0247 Connector (#7109)
- ▶ High Function 9xC19 PDU: Switched, Monitoring (#EPTJ)
- ▶ High Function 9xC19 PDU 3 Phase: Switched, Monitoring (#EPTL)
- ▶ High Function 12xC13 PDU: Switched, Monitoring (#EPTN)
- ▶ High Function 12xC13 PDU 3 Phase: Switched, Monitoring (#EPTQ)

PDU mounting: Only horizontal PDUs are allowed in racks that are hosting Power E870C and Power E880C 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. This placement leaves 2U or more of space at the top or bottom of the rack for cable management. Mounting a horizontal PDU in the middle of the rack is not optimal for cable management.

Feature #7109

Intelligent PDU with Universal UTG0247 Connector is for an intelligent AC power distribution unit (PDU+) that allows the user to monitor the amount of power that is being used by the devices that are plugged in to this PDU+. This AC power distribution unit provides 12 C13 power outlets. It receives power through a UTG0247 connector. It can be used for many different countries and applications by varying the PDU to Wall Power Cord, which must be ordered separately. Each PDU requires one PDU to Wall Power Cord. Supported power cords include the following features: 6489, 6491, 6492, 6653, 6654, 6655, 6656, 6657, and 6658.

Feature #7188

Power Distribution Unit mounts in a 19-inch rack and provides 12 C13 power outlets. Feature 7188 has six 16A circuit breakers, with two power outlets per circuit breaker. System units and expansion units must use a power cord with a C14 plug to connect to the feature 7188. One of the following power cords must be used to distribute power from a wall outlet to the feature 7188: Feature 6489, 6491, 6492, 6653, 6654, 6655, 6656, 6657, or 6658.

Feature #EPTJ

This PDU is an intelligent, switched 200 - 240 volt AC PDU with nine C19 receptacles on the front of the PDU. The PDU is mounted on the rear of the rack, which makes the nine C19 receptacles easily accessible.

Each receptacle has a 20 amp circuit breaker. Depending on country wiring standards, the PDU is single-phase or three-phase wye. (See three-phase feature #EPTK or #EPTL for countries that do not use wye wiring.)

The PDU can be mounted vertically in rack side pockets, or it can be mounted horizontally. If mounted horizontally, it uses 1 EIA (1U) of rack space. (See feature #EPTH for horizontal mounting hardware.)

Device power cords with a C20 plug connect to C19 PDU receptacles and are ordered separately. One country-specific wall line cord also is ordered separately and attaches to a UTG524-7 connector on the front of the PDU.

Supported line cords include features #6489, #6491, #6492, #6653, #6654, #6655, #6656, #6657, #6658, and #6667.

Feature #EPTL

This PDU is an intelligent, switched 208 volt 3-phase AC PDU with nine C19 receptacles on the front of the PDU. The PDU is mounted on the rear of the rack, which makes the nine C19 receptacles easily accessible. Each receptacle has a 20 amp circuit breaker.

The PDU can be mounted vertically in rack side pockets, or it can be mounted horizontally. If mounted horizontally, it uses 1 EIA (1U) of rack space. (See feature #EPTH for horizontal mounting hardware.)

Device power cords with a C20 plug connect to C19 PDU receptacles and are ordered separately. One wall line cord is provided with the PDU (no separate feature number) and includes an IEC60309 60A plug (3P+G).

The PDU supports up to 48 amps. Two RJ45 ports on the front of the PDU enable the client to monitor each receptacle's electrical power usage and to remotely switch any receptacle on or off.

The PDU includes a generic PDU password, which IBM strongly urges clients to change upon installation.

Feature #EPTN

This PDU is an intelligent, switched 200 - 240 volt AC PDU with 12 C13 receptacles on the front of the PDU. The PDU is mounted on the rear of the rack, making the C13 receptacles easily accessible. Each receptacle has a 20 amp circuit breaker. Depending on country wiring standards, the PDU is single-phase or three-phase wye. (See three-phase feature #EPTK or #EPTL for countries that do not use wye wiring.)

The PDU can be mounted vertically in rack side pockets, or it can be mounted horizontally. If mounted horizontally, it uses 1 EIA (1U) of rack space. (See feature #EPTH for horizontal mounting hardware.)

Device power cords with a C14 plug connect to C13 PDU receptacles and are ordered separately. One country-specific wall line cord also is ordered separately and attaches to a UTG524-7 connector on the front of the PDU.

Supported line cords include features #6489, #6491, #6492, #6653, #6654, #6655, #6656, #6657, #6658, and #6667. Two RJ45 ports on the front of the PDU enable the client to monitor each receptacle's electrical power usage and to remotely switch any receptacle on or off.

The PDU includes a generic PDU password, which IBM strongly urges clients to change upon installation.

Feature #EPTQ

This PDU is an intelligent, switched 208 volt 3-phase AC PDU with 12 C13 receptacles on the front of the PDU. The PDU is mounted on the rear of the rack, which makes the 12 C13 receptacles easily accessible. Each receptacle has a 20 amp circuit breaker.

The PDU can be mounted vertically in rack side pockets, or it can be mounted horizontally. If mounted horizontally, it uses 1 EIA (1U) of rack space. (See feature EPTH for horizontal mounting hardware.)

Device power cords with a C14 plug connect to C13 PDU receptacles and are ordered separately. One wall line cord is provided with the PDU (no separate feature number) and includes an IEC60309 60A plug (3P+G).

The PDU supports up to 48 amps. Two RJ45 ports on the front of the PDU enable the client to monitor each receptacle's electrical power usage and to remotely switch any receptacle on or off.

The PDU includes a generic PDU password, which IBM strongly urges clients to change upon installation.

For more information about power cord requirements and power cord feature codes, see the following IBM Power Systems Hardware IBM Knowledge Center website:

http://www.ibm.com/support/knowledgecenter/TI0003M/p8had/p8had_specsheetpdu.htm

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

1.12.6 Rack-mounting rules

For the Power E870C or E880C that is installed in IBM 7014 or FC #055x racks, the following PDU rules apply:

- For PDU #7109 and #7188 when 24 Amp power cord #6654, #6655, #6656, #6657, or #6658 is used, each pair of PDUs can power one Power E870C or Power E880C system node and two I/O expansion drawers, or eight I/O expansion drawers. The 24A PDU cables are used to supply 30A PDUs. The rack configuration with two pairs of 30A PDUs supplying a two system node configuration and two I/O expansion drawers is shown in Figure 1-13.

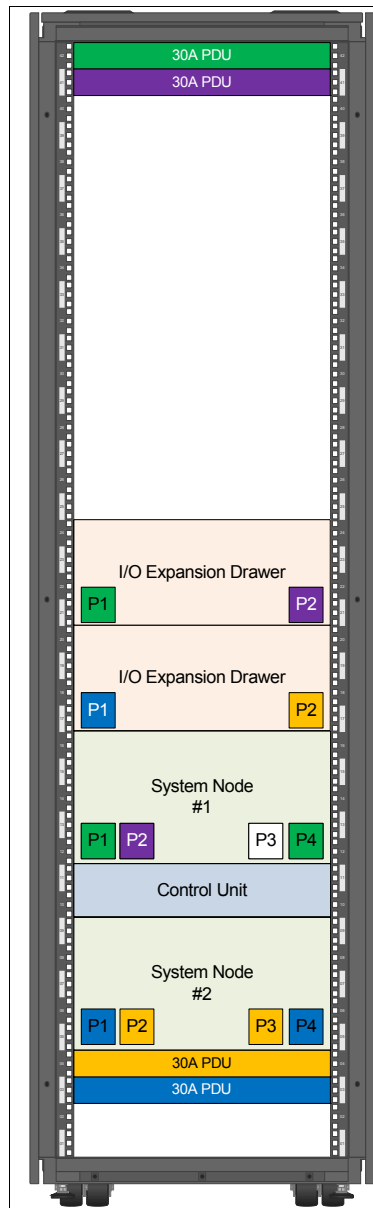


Figure 1-13 Two system node configuration and two I/O expansion drawers supplied by 30A PDUs

- For PDU #7109 and #7188 when three-phase power cords or 48 Amp power cords #6491 or #6492 is used, each pair of PDUs can power up to two Power E870C or Power E880C system nodes and two I/O expansion drawers, or eight I/O expansion drawers. The 48A PDU cables are used to supply 60A PDU. The rack configuration with two pairs of 60A PDUs supplying a two system node configuration, four I/O expansion drawers, and four EXP24S disk drawers is shown in Figure 1-14.

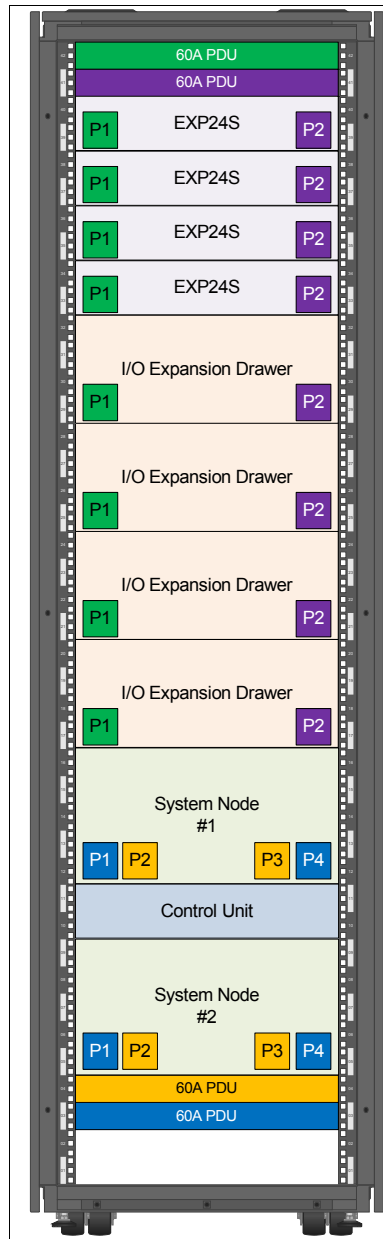


Figure 1-14 Two system nodes configuration with I/O expansions and disk drawers supplied by 60A PDUs

For more information about power cord requirements and power cord feature codes, see the following IBM Power Systems Hardware Knowledge Center website:

http://www.ibm.com/support/knowledgecenter/TI0003M/p8had/p8had_specsheetpdu.htm

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

For rack-integrated systems, a minimum quantity of two PDUs (#7109, #7188, or #7196) is required. These PDUs 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. These outlets are six groups of two outlets that are 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 kilovolt ampere (kVA). The total kVA of all the drawers that are plugged into the PDU must not exceed the power cord limitation.

Each system node that is 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.



Architecture and technical overview

This chapter describes the overall system architecture for the IBM Power System E870C (9080-MME) and IBM Power System E880C (9080-MHE) servers. The bandwidths that are provided throughout this chapter 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 size the performance at the application workload environment level and evaluate performance by using real-world performance measurements and production workloads.

This chapter includes the following topics:

- ▶ 2.1, “Logical diagrams” on page 36
- ▶ 2.2, “IBM POWER8 processor” on page 40
- ▶ 2.3, “Memory subsystem” on page 41
- ▶ 2.4, “Capacity on Demand” on page 53
- ▶ 2.5, “System bus” on page 59
- ▶ 2.6, “Internal I/O subsystem” on page 63
- ▶ 2.7, “PCI adapters” on page 65
- ▶ 2.8, “Internal storage” on page 72
- ▶ 2.9, “External I/O subsystems” on page 74
- ▶ 2.10, “External disk subsystems” on page 81
- ▶ 2.11, “Hardware Management Console” on page 93
- ▶ 2.12, “Operating system support” on page 96
- ▶ 2.13, “Energy management” on page 99

2.1 Logical diagrams

This section contains logical diagrams for the Power E870C and E880C.

The logical system diagram for a single system node of a Power E870C or Power E880C is shown in Figure 2-1.

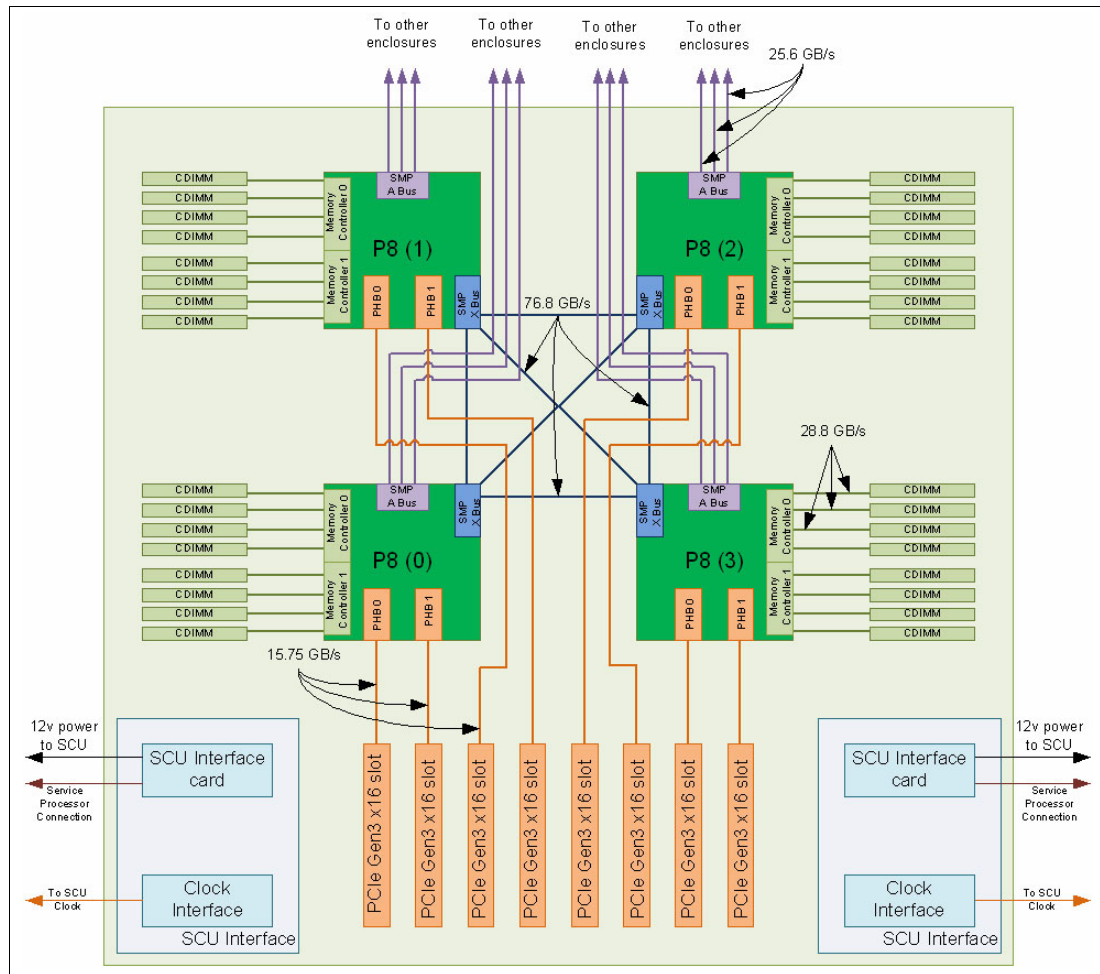


Figure 2-1 Logical system diagram for a system node of a Power E870C or a Power E880C

The logical system diagram for the system control unit of a Power E870C or a Power E880C is shown in Figure 2-2.

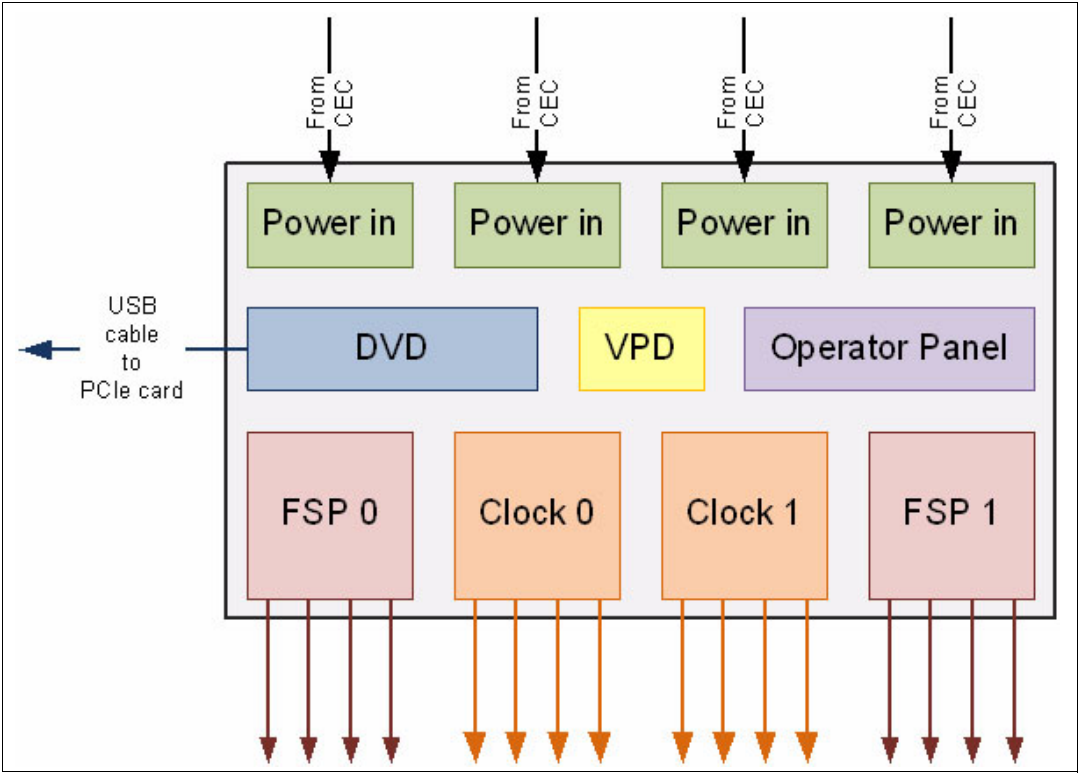


Figure 2-2 Logical system diagram for the system control unit

Flexible symmetric multiprocessing (SMP) cables are used to connect system nodes when a Power E870C or Power E880C is configured with more than one system node. The SMP connection topology for a two-drawer Power E870C or Power E880C system is shown in Figure 2-3.

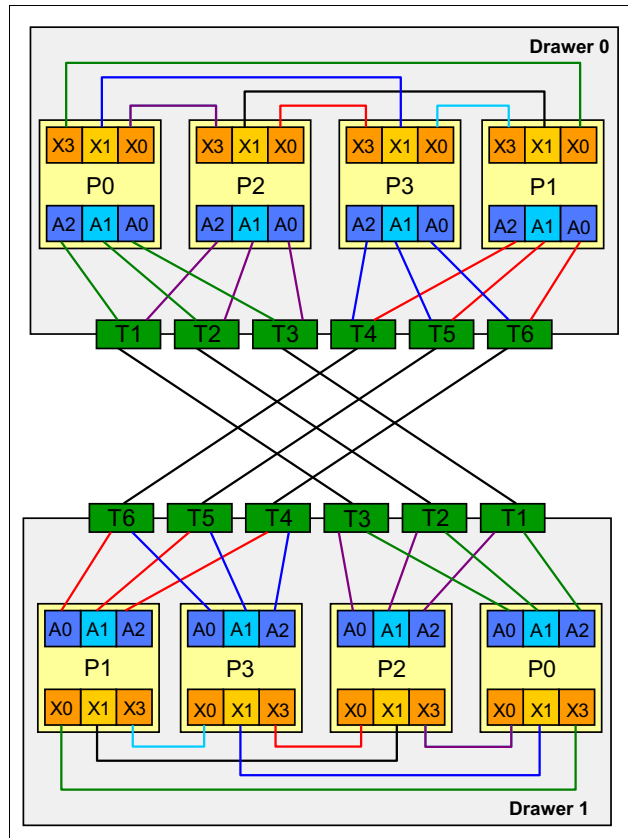


Figure 2-3 SMP connection topology for a two-drawer Power E870C or Power E880C

The SMP connection topology for a three-drawer Power E880C is shown in Figure 2-4.

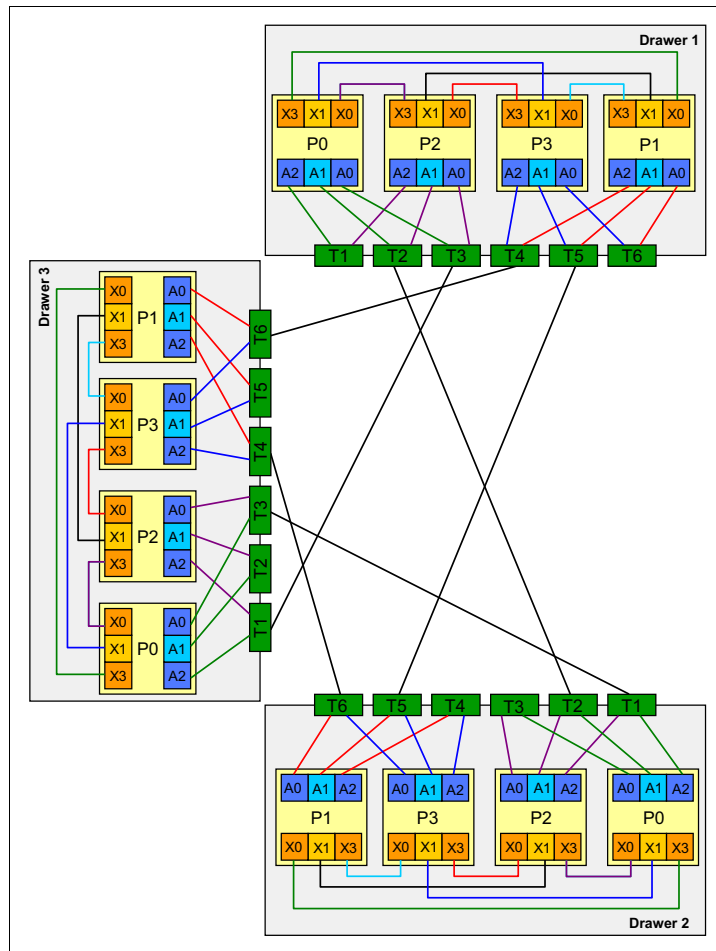


Figure 2-4 SMP connection topology for a three-drawer Power E880C

The SMP connection topology for a four-drawer Power E880C Figure 2-5.

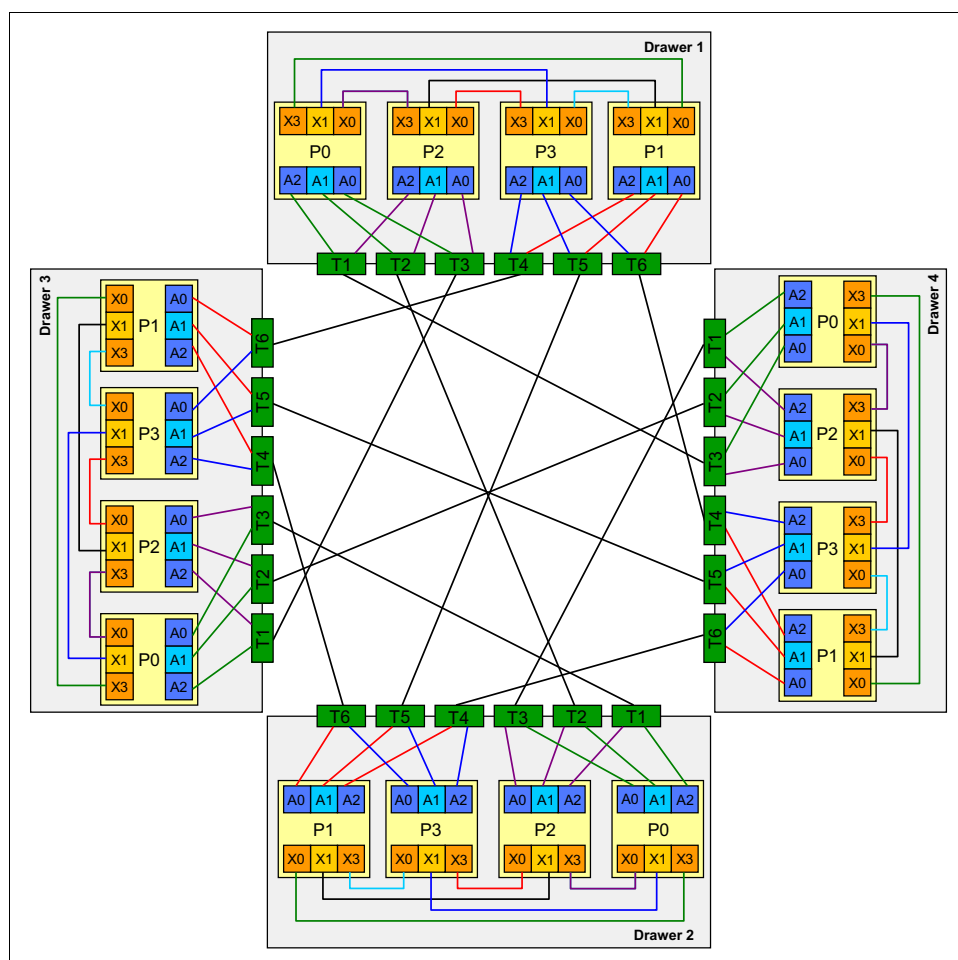


Figure 2-5 SMP connection topology for a four-drawer Power E880C

2.2 IBM POWER8 processor

The POWER8 processor is manufactured by using the IBM 22 nm Silicon-On-Insulator (SOI) technology. Each chip is 649 mm² and contains 4.2 billion transistors. The chip also 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 might 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 and 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.

For more information about the POWER8 processor, see *IBM Power Systems E870 and E880 Technical Overview and Introduction*, REDP-5137, which is available at this website:

<http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp5137.html>

2.3 Memory subsystem

The Power E870C can include up to two system nodes per system with each system node having 32 CDIMM slots that can support 16 GB, 32 GB, 64 GB, 128 GB, and 256 GB CDIMMs that are running at speeds of 1600 MHz. This configuration allows for a maximum system memory of 16 TB for the 256 CDIMM slots in a system that is composed of two system nodes.

The Power E880C can have up to four system nodes per system with each system node having 32 CDIMM slots that can support 16 GB, 32 GB, 64 GB, 128 GB, and 256 GB CDIMMs that are running at speeds of 1600 MHz. This configuration allows for a maximum system memory of 32 TB for the 256 CDIMM slots of a system that is composed of four system nodes.

The memory on the systems is Capacity on Demand capable with which more physical memory capacity can be added and dynamically activated when needed. (At least 50% of the installed memory capacity must be active.)

The Power E870C and E880C servers support an optional feature that is named Active Memory Expansion (#EM82). This feature allows the effective maximum memory capacity to be larger than the true physical memory.

This feature also runs innovative compression and decompression of memory content by using a dedicated coprocessor that is present on each POWER8 processor to provide memory expansion up to 125%, depending on the workload type and its memory usage.

For example, a server with 256 GB of memory that is physically installed effectively can 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 the following set of components that allow for higher bandwidth, lower latency communications, and increased availability:

- ▶ 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 ability saves bandwidth and allows for faster processor-to-memory communications, which also allows for a more robust RAS. For more information, see Chapter 4, “Reliability, availability, serviceability, and manageability” on page 115.

The CDIMMs are in two different form factors: A 152 SDRAM design that is named the Tall CDIMM and an 80 SDRAM design that is named the Short CDIMM. Each design is composed of multiple 4 GB SDRAM devices depending on its total capacity. The CDIMM slots for the Power E870C and Power E880C are tall CDIMMs slots. A filler is added to the short CDIMM, which allows 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.

The CDIMMs that are available for the Power E870C and Power E880C are shown in Figure 2-6.

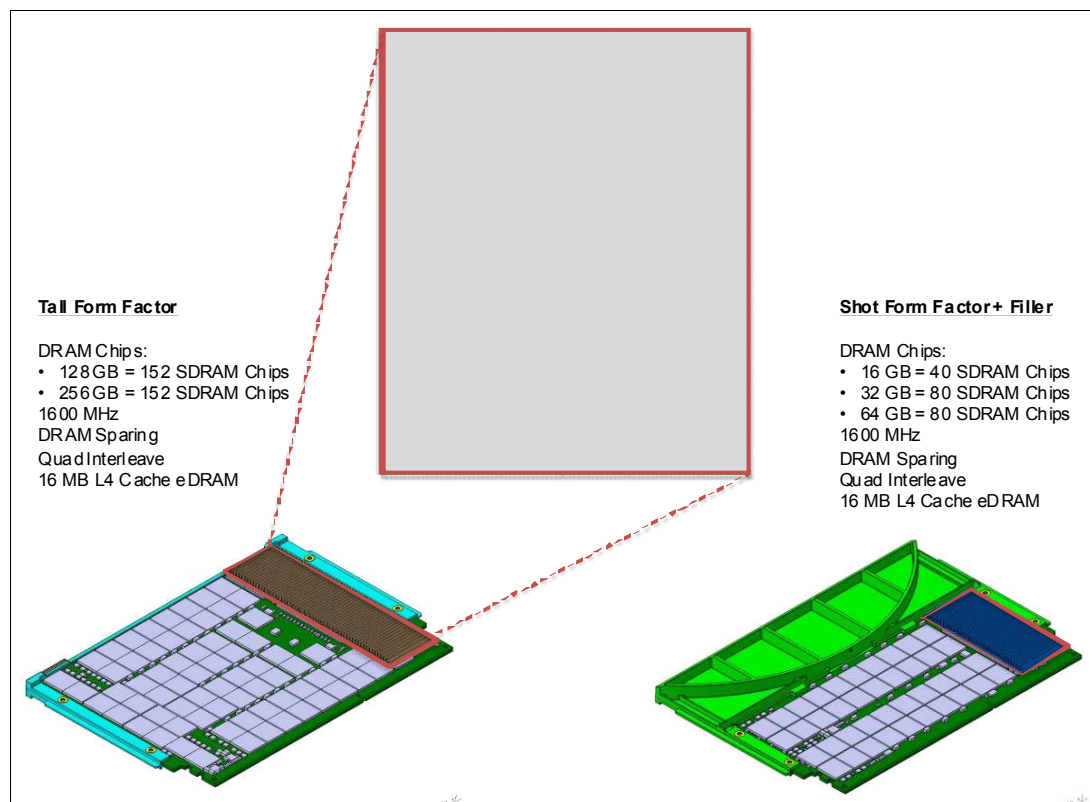


Figure 2-6 Short CDIMM and Tall CDIMM details

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 E870C server has 1 GB of L4 Cache while a fully populated Power E880C has 2 GB of L4 Cache. The L4 Cache performs the following functions that directly affect performance and realize a series of benefits for the Power E870C and Power E880C:

- ▶ 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, which becomes 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 E870C and Power E880C, each memory feature code provides four CDIMMs. Therefore, a maximum of eight memory feature codes per system node are allowed so that all 32 CDIMM slots are filled.

All of the memory CDIMMs' capacity can be upgraded 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.

The following memory options are orderable for the Power E870C and Power E880C:

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

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*. The POWER8 processor with its two memory quads is shown in Figure 2-7.

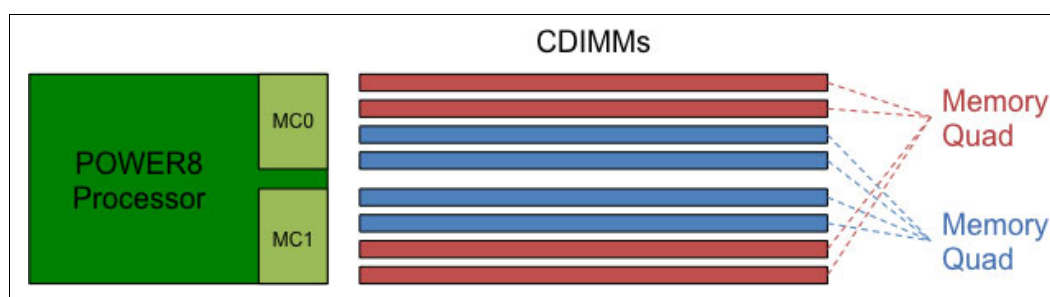


Figure 2-7 Logical diagram of a POWER8 processor and its two memory quads

Consider the following basic rules for memory placement:

- ▶ Each feature code equals a set of four physical CDIMMs, which is 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 specific processor can have only four or eight CDIMMs attached to it.
- ▶ All of the CDIMMs that are connected to the same POWER8 processor must be identical. However, mixing different CDIMM sizes between different POWER8 processors on a system is permitted.
- ▶ At least 50% of the installed memory must be activated via memory activation features.

Note: At the time of this writing the DDR4 CDIMMs featured the following usage guidelines:

- ▶ Cannot be mixed with another size CDIMM on the same system node or server.
- ▶ 50% of the memory slots must be filled.
- ▶ Firmware 840.12 is the required minimum level.

These usage guidelines can change at any time.

The suggested approach is to install memory evenly across all processors and system nodes in the system. Balancing memory across the installed processors allows memory access in a consistent manner and often 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.

The location codes of the memory CDIMMs of a system node and their grouping as memory quads is shown in Figure 2-8.

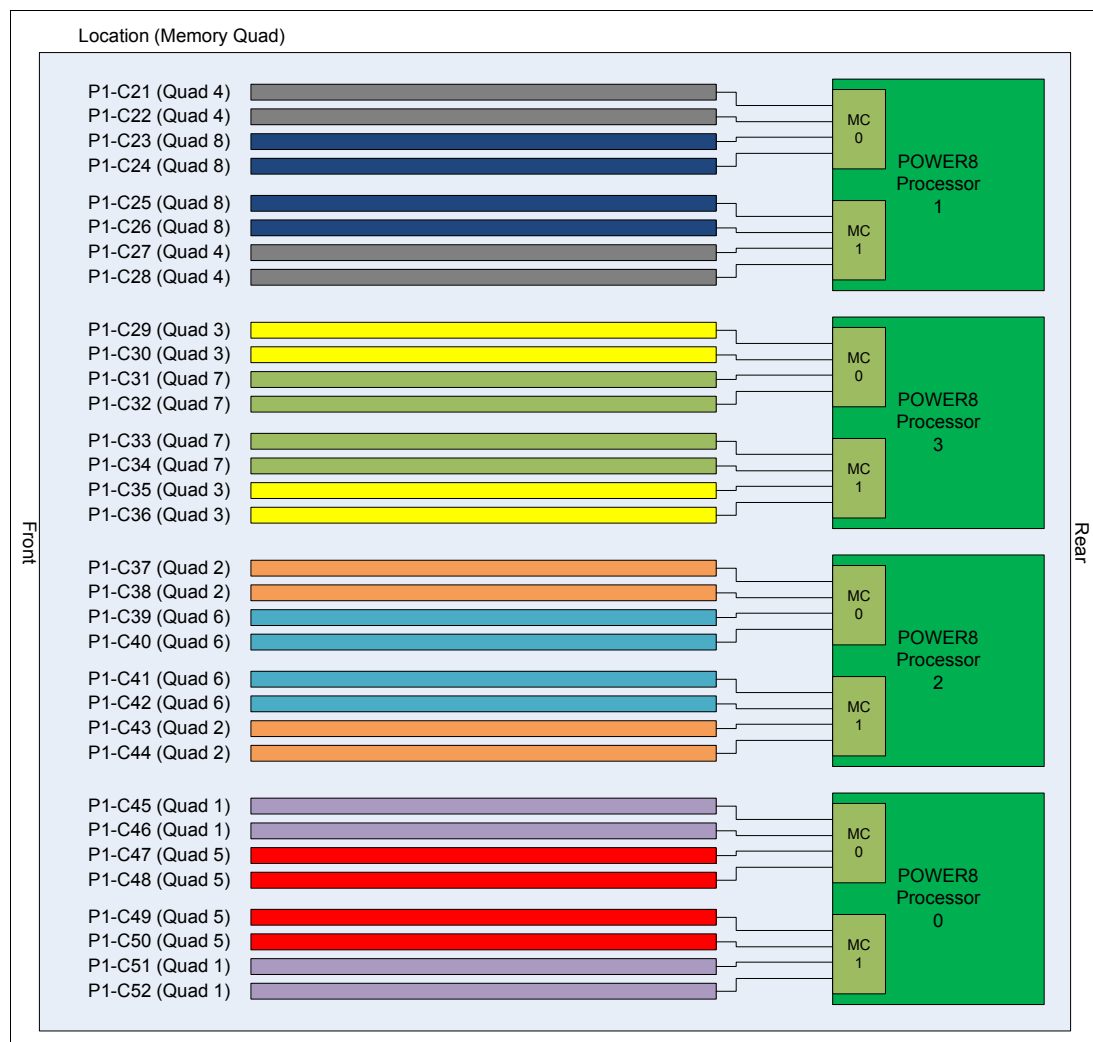


Figure 2-8 System node physical diagram with location codes for CDIMMs

Each system node has eight memory quads that are identified by the different colors that are used in Figure 2-8. The following location codes are used 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

The CDIMM plugging order for a Power E870C or Power E880C with a single system node is listed in Table 2-1.

Table 2-1 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: <ul style="list-style-type: none"> ▶ Memory quads 1 - 4 must be populated. ▶ Memory quads on the same processor must be populated with CDIMMs of the same capacity. 							

The CDIMM plugging order for a Power E870C or Power E880C with two system nodes is listed in Table 2-2.

Table 2-2 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: <ul style="list-style-type: none"> ▶ Memory quads 1 - 8 must be populated. ▶ Memory quads on the same processor must be populated with CDIMMs of the same capacity. 							

The CDIMM plugging order for a Power E880C with three system nodes is listed in Table 2-3.

Table 2-3 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: <ul style="list-style-type: none"> ▶ Memory quads 1 - 12 must be populated. ▶ Memory quads on the same processor must be populated with CDIMMs of the same capacity. 							

The CDIMM plugging order for a Power E880C with four system nodes is listed in Table 2-4.

Table 2-4 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: <ul style="list-style-type: none"> ► 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' capacity can be upgraded on demand and must have a minimum of 50% of their physical capacity activated. For example, the minimum physical installed memory for Power E870C and Power E880C is 256 GB, which requires a minimum of 128 GB activated.

The following activation types can be used to accomplish this activation:

- 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.

These types of memory activations can be in the same system if at least 25% of the memory activations are static. This configuration leads to a maximum of 75% of the memory activations as mobile.

An example of the minimum required activations for a system with 1 TB of installed memory is shown in Figure 2-9.

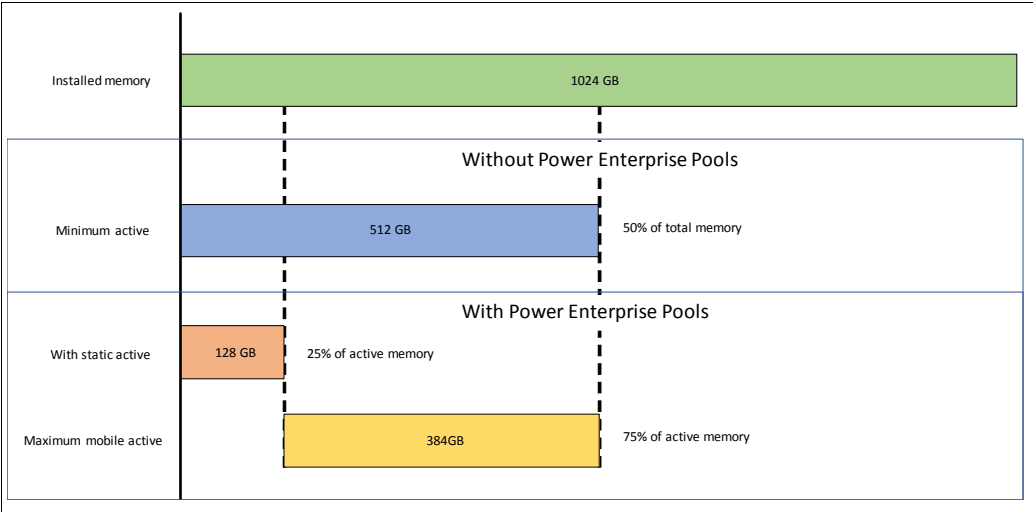


Figure 2-9 Example of the minimum required activations for a system with 1 TB of installed memory

The granularity for static memory activation is 1 GB; for mobile memory activation, the granularity is 100 GB. The feature codes that can be used to achieve the wanted number of activations are listed in Table 2-5.

Table 2-5 Static and mobile memory activation feature codes

Feature code	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. To enable mobile memory activations, the systems must be part of a power enterprise pool and include feature code #EB35 as configured. For more information about power enterprise pools, see 2.4.2, “Power enterprise pools and Mobile Capacity on Demand” on page 55.

2.3.4 Memory throughput

The peak memory and I/O bandwidths per system node increased over 300% compared to previous POWER7 processor-based servers. This increase provides the next generation of data-intensive applications with a platform that can handle the needed amount of data.

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

Power E870C

The maximum bandwidth estimates for a single core on the Power E870C system are listed in Table 2-6.

Table 2-6 Power E870C single core bandwidth estimates

Single core	Power E870C
	1 core @ 4.024 GHz
L1 (data) cache	193.15 GBps
L2 cache	193.15 GBps
L3 cache	257.54 GBps

The bandwidth figures for the caches are calculated by using the following formulas:

- 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 core clock. The following formula is used:

$$4.024 \text{ GHz Core: } (2 * 16 \text{ B} + 1 * 16 \text{ B}) * 4.024 \text{ GHz} = 193.15 \text{ GBps}$$

- 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 core clock. The following formula is used:

$$4.024 \text{ GHz Core: } (1 * 32 \text{ B} + 1 * 16 \text{ B}) * 4.024 \text{ GHz} = 193.15 \text{ GBps}$$

- 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 core clock. The following formula is used:

$$4.024 \text{ GHz Core: } (1 * 32 \text{ B} + 1 * 32 \text{ B}) * 4.024 \text{ GHz} = 257.54 \text{ GBps}$$

For each system node of a Power E870C that includes four processors and all its memory CDIMMs filled, the overall bandwidths are listed in Table 2-7.

Table 2-7 Power E870C system node bandwidth estimates

System node bandwidths	Power E870C
	32 cores @ 4.024 GHz
L1 (data) cache	6,181 GBps
L2 cache	6,181 GBps
L3 cache	8,241 GBps
Total Memory	922 GBps
PCIe Interconnect	252.064 GBps
Intra-node buses (two system nodes)	922 GBps

For PCIe Interconnect, each POWER8 processor has 32 PCIe lanes that are running at 7.877 Gbps full-duplex. The following bandwidth formula is used:

$$32 \text{ lanes} * 4 \text{ processors} * 7.877 \text{ Gbps} * 2 = 252.064 \text{ GBps}$$

Rounding: The bandwidths that are listed in this section might appear slightly differently in other materials because figures are rounded.

For the entire Power E870C system that includes two system nodes, the overall bandwidths are listed in Table 2-8.

Table 2-8 Power E870C total bandwidth estimates

Total bandwidths	Power E870C
	64 cores @ 4.024 GHz
L1 (data) cache	12,362 GBps
L2 cache	12,362 GBps
L3 cache	16,484 GBps
Total Memory	1,844 GBps
PCIe Interconnect	504.128 GBps
Inter-node buses (two system nodes)	307 GBps
Intra-node buses (two system nodes)	1,844 GBps

Power E880C

The maximum bandwidth estimates for a single core on the Power E880C system are listed in Table 2-9.

Table 2-9 Power E880C single core bandwidth estimates

Single core	Power E880C	Power E880C	Power E880C
	1 core @ 4.024 GHz	1 core @ 4.190 GHz	1 core @ 4.350 GHz
L1 (data) cache	193.15 GBps	201.12 GBps	208.80 GBps
L2 cache	193.15 GBps	201.12 GBps	208.80 GBps
L3 cache	257.54 GBps	268.19 GBps	278.40 GBps

The bandwidth figures for the caches are calculated by using the following formulas:

- 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 core clock. The following formulas are used:
 - 4.024 GHz Core: $(2 * 16 \text{ B} + 1 * 16 \text{ B}) * 4.024 \text{ GHz} = 193.15 \text{ GBps}$
 - 4.190 GHz Core: $(2 * 16 \text{ B} + 1 * 16 \text{ B}) * 4.190 \text{ GHz} = 201.12 \text{ GBps}$
 - 4.350 GHz Core: $(2 * 16 \text{ B} + 1 * 16 \text{ B}) * 4.350 \text{ GHz} = 208.80 \text{ GBps}$
- 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 core clock. The following formulas are used:
 - 4.024 GHz Core: $(1 * 32 \text{ B} + 1 * 16 \text{ B}) * 4.024 \text{ GHz} = 193.15 \text{ GBps}$
 - 4.190 GHz Core: $(1 * 32 \text{ B} + 1 * 16 \text{ B}) * 4.190 \text{ GHz} = 201.12 \text{ GBps}$
 - 4.350 GHz Core: $(1 * 32 \text{ B} + 1 * 16 \text{ B}) * 4.350 \text{ GHz} = 208.80 \text{ GBps}$
- 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 core clock. The following formulas are used:
 - 4.024 GHz Core: $(1 * 32 \text{ B} + 1 * 32 \text{ B}) * 4.024 \text{ GHz} = 257.54 \text{ GBps}$
 - 4.190 GHz Core: $(1 * 32 \text{ B} + 1 * 32 \text{ B}) * 4.190 \text{ GHz} = 268.19 \text{ GBps}$
 - 4.350 GHz Core: $(1 * 32 \text{ B} + 1 * 32 \text{ B}) * 4.350 \text{ GHz} = 278.40 \text{ GBps}$

For each system node of a Power E880C populated with four processors and all its memory CDIMMs filled, the overall bandwidths are listed in Table 2-10.

Table 2-10 Power E880C system node bandwidth estimates

System node bandwidths	Power E880C	Power E880C	Power E880C
	48 cores @ 4.024 GHz	40 cores @ 4.190 GHz	32 cores @ 4.350 GHz
L1 (data) cache	9,271 GBps	8,044.8 GBps	6,682 GBps
L2 cache	9,271 GBps	8,044.8 GBps	6,682 GBps
L3 cache	12,362 GBps	10,727.6 GBps	8,909 GBps
Total Memory	922 GBps	922 GBps	922 GBps
PCIe Interconnect	252.064 GBps	252.064 GBps	252.064 GBps
Intra-node buses (two system nodes)	922 GBps	922 GBps	922 GBps

For PCIe Interconnect, each POWER8 processor features 32 PCIe lanes that are running at 7.877 Gbps full-duplex. The bandwidth is calculated by using the following formula:

$$32 \text{ lanes} * 4 \text{ processors} * 7.877 \text{ Gbps} * 2 = 252.064 \text{ GBps}$$

Rounding: The bandwidths that are listed in this section might appear slightly differently in other materials because figures are rounded.

For the entire Power E880C system populated with four system nodes, the overall bandwidths are listed in Table 2-11.

Table 2-11 Power E880C total bandwidth estimates

Total bandwidths	Power E880C	Power E880C	Power E880C
	192 cores @ 4.024 GHz	160 cores @ 4.190 GHz	128 cores @ 4.350 GHz
L1 (data) cache	37,084 GBps	32,179.2 GBps	26,726 GBps
L2 cache	37,084 GBps	32,179.2 GBps	26,726 GBps
L3 cache	49,448 GBps	42,190.4 GBps	35,635 GBps
Total Memory	3,688 GBps	3,688 GBps	3,688 GBps
PCIe Interconnect	1008.256 GBps	1008.256 GBps	1008.256 GBps
Inter-node buses (four system nodes)	307 GBps	307 GBps	307 GBps
Intra-node buses (four system nodes)	3,688 GBps	3,688 GBps	3,688 GBps

2.3.5 Active Memory Mirroring

The Power E870C and Power E880C systems can 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 enables the system to continue to operate uninterrupted.

Active Memory Mirroring (AMM) is included with all Power E870C and Power E880C systems at no extra charge. It can be enabled, disabled, or reenabled depending on the user's requirements.

The Hypervisor code logical memory blocks are mirrored on distinct CDIMMs to allow for more usable memory. Because no specific CDIMM hosts the Hypervisor memory blocks, mirroring is done at the logical memory block level, not at the CDIMM level. To enable the AMM feature, the server must include enough free memory to accommodate the mirrored memory blocks.

In addition to the Hypervisor code, the following components are vital to the server operation are also mirrored:

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

Whether the Active Memory Mirroring option is enabled can be checked and its status can be changed through the Hardware Management Console (HMC) under the Advanced Tab on the CEC Properties panel as shown in Figure 2-10.

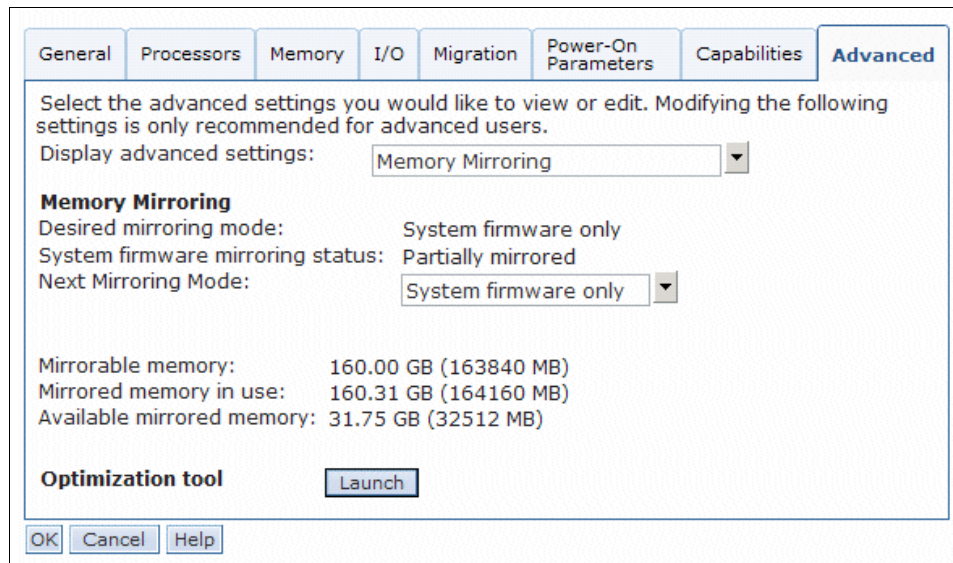


Figure 2-10 CEC Properties panel on an HMC

After one of the CDIMMs that contain Hypervisor data fails, all of the server operations remain active and flexible service processor (FSP) isolates the failing CDIMMs. Systems stay in the partially mirrored state until the failing CDIMM is replaced.

The following components 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 that is used to hold the contents of platform dumps

Partition data: Active Memory Mirroring does *not* mirror partition data. It was designed to mirror only the hypervisor code and its components, which allows 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 Special Uncorrectable Error (SUE) handling methods in the hardware, firmware, and operating system.

2.3.6 Memory Error Correction and Recovery

The memory 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 that are detected at the memory controller and the memory buffer.

Memory is also periodically scrubbed to correct soft errors and solid single-cell errors that are reported to the Hypervisor, which supports operating system deallocation of a page that is associated with a hard single-cell fault.

For more information about memory RAS, see 4.3.10, “Memory protection” on page 124.

2.3.7 Special Uncorrectable Error handling

SUE handling prevents an uncorrectable error in memory or cache from immediately causing the system to end. Instead, 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, stopping can be limited to the program, kernel, or Hypervisor owning the data, or freeze of the I/O adapters that are controlled by an I/O hub controller if data is to be transferred to an I/O device.

2.4 Capacity on Demand

Several types of Capacity on Demand (CoD) offerings are available on the Power 870 and Power E880C servers to help meet changing resource requirements in an on-demand environment. These offerings use resources that are installed on the system but are not activated.

2.4.1 Capacity Upgrade on Demand

Power E870C and Power E880C systems include several 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 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 are unavailable for use until you activate them. Inactive processor cores and memory units can be permanently activated by purchasing an activation feature that is named Capacity Upgrade on Demand (CUoD). The provided activation code is entered into your server.

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

CUoD can feature several applications to allow for a more flexible environment. One of its benefits is to allow a specific company to reduce the initial investment on a system. Traditional projects that use other technologies require that a system is acquired with all of the resources available to support the entire lifecycle of the project. This requirement might incur in costs that are necessary only in later stages of the project, usually with effects on software licensing costs and software maintenance.

By using CUoD, a company can start with a system with enough installed resources to support the entire project lifecycle but only with enough active resources that are necessary for the initial project phases. More resources can be added with the project, adjusting the hardware platform with the project needs. This addition allows for a company to reduce the initial investment in hardware and acquire only software licenses that are needed on each project phase, which reduces the Total Cost of Ownership and Total Cost of Acquisition of the solution. A comparison between two scenarios (a fully activated system versus a system with CUoD resources being activated along with the project timeline) is shown in Figure 2-11.

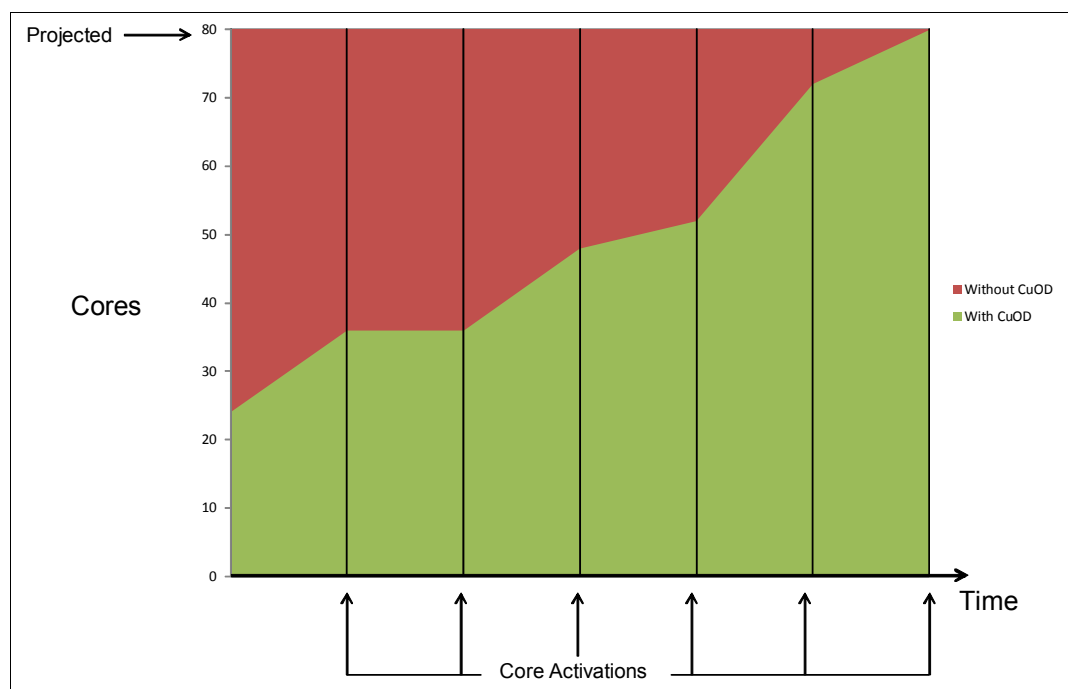


Figure 2-11 Active cores scenarios comparison during a project lifecycle

The static processor activation features that are available for the Power E870C and Power E880C are listed in Table 2-12.

Table 2-12 IBM Power Systems CUoD static processor activation features

System	Processor feature	Processor core static activation feature
Power E870C	#EPBA (4.02 GHz Processor Card)	EPBJ
Power E880C	#EPBB (4.35 GHz Processor Card)	EPBK
Power E880C	#EPBS (4.19 GHz Processor Card)	EPBU
Power E880C	#EPBD (4.02 GHz Processor Card)	EPBM

The static memory activation features that are available for the Power E870C and Power E880C are listed in Table 2-13.

Table 2-13 IBM Power Systems CUoD static memory activation features

System	Description	Feature code
Power E870C, Power E880C	Activation of 1 GB DDR3 POWER8 memory	EMA5
Power E870C, Power E880C	Activation of 100 GB DDR3 POWER8 memory	EMA6

2.4.2 Power enterprise pools and Mobile Capacity on Demand

Although static activations are valid for a single system, some customers can benefit from moving processor and memory activations among different servers because of 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. By using Mobile CoD activation codes, the systems administrator can perform tasks without contacting IBM.

The following types of power enterprise pools are available:

- ▶ Power 770 (9117-MMD), Power E870 (9119-MME), Power E870C (9080-MME), and Power E880C (9080-MHE) class systems
- ▶ Power 780 (9117-MHD), Power 795 (9119-FHB), Power E880 (9119-MHE), Power E870C (9080-MME), and Power E880C (9080-MHE) class systems

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

Mobile CoD features the following basic rules:

- ▶ The Power 770 and Power 780 systems require a minimum of four static processor activations.
- ▶ The Power 870, Power 870C, Power E880, and Power 880C 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 larger).
- ▶ For all systems, 25% of the active memory capacity must include static activations.

All of 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 systems that are not part of a power enterprise pool. Systems can belong to only one power enterprise pool at a time. Although powering down an HMC does not limit the assigned resources of participating systems in a pool, it 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 systems can be removed from the pool.
- ▶ New resources can be added to the pool or resources can be removed from the pool.
- ▶ Pool information can be viewed, including pool resource assignments, compliance, and history logs.

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

The mobile processor activation features that are available for the Power E870C and Power E880C are listed in Table 2-14.

Table 2-14 Mobile processor activation features

System	Description	CUoD mobile processor core activation feature
Power E870C	1-Core Mobile activation	#EP2S
Power E880C	1-Core Mobile activation	#EP2T

The mobile memory activation feature that is available for the Power E870C and Power E880C is listed in Table 2-15.

Table 2-15 Mobile memory activation features

System	Description	Feature code
Power E870C, Power E880C	100 GB Mobile memory activation	#EMA7

For more information about power enterprise pools, see *Power Enterprise Pools on IBM Power Systems*, REDP5101, which is available at this website:

<http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp5101.html>

2.4.3 Elastic Capacity on Demand

Note: Some websites 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 named On/Off CoD.

When you order an Elastic CoD feature, you receive an enablement code with which a system operator can make requests for more 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. Prepaid and post-pay options are available.

Charges are based on usage reporting that is collected monthly. Processors and memory can be activated and turned off an unlimited number of times when more 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 is on the server records the usage activity. This usage data must be sent to IBM monthly. A bill is then generated based on the total amount of processor and memory resources that is used in increments of processor and memory (1 GB) days.

Power E870C and Power E880C support the 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 temporary capacity is used on your server, you must enable your server. To enable the server, an enablement feature (MES only) must be ordered and the required contracts must be in place.

If a Power E870C or Power E880C server uses IBM i and 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 Elastic CoD process consists of the following steps:

- Enablement

Before requesting temporary capacity on a server, you must enable it for Elastic CoD. To perform this enablement, order an enablement feature and sign the required contracts. IBM generates an enablement code, mails it to you, and posts it on the Internet 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, order another processor enablement code to reset the number of days that you can request back to 360.

A *memory enablement* code allows you to request up to 999 memory days of temporary capacity. If you reach the limit of 999 memory days, order 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 that is signed by the client before the enablement code is received 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, see this website:

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

2.4.4 Utility Capacity on Demand

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

By using 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, a Utility CoD processor minute is charged and this level of performance is available for the next minute of use.

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

Registration and usage reporting for utility CoD is made by 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 E870C or Power E880C server uses the IBM i operating system and 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 about registration, enablement, and use of Utility CoD, see this website:

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

2.4.5 Trial Capacity on Demand

A *standard request* for Trial CoD requires you to complete a form that includes contact information and vital product data (VPD) from your Power E870C or Power E880C 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 E870C or Power E880C system with inactive CoD resources. An exception request activates 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.

For more information about requesting a Standard or an Exception Trial, see this website:

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

2.4.6 Software licensing and CoD

For more information about software licensing considerations with the various CoD offerings, see the most recent revision of the *Power Systems Capacity on Demand User's Guide*, which is available at this website:

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

2.5 System bus

This section provides more information about internal buses.

2.5.1 PCI Express Gen3

The internal I/O subsystem on the Power E870C and Power E880C 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) that are running at 7.877 Gbps full-duplex and provides 31.508 GBps of I/O connectivity to the PCIe slots.

The connections are shown in Figure 2-12.

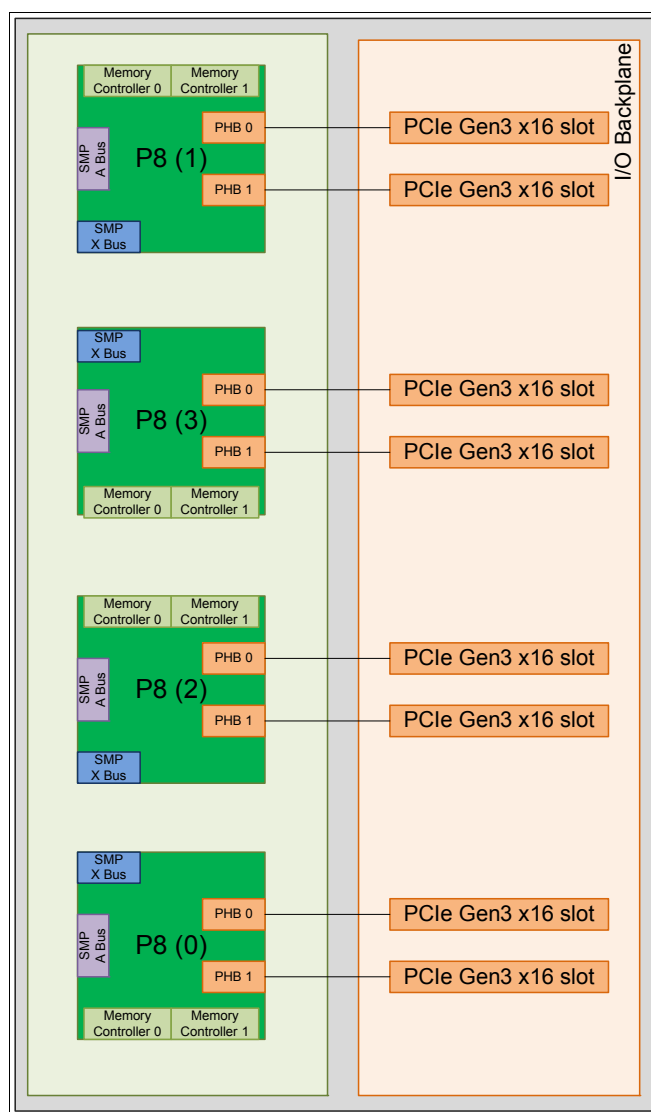


Figure 2-12 System nodes PCIe slots directly attached to PCIe controllers on POWER8 chips

In addition to the slots that are directly attached to the processors PCI Gen3 controllers, the systems allow for more PCIe adapters on external PCIe Expansion Drawers and disks on external drawers that are connected through PCIe SAS adapters.

Figure 2-13 on page 61 shows the I/O connectivity options that are available for the Power E870C and Power E880C. The system nodes allow for eight PCIe Gen3 x16 slots. Slots can be added by attaching PCIe Expansion Drawers and SAS disks can be attached to EXP24S SFF Gen2 Drawers. The EXP24S can be attached to SAS adapters on the system nodes or the PCIe Expansion Drawer.

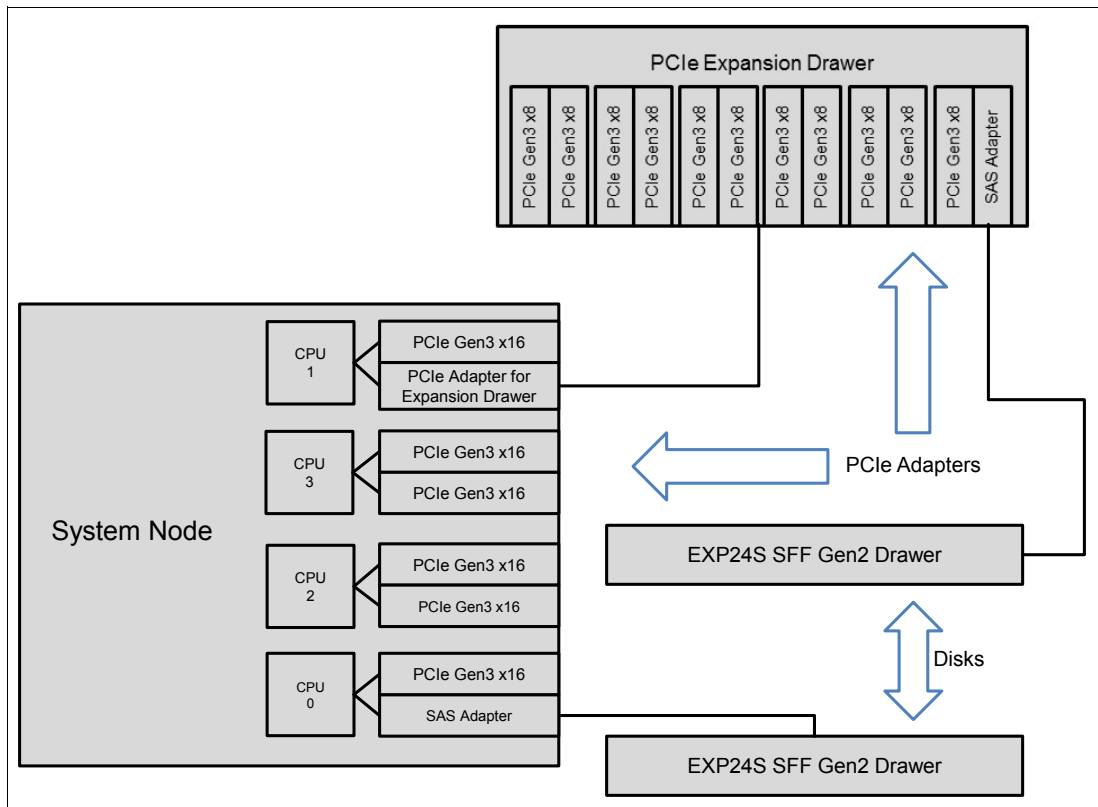


Figure 2-13 I/O connectivity options available for Power E870C and Power E880C

For more information about adapters and their supported slots, see 2.7, “PCI adapters” on page 65.

Disk support: Support is not available for disks that are directly installed on the system nodes and PCIe Expansion Drawers. If directly attached SAS disk are required, they must 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 74.

2.5.2 Service Processor Bus

The redundant service processor bus connectors are 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 through these cables.

Unlike the previous generations in which a specific pair of enclosures hosts the service processors, redundant service processor cards are installed on the control unit and redundant clock cards on Power E870C and Power E880C as a standard.

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

The cables that are necessary for each system node are grouped under a single feature code, which allows 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.

The available feature codes are listed in Table 2-16.

Table 2-16 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

The following cable set feature codes are incremental and depend on the number of installed drawers:

- ▶ 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

For more information about system connection topology, see 2.1, “Logical diagrams” on page 36.

2.6 Internal I/O subsystem

The internal I/O subsystem is on the I/O planar, which supports eight PCIe Gen3 x16 slots. All PCIe slots are hot-pluggable and are 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 125. The slot configuration of Power E870C and Power E880C system nodes is listed in Table 2-17.

Table 2-17 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 this writing, there are no supported CAPI adapters for the Power E870C and E880C system node.

The physical locations of the slots are shown in Figure 2-14.

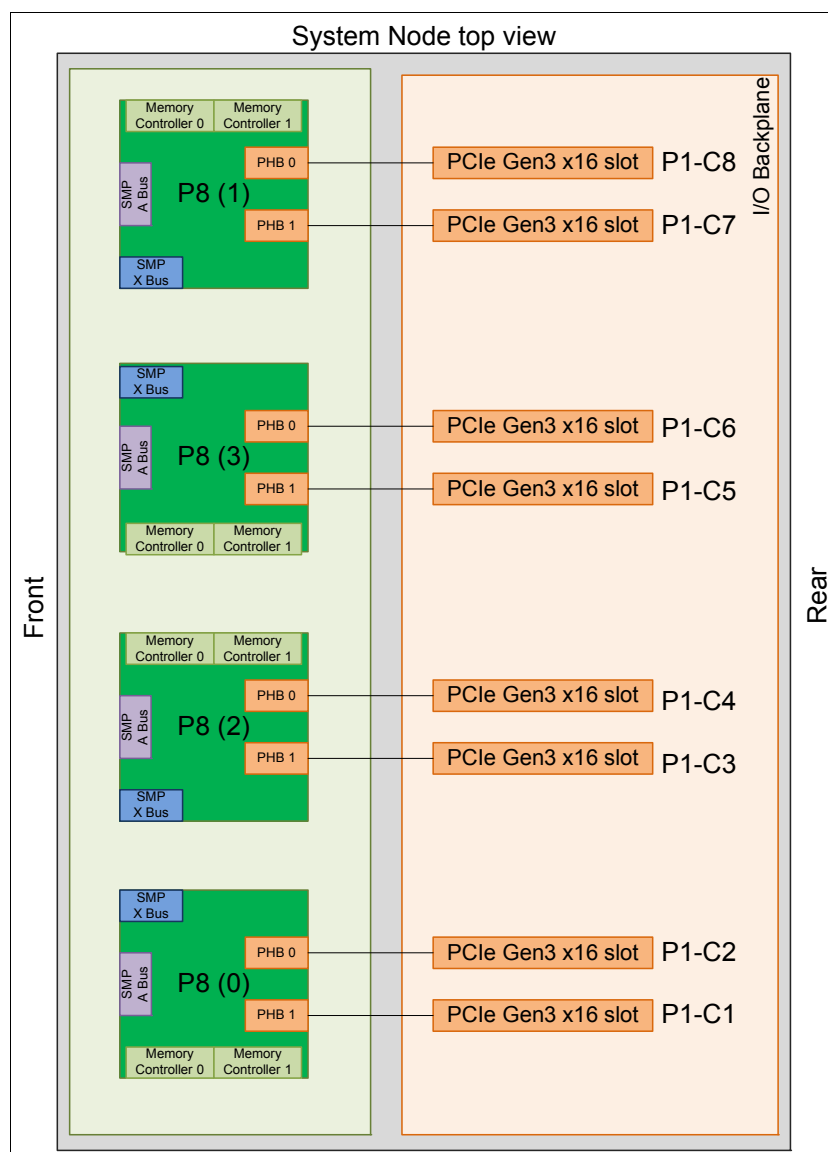


Figure 2-14 System node top view and PCIe slot location codes

2.6.1 Blind-swap cassettes

The Power E870C and Power E880C 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, which facilitates more airflow through the cassette and 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 that are installed in standard PCIe slots.

The system control unit has one USB port that is dedicated to the DVD drive and 4 Ethernet ports that are used for HMC communications. Because there is no serial port, an HMC is mandatory for system management. The FSP's virtual console is on the HMC.

The location of the USB and HMC Ethernet ports is shown in Figure 2-15.

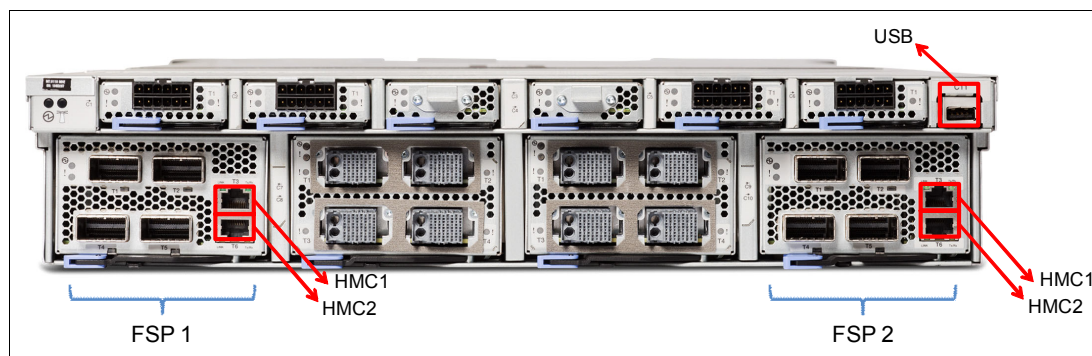


Figure 2-15 Physical location of the USB and HMC ports on the system control unit

For more information about connection and use of the DVD, see 2.8.1, “Media features” on page 73.

2.7 PCI adapters

This section describes the types and functions of the PCI cards that are supported by Power E870C and Power E880C systems.

2.7.1 PCI Express

PCI Express (PCIe) uses a serial interface and allows for point-to-point interconnections between devices (by 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 the number of lanes.

The PCIe interfaces that are supported on this server are PCIe Gen3, which are 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. However, larger speed adapters are not compatible in smaller connector sizes (for example, a x16 adapter cannot go in an x8 PCIe slot connector).

IBM POWER8 processor-based servers can support the following 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, which are used in the PCIe Gen3 I/O expansion drawer (#EMX0).

Low-profile PCIe adapters are supported only in low-profile PCIe slots. 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 this 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 any update prerequisites are available to install. To make this determination, see the following 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. Operating system support (AIX, IBM i, and Linux) for each of the adapters is listed in the tables.

2.7.2 LAN adapters

To connect the Power E870C and Power E880C servers to a local area network (LAN), you can use the LAN adapters that are supported in the PCIe slots of the system. The available LAN adapters are listed in Table 2-18. (Information about FCoE adapters is listed in Table 2-22 on page 70.)

Table 2-18 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
EC27	EC27	PCIe2 LP 2-Port 10 GbE RoCE SFP+ Adapter	CEC	AIX, Linux

Feature Code	CCIN	Description	Placement	OS support
EC28	EC27	PCIe2 2-Port 10 GbE RoCE SFP+ Adapter	I/O drawer	AIX, IBM i, Linux
EC29	EC29	PCIe2 LP 2-Port 10 GbE RoCE SR Adapter	CEC	AIX, IBM i, Linux
EC2G		PCIe2 LP 2-port 10 GbE SFN6122F Adapter	CEC	Linux
EC2J		PCIe2 2-port 10 GbE SFN6122F Adapter	I/O drawer	Linux
EC2M	57BE	PCIe3 LP 2-port 10 GbE NIC&RoCE SR Adapter	CEC	AIX, IBM i, Linux
EC2N		PCIe3 2-port 10 GbE NIC&RoCE SR Adapter	I/O drawer	AIX, IBM i, Linux
EC30	EC29	PCIe2 2-Port 10 GbE 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
EN0M	2CC0	PCIe2 4-port(10 Gb FCoE and 1 GbE) LR&RJ45 Adapter	I/O drawer	AIX, IBM i, Linux
EN0N	2CC0	PCIe2 LP 4-port(10 Gb FCoE and 1 GbE) LR&RJ45 Adapter	CEC	AIX, IBM i, Linux
EN0S	2CC3	PCIe2 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter	I/O drawer	AIX, IBM i, Linux
EN0T	2CC3	PCIe2 LP 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter	CEC	AIX, IBM i, Linux
EN0U	2CC3	PCIe2 4-port (10 Gb+1 GbE) Copper SFP+RJ45 Adapter	I/O drawer	AIX, IBM i, Linux
EN0V	2CC3	PCIe2 LP 4-port (10 Gb+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

Feature Code	CCIN	Description	Placement	OS support
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

The available graphics accelerator adapters are listed in Table 2-19. An adapter can be configured to operate in 8-bit or 24-bit color modes. The adapter supports analog and digital monitors.

Table 2-19 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

The SAS adapters that are available for Power E870C and Power E880C systems are listed in Table 2-20.

Table 2-20 Available SAS adapters

Feature Code	CCIN	Description	Placement	OS support
5901	57B3	PCIe Dual-x4 SAS Adapter	I/O drawer	AIX, Linux
EJ0J	57B4	PCIe3 RAID SAS Adapter Quad-port 6 Gb x8	I/O drawer	AIX, Linux
EJ0L	57CE	PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6 Gb x8	I/O drawer	AIX, Linux
EJ0M		PCIe3 LP RAID SAS ADAPTER	CEC	AIX, Linux
EJ10	57B4	PCIe3 SAS Tape/DVD Adapter Quad-port 6 Gb x8	I/O drawer	AIX, Linux
EJ11	57B4	PCIe3 LP SAS Tape/DVD Adapter Quad-port 6 Gb x8	CEC	AIX, Linux
EJ14	57B1	PCIe3 12 GB Cache RAID PLUS SAS Adapter Quad-port 6 Gb x8	I/O drawer	AIX, IBM i, Linux
EJ1P	57B3	SAS Tape/DVD Dual-port 3Gb x8 Adapter	CEC	AIX, IBM i, Linux

2.7.5 Fibre Channel adapter

The systems support direct or SAN connection to devices that use Fibre Channel adapters. The available Fibre Channel adapters, which all have LC connectors, are listed in Table 2-21.

Table 2-21 Available Fibre Channel adapters

Feature Code	CCIN	Description	Placement	OS support
5273	577D	PCIe LP 8 Gb 2-Port Fibre Channel Adapter	CEC	AIX, IBM i, Linux
5276	5774	PCIe LP 4 Gb 2-Port Fibre Channel Adapter	CEC	AIX, IBM i, Linux
5729	5729	PCIe2 8 Gb 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 16 Gb 2-port Fibre Channel Adapter	I/O drawer	AIX, Linux
EN0B	577F	PCIe2 LP 16 Gb 2-port Fibre Channel Adapter	CEC	AIX, IBM i, Linux
EN0F	578D	PCIe2 LP 8 Gb 2-Port Fibre Channel Adapter	CEC	AIX, Linux
EN0G		PCIe2 8 Gb 2-Port Fibre Channel Adapter	I/O drawer	AIX, Linux
EN0Y	EN0Y	PCIe2 LP 8 Gb 4-port Fibre Channel Adapter	CEC	AIX, IBM i, Linux
EN12		PCIe2 8 Gb 4-port Fibre Channel Adapter	I/O drawer	AIX, Linux

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

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.

Fibre Channel and network connections and FCoE connections are compared in Figure 2-16.

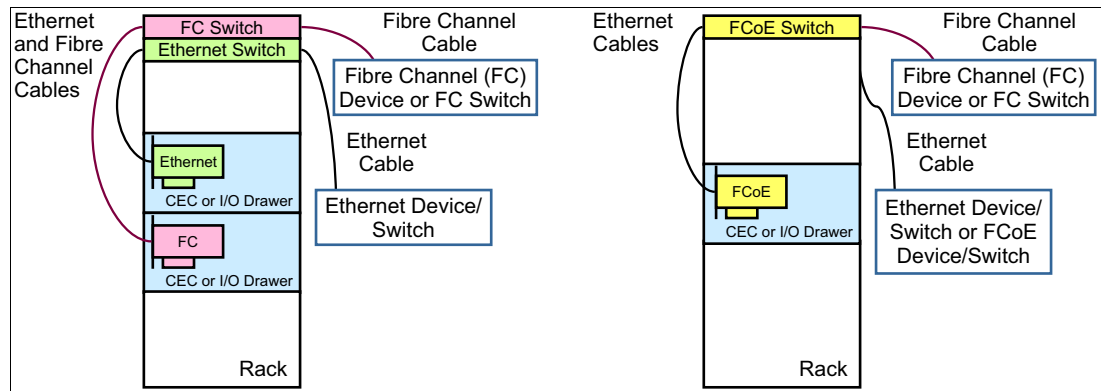


Figure 2-16 Comparing Fibre Channel and network connections and FCoE connections

The available FCoE adapters are listed in Table 2-22. The adapters are high-performance Converged Network Adapters (CNAs) that use SR optics. Each port can simultaneously provide network interface card (NIC) traffic and Fibre Channel functions.

Table 2-22 Available FCoE adapters

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

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

Note: Adapters EN0J and EN0L support SR-IOV when minimum firmware and software levels are met.

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 that is in the system node or a PCIe Gen3 I/O drawer.

The available USB adapters are listed in Table 2-23.

Table 2-23 Available asynchronous and USB adapters

Feature Code	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 that are necessary for present and future server systems with levels significantly better than can be achieved by using bus-oriented I/O structures.

InfiniBand is an open set of interconnect standards and specifications. The main InfiniBand specification is published by the InfiniBand Trade Association and is available at this website:

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

InfiniBand is based on a switched fabric architecture of serial point-to-point links. These InfiniBand links can be connected to host channel adapters (HCAs), which are used primarily in servers, or target channel adapters (TCAs), which are 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 of 2.5 Gbps - 120 Gbps.

The architecture defines a layered hardware protocol and a software layer to manage the initialization and communication between devices. Each link can support multiple transport services for reliability and multiple prioritized virtual communication channels.

The available InfiniBand adapters are listed in Table 2-24.

Table 2-24 Available InfiniBand adapters

Feature Code	Description	Placement	OS support
EC3E	PCIe3 LP 2-port 100 Gb EDR InfiniBand Adapter x16	CEC	Linux
EC3T	PCIe3 LP 1-port 100 Gb EDR InfiniBand Adapter x16	I/O drawer	Linux

2.7.9 Cryptographic Coprocessor

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

The IBM PCIe Cryptographic Coprocessor adapter includes 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 that use Europay, MasterCard, and Visa
- ▶ Cryptographic key generation and random number generation
- ▶ Generation, verification, and translation PIN processing
- ▶ Encrypt and decrypt by using AES and DES keys

For more information about the latest firmware and software updates, see this website:

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

The cryptographic adapter that is available for the server is listed in Table 2-25.

Table 2-25 Available cryptographic adapters

Feature Code	Description	Placement	OS support
EJ23	PCIe3 Crypto Coprocessor BSC-Gen3 4767	I/O drawer	AIX, IBM i, Linux

2.7.10 CAPI adapters

The CAPI adapter for JAVA and IBM WebSphere® acceleration is listed in Table 2-26.

Table 2-26 Available CAPI adapters

Feature code	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 E870C and Power E880C do not allow for conventional physical storage. All storage must be provided externally by using I/O expansion drawers or SAN. At the time of this writing, the only external I/O expansion drawer that can be used in Power E870C and Power E880C is the EXP24S, which is attached by using SAS ports to a SAS PCIe adapter that is installed in a system node drawer or in a PCIe expansion slot that is in an I/O expansion drawer.

The system control unit includes a DVD drive, which is connected to an external USB port on the rear of the unit. To use the DVD drive, at least one PCIe USB adapter must be installed and connected by using a 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 Media features

One DVD media bay is available per system, which is 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 by using a USB cable to a USB PCIe adapter, which is installed in one of the available PCIe slots on the system nodes or I/O expansion drawers. The DVD connection is shown in Figure 2-17.

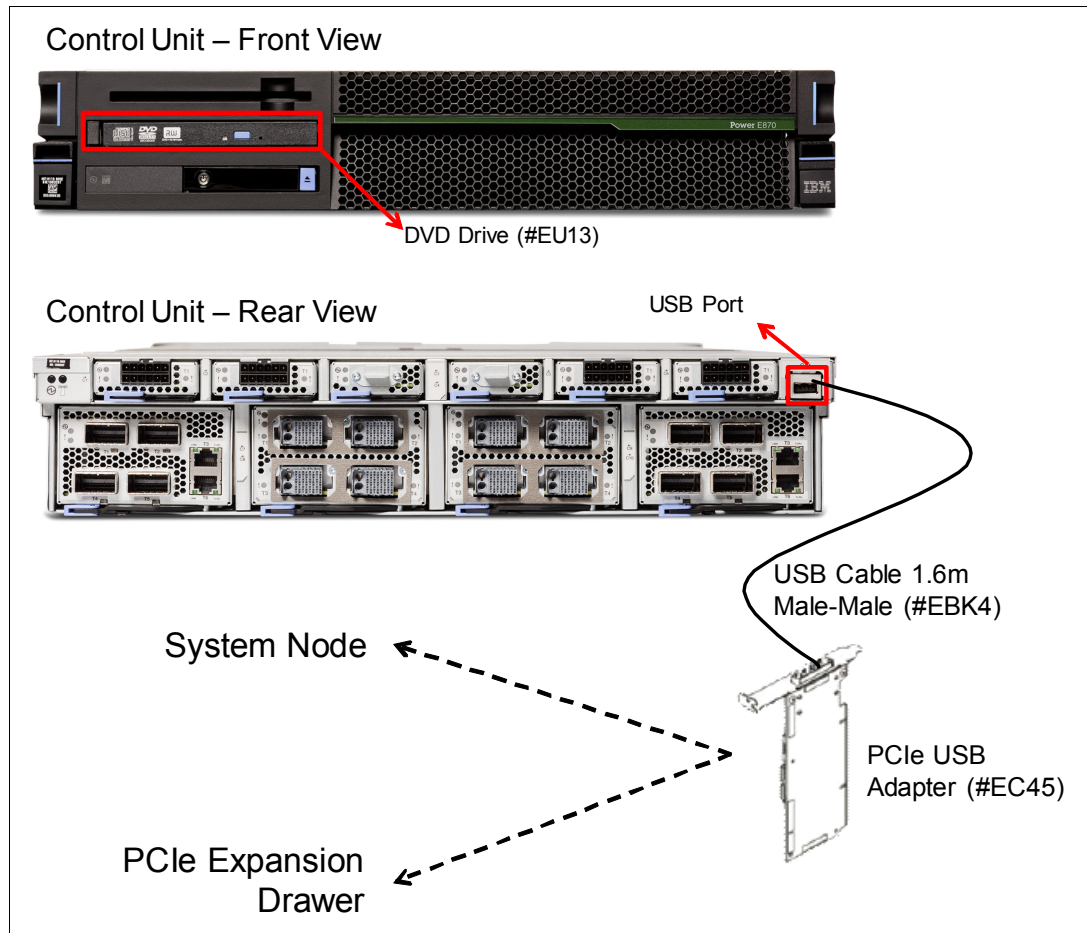


Figure 2-17 DVD drive physical location on control unit and suggested cabling

The following basic rules for DVD drives apply for these systems:

- ▶ 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 independent of another components and can be freely moved between partition.

Note: Several daily functions in which a DVD drive was needed can now be run by using other methods, such as Virtual Media Repository on the Virtual I/O Server or the Remote Virtual I/O Server installation on HMC.

The Power E870C and Power E880CL support the RDX USB External Docking Station for Removable Disk Cartridge (#EUA4). The USB External Docking Station accommodates RDX removable disk cartridge of any capacity. The disks are in a protective rugged cartridge enclosure that plug into the docking station. The docking station holds one removable rugged disk drive/cartridge at a time. The rugged removable disk cartridge and docking station backs up similar to tape drive. This can be an excellent alternative to DAT72, DAT160, 8 mm, and VXA-2 and VXA-320 tapes.

Table 2-27 shows the available media device feature codes for the Power E870C and Power E880C servers.

Table 2-27 Media device feature code descriptions for Power E870C and Power E880C

Feature code	Description
EC13	SATA Slimline DVD-RAM Drive
EUA4	RDX USB External Docking Station for Removable Disk Cartridge

2.9 External I/O subsystems

This section describes the PCIe Gen3 I/O expansion drawer that can be attached to the Power E870C and Power E880C.

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 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 includes two AOC cables.

Concurrent repair and adding or removing PCIe adapters are 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 are installed 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.

The back view of the PCIe Gen3 I/O expansion drawer is shown in Figure 2-18.

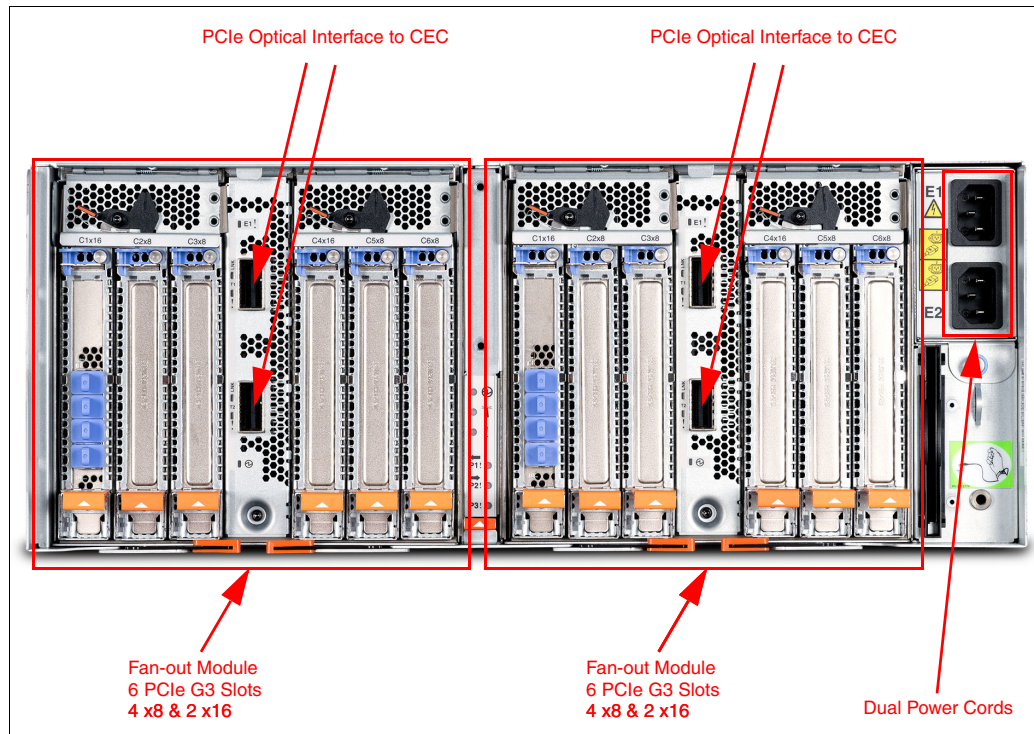


Figure 2-18 Rear view of the PCIe Gen3 I/O expansion drawer

2.9.2 PCIe Gen3 I/O expansion drawer optical cabling

I/O drawers are connected to the adapters in the system node by using the following 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 fanout module must be cabled to the top port of the #EJ07 port. Likewise, the bottom two ports must be cabled together. Complete the following steps:

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

The connector locations for the PCIe Gen3 I/O expansion drawer are shown in Figure 2-19.

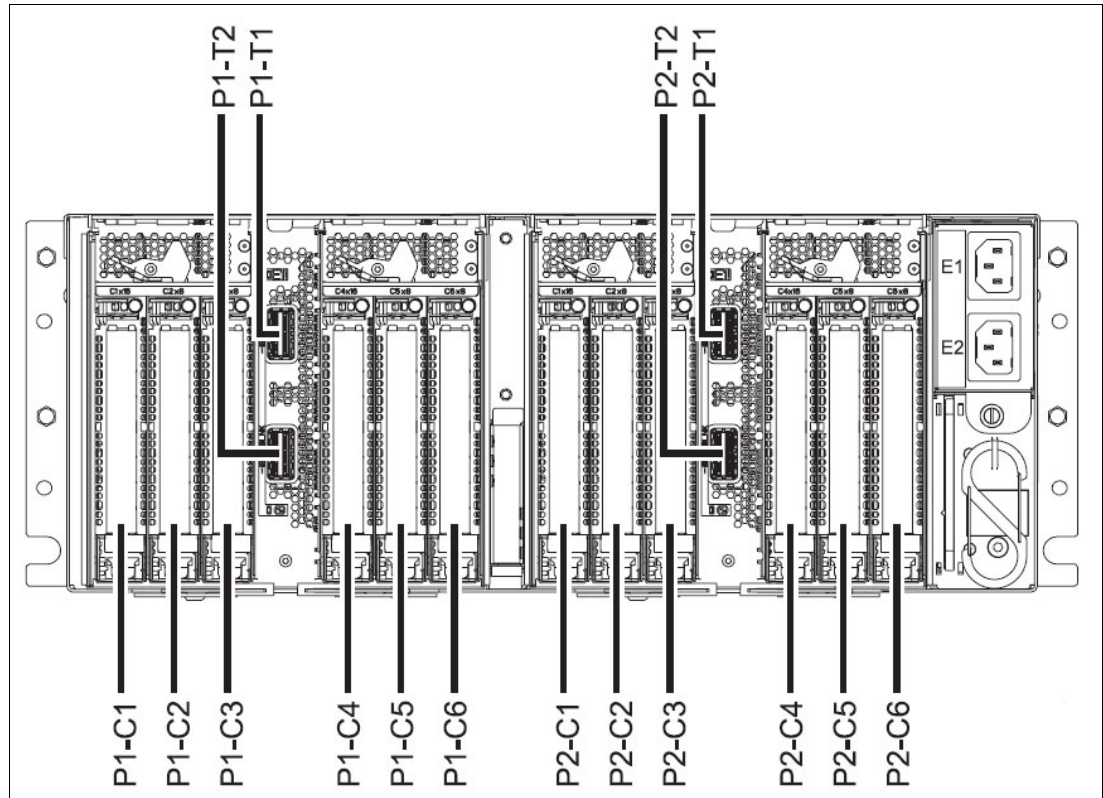


Figure 2-19 Connector locations for the PCIe Gen3 I/O expansion drawer

The typical optical cable connections are shown in Figure 2-20.

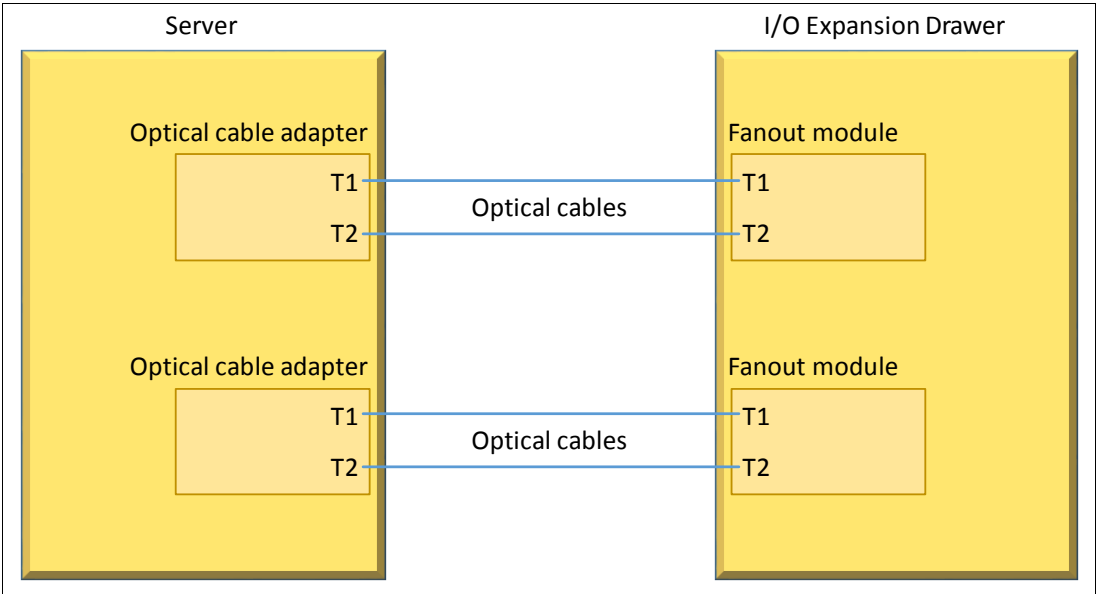


Figure 2-20 Typical optical cable connection

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 E870C and Power E880C system node. However, we advise that you use the PCIe adapter slot priority information while selecting slots for installing PCIe3 Optical Cable Adapter (#EJ07).

The PCIe adapter slot priorities in the Power E870C and Power E880C system are shown in Table 2-28.

Table 2-28 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

Examples of supported configurations are shown in Figure 2-21, Figure 2-22 on page 79, and Figure 2-23 on page 80. For simplification, not every possible combination of the I/O expansion drawer to server attachments is shown.

An example of a single-system node and two PCI Gen3 I/O expansion drawers is shown in Figure 2-21.

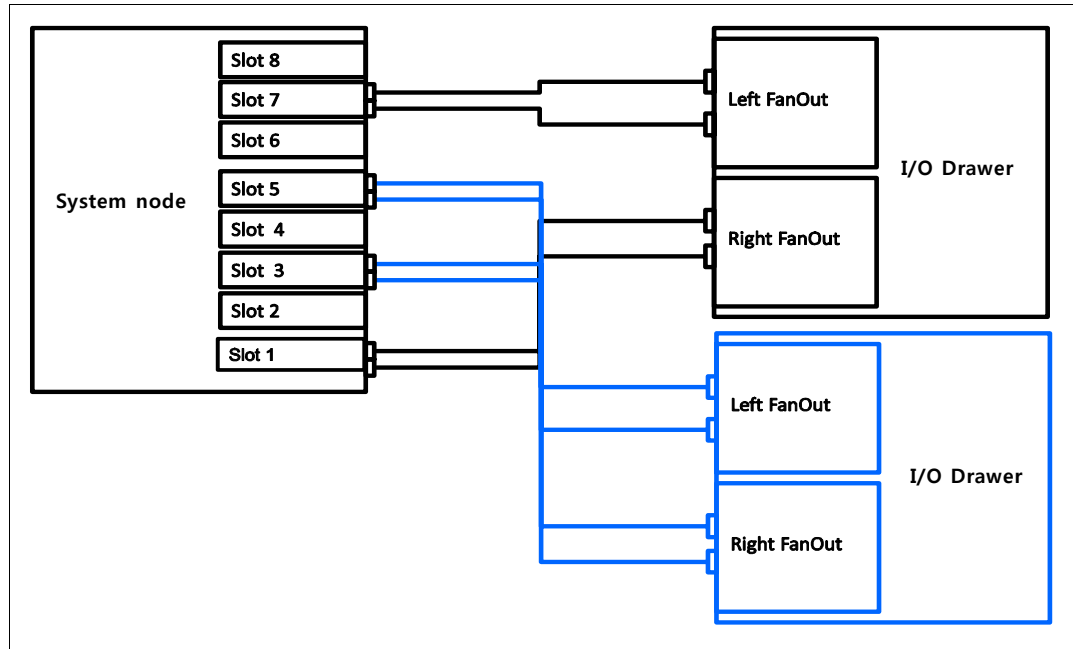


Figure 2-21 Example of a single system node and two I/O drawers

An example of two system nodes and two PCI Gen3 I/O expansion drawers is shown in Figure 2-22.

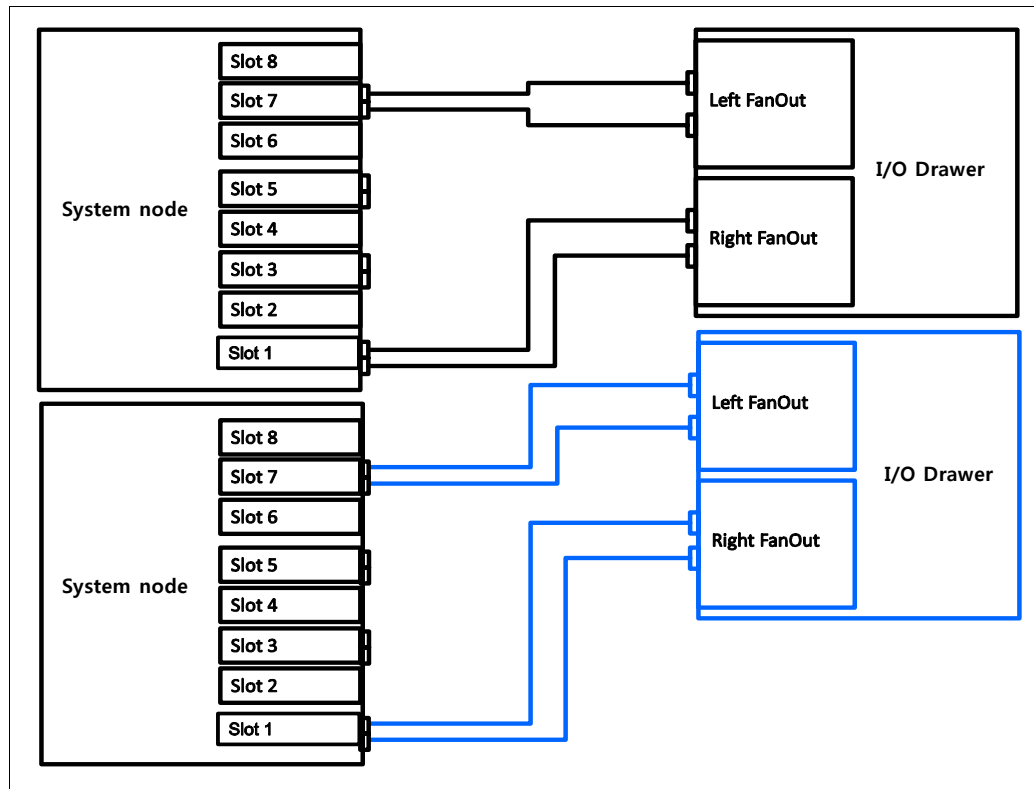


Figure 2-22 Example of two system nodes and two I/O drawers

An example of two system nodes and four PCI Gen3 I/O expansion drawers is shown in Figure 2-23.

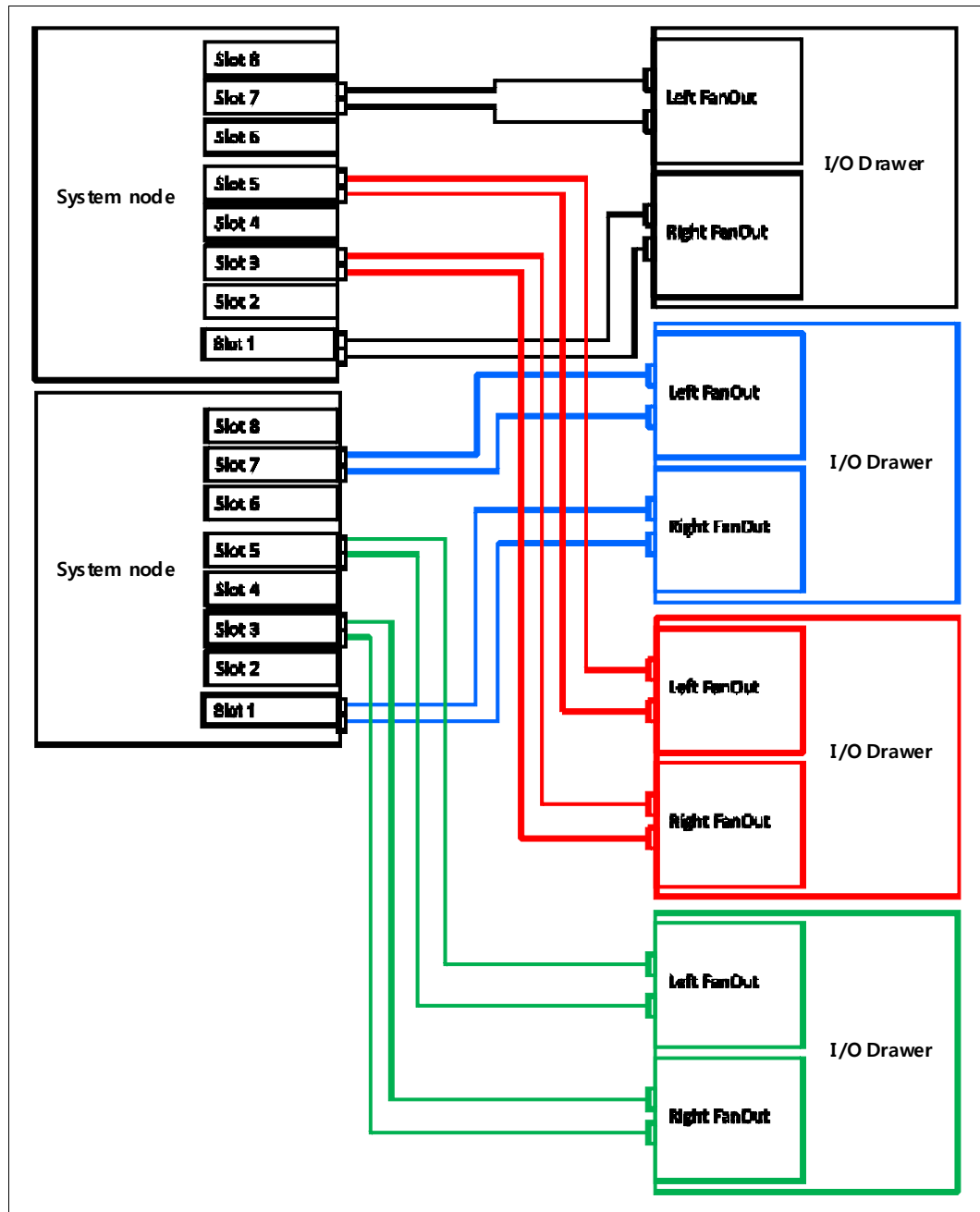


Figure 2-23 Example of two system nodes and four I/O drawers

2.9.3 PCIe Gen3 I/O expansion drawer SPCN cabling

A system power control network (SPCN) is *not* used to control and monitor the status of power and cooling within the I/O drawer. Instead, 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 E870C and Power E880C 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 E870C and Power E880C server.

2.10.1 EXP24S SFF Gen2-bay Drawer

The EXP24S SFF Gen2-bay Drawer (#5887) is an expansion drawer with 24 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 use, the system node of Power E870C and Power E880C system does not include 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 that are consistent with previously used IBM Power Systems. IBM i configurations require the drives to be protected (RAID or mirroring). Although protecting the drives is highly advised, it is 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). The front of the unit and the groups of disks on each mode is shown in Figure 2-24.

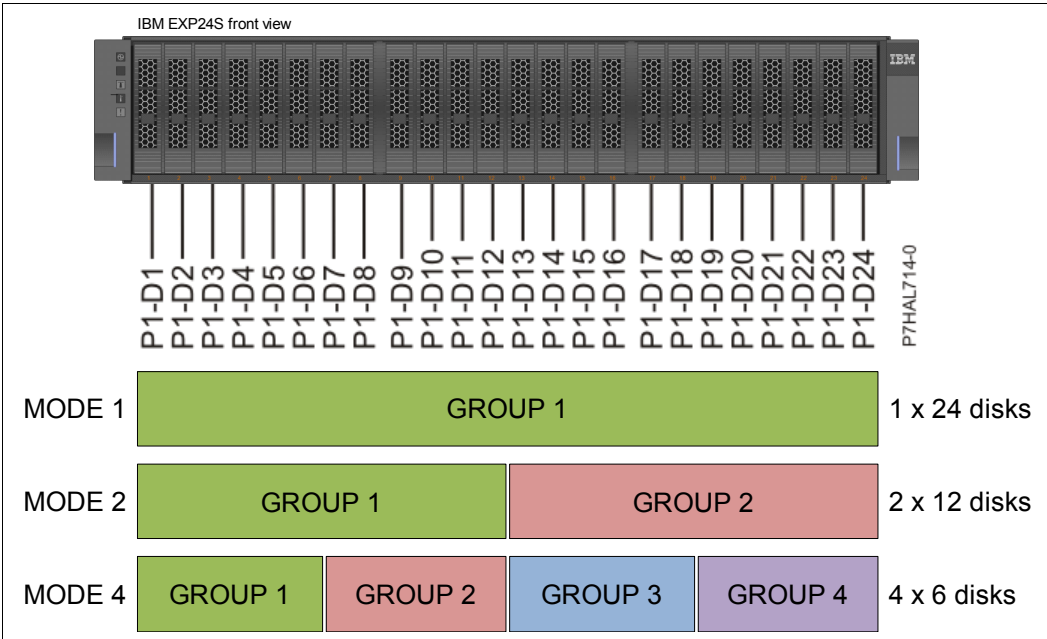


Figure 2-24 EXP24S front view with location codes and disk groups

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

<http://w3.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5121>

As shown in Figure 2-25 on page 83, 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).

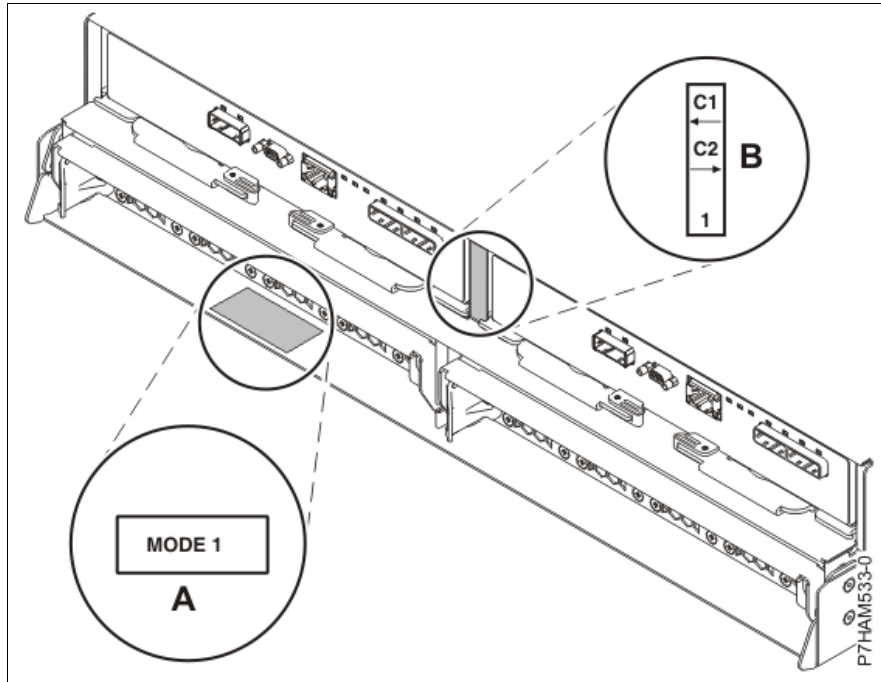


Figure 2-25 Mode sticker locations at the rear of the 5887 disk drive enclosure

The EXP24S SAS ports are attached to a SAS PCIe adapter or pair of adapters by using SAS YO or X cables. Cable length varies depending on the feature code and proper length is calculated while considering routing for proper airflow and ease of handling. Both types of SAS cables are shown in Figure 2-26.

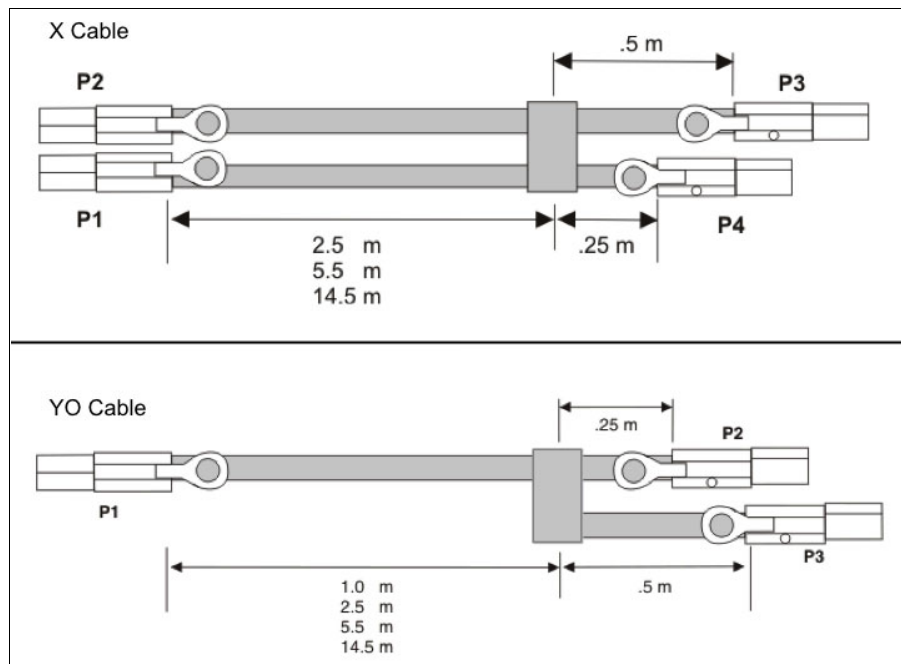


Figure 2-26 Diagram of SAS cables types X and YO

The following SAS adapters support the EXP24S:

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

The EXP24S drawer can support up to 24 SAS SFF Gen-2 disks. The available disk options are listed in Table 2-29.

Table 2-29 Available disks for the EXP24S

Feature Code	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
1738	856 GB 10 K RPM SAS SFF-2 Disk Drive (IBM i)	IBM i
1752	900 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX, Linux

Feature Code	Description	OS support
1948	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	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	1.1 TB 10 K RPM SAS SFF-2 Disk Drive (IBM i)	IBM i
ESD3	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	571 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224	IBM i
ESEV	600 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX, Linux
ESEY	283 GB 15 K RPM SAS SFF-2 4 K Block - 4224 Disk Drive	IBM i
ESEZ	300 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	AIX, Linux
ESF2	1.1 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224	IBM i
ESF3	1.2 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX, Linux
ESFN	571 GB 15 K RPM SAS SFF-2 4 K Block - 4224 Disk Drive	IBM i
ESFP	600 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	AIX, Linux
ESFS	1.7 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224	IBM i
ESFT	1.8 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX, Linux

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

- 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.

The rear connectors of the EXP24S drawer, how they relate with the modes of operation, and disk grouping are shown in Figure 2-27.

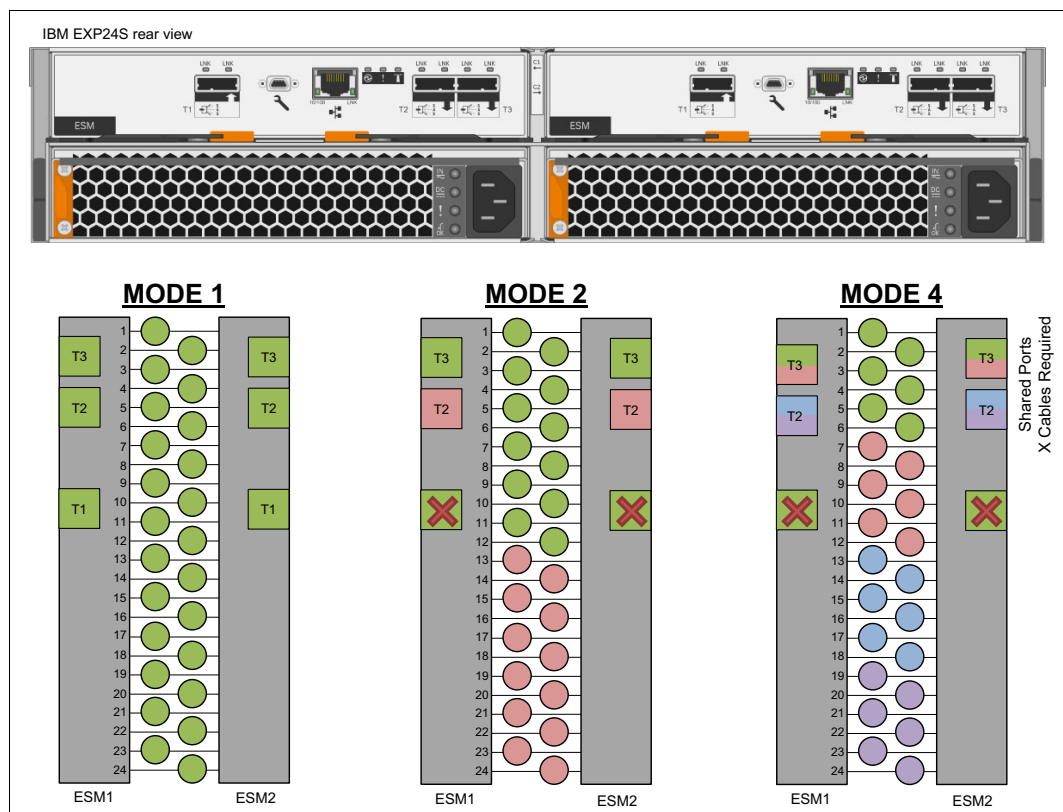


Figure 2-27 Rear view of EXP24S with the 3 modes of operation and the disks assigned to each port

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 by any of the supported SAS adapters or controllers.

For more information about the most common configurations for EXP24S with IBM Power Systems, see 2.10.2, “EXP24S common usage scenarios” on page 86. (Not all possible scenarios are included.)

For more information about SAS cabling and cabling configurations, search for “Planning for serial-attached SCSI cables” in the following IBM Knowledge Center website:

http://www.ibm.com/support/knowledgecenter/TI0003M/p8had/p8had_sascabing.htm

2.10.2 EXP24S common usage scenarios

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

Note: Not all possible scenarios are included. For more information about supported scenarios, see the *Planning for serial-attached SCSI cables* guide in the IBM Knowledge Center.

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, which allows for up to 24 disks to be attached to the server. The connection diagram and components of the solution are shown in Figure 2-28.

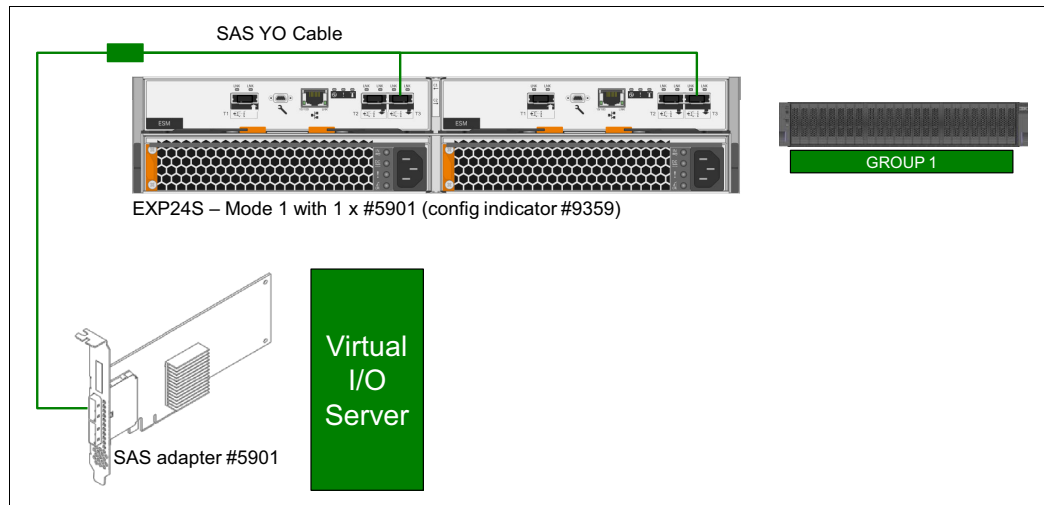


Figure 2-28 Scenario 1: Basic non-redundant connection

The following feature codes are required for this scenario:

- ▶ 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, which allows for up to 24 disks to be attached to the server. The connection diagram and components of the solution are shown in Figure 2-29.

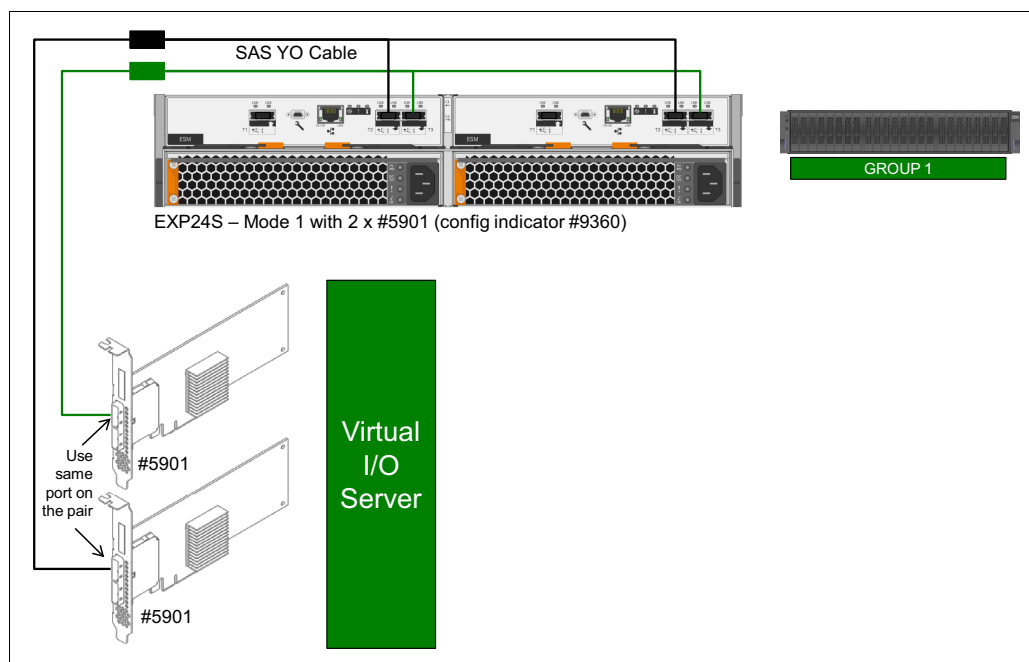


Figure 2-29 Scenario 2: Basic redundant connection

The following feature codes are required for this scenario:

- ▶ 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 that are used on the SAS adapters must be the same for both adapters of the pair. There is no SSD support for 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, which allows for up to 12 disks to be attached to each Virtual I/O Server. The connections and components of the solution are shown in Figure 2-30.

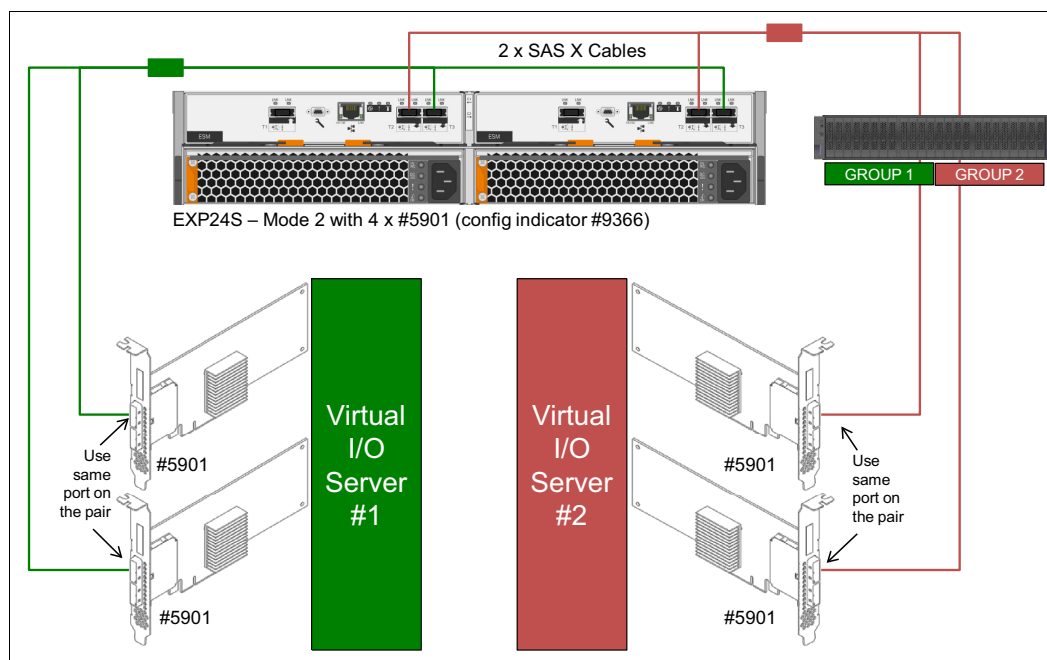


Figure 2-30 Dual Virtual I/O Servers sharing a single EXP24S

The following feature codes are required for this scenario:

- ▶ 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 that are used on the SAS adapters must be the same for both adapters of the pair. There is no SSD support for 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, which is allowed for up to 24 disks to be attached to each Virtual I/O Server (2 per drawer).

When compared to scenario 3, this scenario has the benefit to allow disks from different EXP24S drawers to be mirrored, which allows for hot maintenance of the entire EXP24S drawers if all data is properly mirrored. The connections and components of the solution are shown in Figure 2-31.

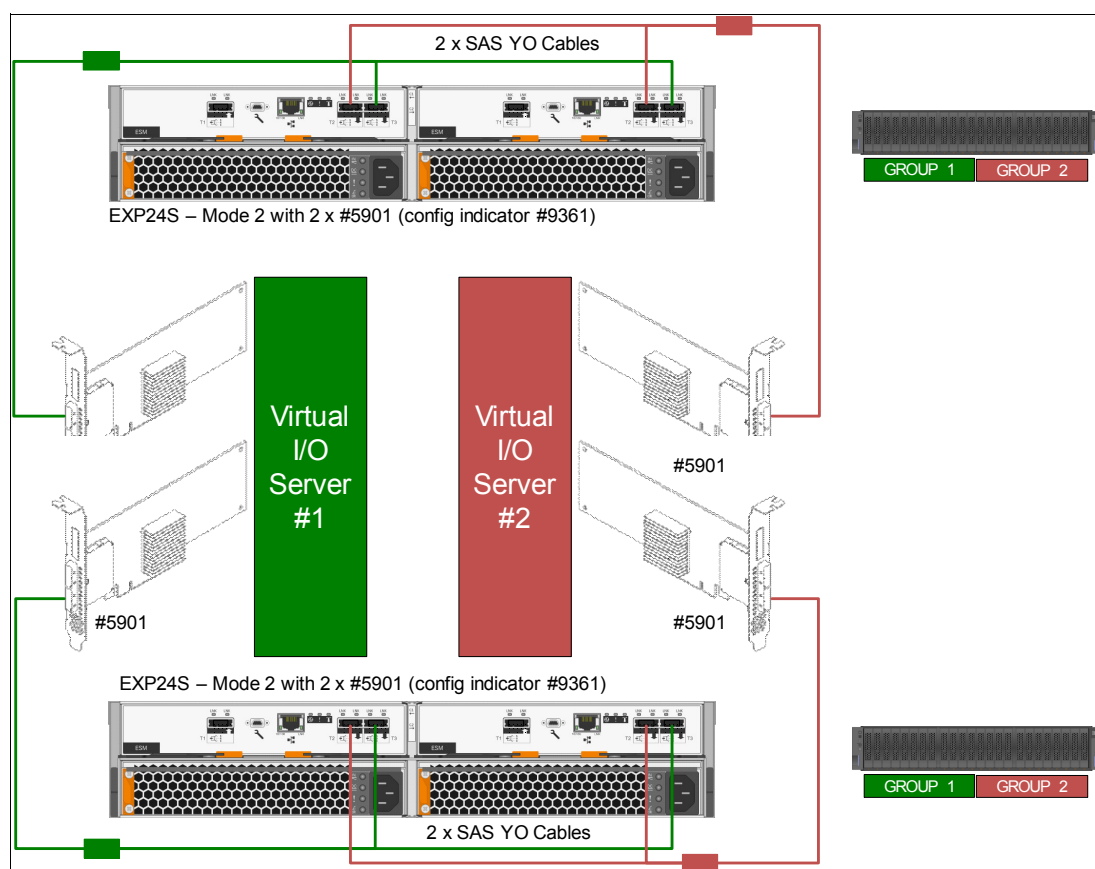


Figure 2-31 Dual Virtual I/O Servers sharing two EXP24S

The following feature codes are required for this scenario:

- ▶ 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 for 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, which allows for up to 12 disks to be attached to each Virtual I/O Server (6 per drawer). This scenario includes the benefit to allow disks from different EXP24S drawers to be mirrored, which allows for hot maintenance of the entire EXP24S drawers if all data is properly mirrored. The connections and components of the solution are shown in Figure 2-32.

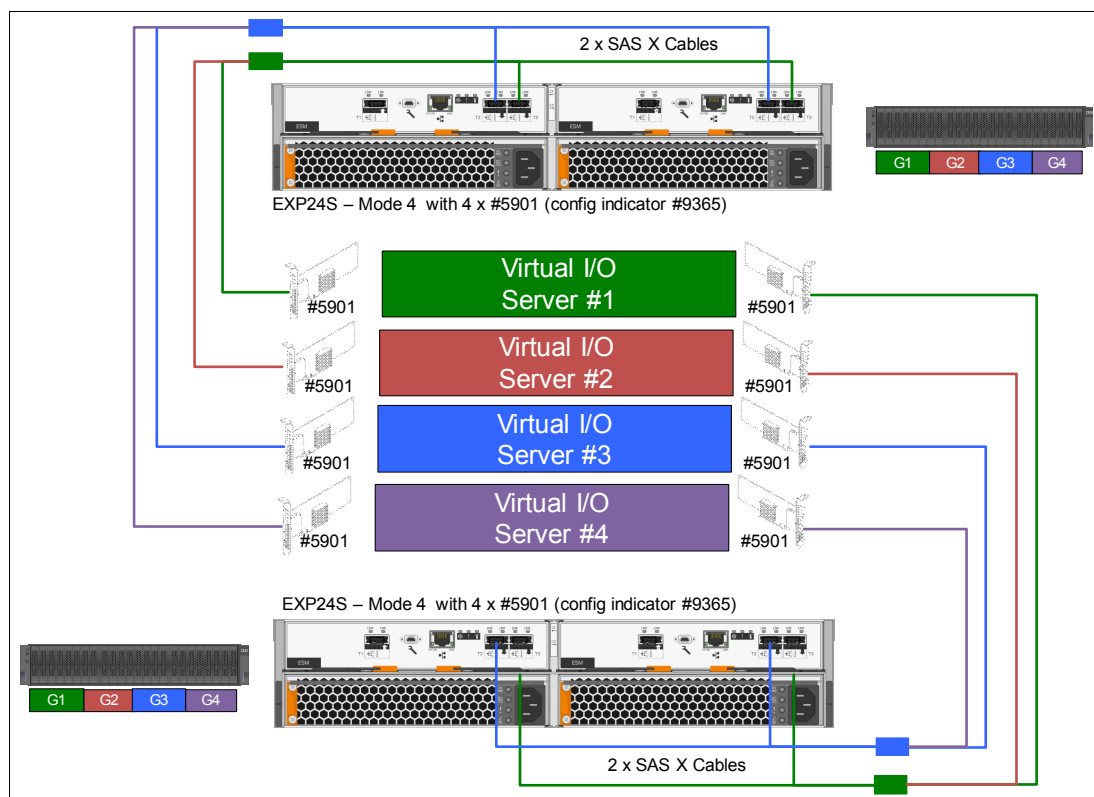


Figure 2-32 Four Virtual I/O Servers sharing two EXP24S

The following feature codes are required for this scenario:

- ▶ 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 for this scenario.

Other scenarios

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

<http://www.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 this website:

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

The following section highlights some of the offerings.

IBM Storwize family

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

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

IBM FlashSystem family

The IBM FlashSystem® family 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 this website:

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

IBM XIV Storage System

The IBM XIV® Storage System hardware is part of the Spectrum Accelerate family and is a high-end disk storage system that helps 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 Storage Systems extend ease of use with integrated management for large and multi-site XIV deployments, which reduces operational complexity and enhances 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® storage subsystem is a high-performance, high-capacity, and secure storage system that delivers the highest levels of performance, flexibility, scalability, resiliency, and total overall value for the most demanding, heterogeneous storage environments. The system effectively and efficiently manages a broad scope of storage workloads in today's complex data center.

The IBM System Storage DS8000 also includes a range of features that automate performance optimization and application quality of service. It also provides 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

An HMC is a dedicated appliance with which administrators 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 IBM 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.

The HMC can be a hardware or a virtual appliance, which are described next. You can use IBM Power Systems HMCs in any of the following configurations:

- ▶ Only hardware appliance HMCs
- ▶ Only virtual appliance HMCs
- ▶ A combination of hardware appliance and virtual appliance HMCs

Requirements: When any HMC is used with the Power E870C and Power E880C servers, the HMC code must be running at V850.10 level, or later. The minimum firmware level for the Power E870C and Power E880C is V840.30, or later.

2.11.1 Hardware appliance HMC

The 7042-CR9 hardware appliance HMC is a dedicated rack-mounted workstation. It can be used to manage any of the systems that are supported by the version 8 HMC. It provides hardware, service, and basic virtualization management for your IBM Power Systems servers.

HMC RAID 1 support

A high availability feature is available for the hardware appliance HMC. By default, the 7042-CR9 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 configuration 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 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) desktside HMC. The latest HMC model is the 7042-CR9.

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

2.11.2 Virtual appliance HMC

The 5765-HMV virtual appliance HMC can be used to manage any of the systems that are supported by the version 8 HMC. It provides hardware, service, and basic virtualization management for your IBM Power Systems servers. The virtual HMC runs as a virtual machine on an x86 server that is virtualized by VMware ESXi or Red Hat KVM.

It offers the same functionality as the hardware appliance HMC. However, the virtual appliance HMC features the following differences from the hardware appliance HMC:

- ▶ An activation engine, which provides unique configuration during initial deployment.
- ▶ The way the license acceptance dialog is presented.
- ▶ Support for multiple virtual disks for more data storage.
- ▶ Although formatting physical media is not supported, it is supported by using a virtual device that is attached to the VM.

At the time of this writing, the following requirements must be met:

- ▶ Hardware:
 - x86 64-bit hardware with hardware virtualization assists (Intel VT-x or AMD-V)
 - Resources for the HMC virtual appliance VM: Four CPUs, 8 GB of memory, 160 GB of disk space, and two network interfaces
- ▶ Software requirements: VMware ESXi V5, or later, or Red Hat Enterprise Linux 6.x with KVM

2.11.3 HMC code level

When the Power E870C and Power E880C servers is used with any HMC, the HMC code must be a minimum level of V850.10, or later. The minimum firmware level for the Power E870C and Power E880C is V840.30, or later.

When the Power E870C and Power E880C server is used to access the cloud-based HMC Apps as a Service, the HMC code must be at a minimum level of V860.

If you are attaching an HMC to a new server or adding a function to a 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, see 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 might verify entitlement through customer number, serial number electronic restrictions, or any other means or methods that are employed by IBM at its discretion.

2.11.4 HMC connectivity to the POWER8 processor-based systems

POWER8 processor-based servers and their predecessor systems that are managed by a hardware or virtual appliance HMC require Ethernet connectivity between the HMC and the server's service processor. If dynamic LPAR, Live Partition Mobility, or PowerVM Active Memory Sharing operations are required on the managed partitions, Ethernet connectivity also 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 the HMC1 or HMC2 ports of the managed server, although other network configurations are possible. You can 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.

A simple network configuration to enable the connection from the HMC to the server and to allow for dynamic LPAR operations is shown in Figure 2-33. For more information about HMC and the possible network connections, see *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491:

<http://www.redbooks.ibm.com/abstracts/sg247491.html>

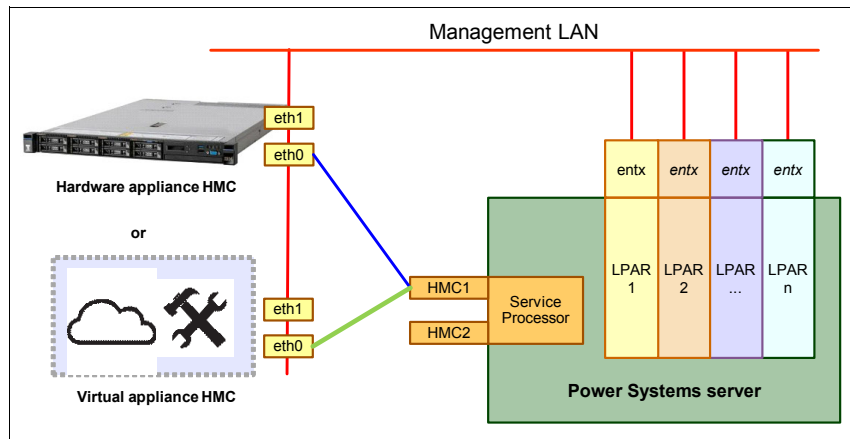


Figure 2-33 Network connections from the HMC to service processor and LPARs

By default, the service processor HMC ports are configured for dynamic IP address allocation. The HMC can be configured as a DHCP server, which provides an IP address at the time that the managed server is powered on. In this case, the 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 also is an option and can be configured by using the ASMI menus.

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

- ▶ Run at a speed of 1 Gbps.
- ▶ Are visible to the service processor only 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.

The following network configuration are used 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 61.

2.11.5 High availability HMC configuration

The HMC is an important hardware component. Although IBM 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 that uses 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. Consider the following points:

- ▶ 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 communicate with the server's partitions over a public network to allow for full synchronization and functionality.

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 E870C and Power E880C 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 more information about available software on IBM Power Systems, see this IBM Power Systems Software™ website:

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

2.12.1 Virtual I/O Server

The following minimum levels of Virtual I/O Server for the Power E870C and Power E880C are required:

- ▶ VIOS 2.2.3.4 with iFix IV63331
- ▶ VIOS 2.2.2.6

IBM regularly updates the Virtual I/O Server code. For more information about the latest updates, see this website:

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

2.12.2 IBM AIX operating system

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

IBM periodically releases maintenance packages (service packs or technology levels) for the AIX operating system. For more information about these packages, downloading, and obtaining the CD-ROM, see this website:

<http://www.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, see this website:

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

The minimum required levels for AIX are listed in Table 2-30.

Table 2-30 Supported AIX levels on the E870C and E880C

AIX Version	With Virtual I/O Server	Without Virtual I/O Server
6.1	<ul style="list-style-type: none">▶ 6100-08 Technology Level and Service Pack 1, or later▶ 6100-09 Technology Level and Service Pack 1, or later	<ul style="list-style-type: none">▶ 6100-08 Technology Level and Service Pack 6, or later▶ 6100-09 Technology Level and Service Pack 4, and APAR IV63331, or later
7.1	<ul style="list-style-type: none">▶ 7100-02 Technology Level and Service Pack 1 or later▶ 7100-03 Technology Level and Service Pack 1 or later	<ul style="list-style-type: none">▶ 7100-02 Technology Level Service Pack 6, or later▶ 7100-03 Technology Level Service Pack 4, and APARs IV63332, or later▶ 7100-04 Technology Level, or later
7.2	7200-00 Technology Level, or later	7200-00 Technology Level, or later

2.12.3 IBM i operating system

The IBM i operating system is supported on the Power E870C and Power E880C with the following minimum required levels:

- ▶ IBM i 7.1 with 7.1.0 machine code RS710-S, or later
- ▶ IBM i 7.2 TR4 or later
- ▶ IBM i 7.3, or later

IBM periodically releases maintenance packages (service packs or technology levels) for the IBM i operating system. For more information about these packages, downloading, and obtaining the CD-ROM, see this website:

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

For more information about hardware features compatibility and the corresponding AIX and IBM i Technology Levels, see the following IBM Prerequisite website:

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 E870C and Power E880C are listed in Table 2-31.

Table 2-31 Supported Linux versions on the E870C and E880C

Linux vendor	Big endian	Little endian
SUSE	SUSE Linux Enterprise Server 11 Service Pack 3, or later	SUSE Linux Enterprise Server 12 and later Service Packs
RedHat	Red Hat Enterprise Linux 6.5 for POWER, or later	Red Hat Enterprise Linux 7.1, or later
Ubuntu	N/A	<ul style="list-style-type: none">▶ Ubuntu 14.04.3, or later▶ Ubuntu 16.04, or later

If you want to configure Linux partitions in virtualized IBM 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 that are 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, see the following resources:

- ▶ Features and external devices that are supported by Linux:
<http://www.ibm.com/systems/p/os/linux/index.html>
- ▶ SUSE Linux Enterprise Server:
<http://www.novell.com/products/server>
- ▶ Red Hat Enterprise Linux Advanced Server:
<http://www.redhat.com/rhel/features>
- ▶ Ubuntu:
<http://www.ubuntu.com/server>

2.12.5 Supported Java versions

Java is supported on POWER8 servers. For best use of the performance capabilities and most recent improvements of POWER8 technology, upgrade Java based applications to Java 8, Java 7, or Java 6. For more information, see the following websites:

- ▶ <http://www.ibm.com/developerworks/java/jdk/aix/service.html>
- ▶ <http://www.ibm.com/developerworks/java/jdk/linux/download.html>

2.13 Energy management

The Power E870C and Power E880C systems include features to help clients become more energy-efficient. IBM 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 includes 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, which reduces 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%, which is 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 start when the system begins 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, which implies that as the frequency of a processor increases, its usage decreases given a constant workload. Dynamic power saver mode uses this relationship to detect opportunities to save power that are 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 reduces 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. The maximum available frequency boost for different speed processors in the Power E870C and E880C is listed in Table 2-32.

Table 2-32 Maximum frequency boosts for Power E870C and E880C processors

System	Cores per chip	Nominal speed	Maximum boost speed
Power E870C	8	4.024 GHz	4.123 GHz
Power E880C	8	4.356 GHz	4.522 GHz
Power E880C	10	4.190 GHz	4.456 GHz
Power E880C	12	4.024 GHz	4.256 GHz

The frequency boost figures are maximums and depend on the environment in which the servers are installed. Maximum boost frequencies might 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 specific 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, which significantly degrades performance.

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 *margin*ed 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 often 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 and soft power capping. Soft power capping extends the allowed energy capping range beyond a region that can be ensured in all configurations and conditions. If the energy management goal is to meet a particular consumption limit, soft power capping is the mechanism to use.

► Processor core nap mode

IBM POWER8 processor uses a low-power mode that is named *nap* that stops processor execution when there is no work to do on that processor core. The latency of exiting nap mode is small, which often does not generate any affect on running applications. 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, which is 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 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 wink mode

The most amount of energy can be saved when a whole POWER8 chiplet enters *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. IBM Power Systems are designed to operate in worst-case environments, such as 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 (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 slots 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. This configuration reduces the number of servers that are powered on, which also reduces 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 2-34.

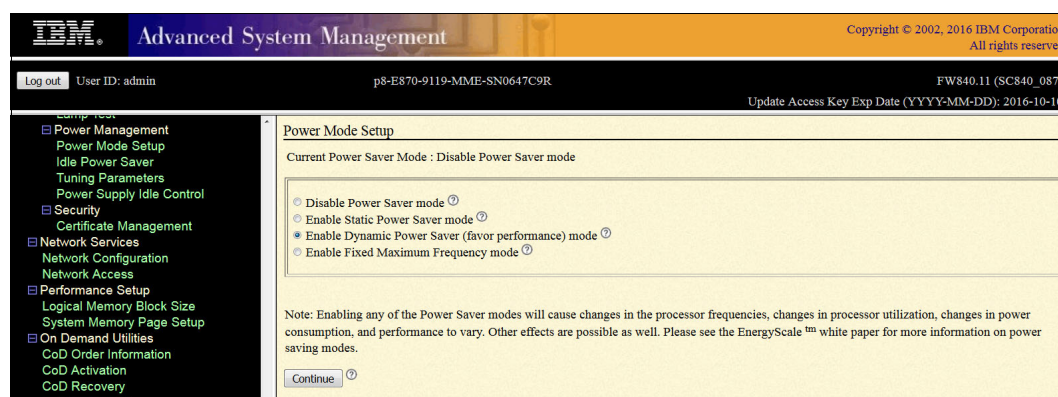


Figure 2-34 Setting the power mode in ASMI

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 OCC that uses 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+, which enables reaction under the timescale of a typical OS time slice and allows 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 voltage and frequency for workload variation enables better increase in power savings versus optimizing frequency only.

2.13.3 Energy consumption estimation

The following energy-related values are important for IBM Power Systems:

- Maximum power consumption and power source loading values

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

<http://www.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm>

At this website, search for and enter the model number and “server specifications”. For example, for the Power E870C and Power E880C system, search for “9080-MME and 9080-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, which is available at this website:

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

In this tool, select the type and model for the system and enter some information about the configuration and wanted CPU usage. The tool shows the estimated energy consumption and the waste heat at the wanted usage and at full usage.



Private and Hybrid Cloud features

The Power E870C and the Power E880C servers include the cloud management software and services to assist with the move to the cloud, both private and hybrid. This feature allows for fast and automated virtual machine (VM) deployments that are based on prebuilt image templates and self-service capabilities, all with an intuitive interface.

This chapter provides an overview of this software and services and includes the following topics:

- ▶ 3.1, “Private cloud software” on page 106
- ▶ 3.2, “Hybrid cloud support” on page 110
- ▶ 3.3, “Geographically Dispersed Resiliency for Power” on page 112
- ▶ 3.4, “IBM Power to Cloud Rewards Program” on page 113

3.1 Private cloud software

Companies are creating private clouds that join characteristics from internal private environments and public clouds to create a more flexible, agile, secure, and customized infrastructure environment.

Having a single point to control this infrastructure is key to achieving these goals. It is here that the private cloud software comes into place.

The following sections describe the software that is used to allow for the Power E870C and Power E880C to be part of a private cloud. Also described are the tools that are available to ease the deployment and management of VMs in a private cloud.

3.1.1 IBM Cloud PowerVC Manager

Managing a private cloud requires software tools to help create a virtualized pool of compute resources, provide a self-service portal for users and policies for resource allocation, control, security, and metering data for resource billing. Management tools for private clouds often are service-driven instead of resource-driven because cloud environments typically are highly virtualized and organized concerning portable workloads.

The OpenStack based IBM Cloud PowerVC Manager provides the self-service cloud portal for IBM Power Systems. This portal enables users to quickly request cloud resources and reliably deploy VMs with approval policies to maintain control over provisioning of cloud resources.

IBM Cloud PowerVC Manager provides the following features:

- ▶ Create VMs and resize the VMs CPU and memory and attach disk volumes to those VMs
- ▶ Import VMs and volumes so that they can be managed by PowerVC
- ▶ Monitor the use of resources in your environment
- ▶ Migrate VMs while they are running (live migration between physical servers)
- ▶ Improve resource usage to reduce capital expense and power consumption
- ▶ Increase agility and execution to quickly respond to changing business requirements
- ▶ Increase IT productivity and responsiveness
- ▶ Simplify IBM Power Systems virtualization management
- ▶ Accelerate repeatable, error-free virtualization deployments

IBM Cloud PowerVC Manager can manage AIX, IBM i, and Linux VMs that are running under PowerVM virtualization and Linux VMs that are running under PowerKVM virtualization.

After it is Openstack based, IBM Cloud PowerVC Manager provides upward integration with hybrid cloud orchestration products. This feature allows for IBM Power Systems to take part in a heterogeneous private cloud infrastructure. For more information, see 3.2.1, “Hybrid infrastructure management tools” on page 110.

IBM Cloud PowerVC Manager extends Openstack with the extra components, so IBM Power Systems can be easily managed as with any other platform. The Openstack components and IBM additions are shown in Figure 3-1.

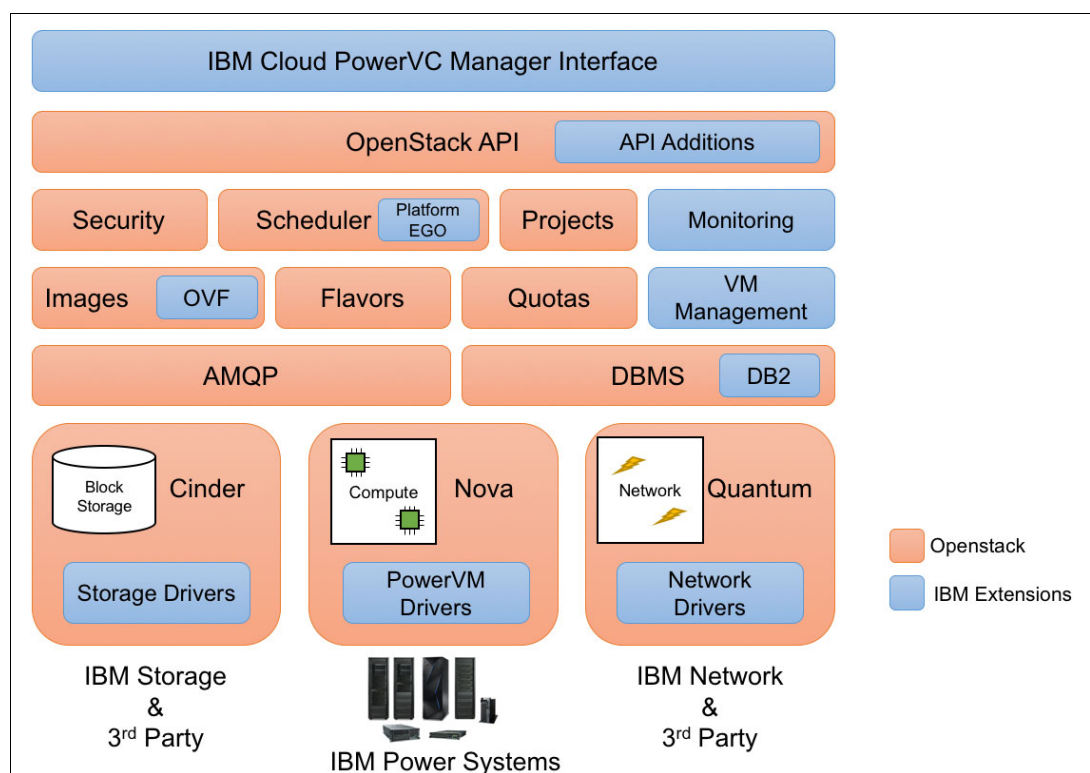


Figure 3-1 Openstack components and IBM Cloud PowerVC Manager additions

For more information about IBM Cloud PowerVC Manager, see *IBM PowerVC Version 1.3.1 Introduction and Configuration*, SG24-8199:

<http://www.redbooks.ibm.com/abstracts/sg248199.html>

3.1.2 Cloud-based HMC Apps as a Service

The new HMC Apps as a Service provides powerful insights into your IBM Power Systems infrastructure by using a set of hosted as-a-service applications, with no extra software or infrastructure setup.

By using these new applications, clients can aggregate IBM Power Systems performance, auditing, and inventory data from across their enterprise, which removes the burden of manual collection and collation. These IBM-developed applications are hosted in a secure, multi-tenant cloud and provide health scores, search and filtering, and threshold-based alerts that can be accessed through a secure portal from desktops or mobile devices.

The applications gather data that is sent by the customer HMCs over the Internet by using the built-in cloud connector that is included with HMC V8R8.6.0 and NovaLink v1.0.0.4.

A company with three datacenters that are consolidating information that is provided by HMC and Novalink into HMC Apps as a Service is shown in Figure 3-2.

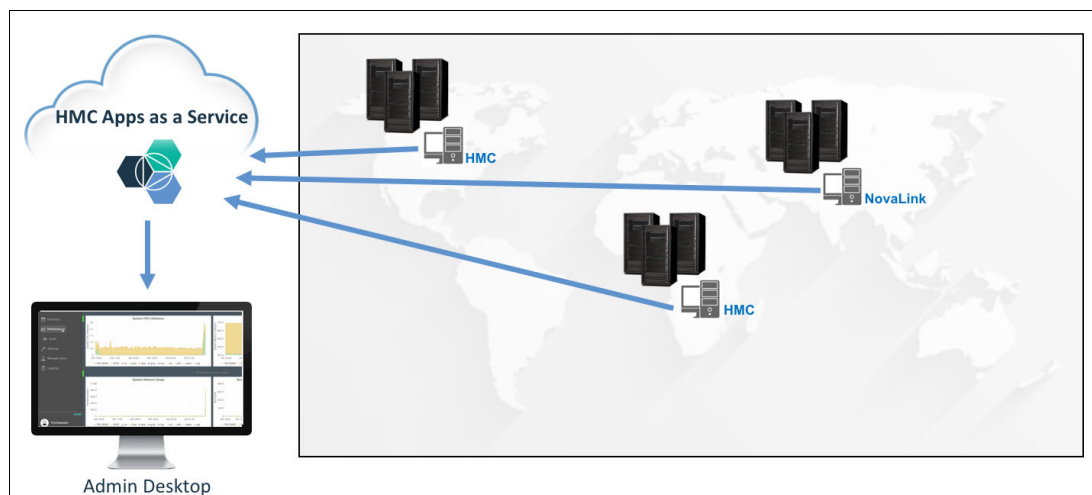


Figure 3-2 Logical diagram of HMC Apps as a Service consolidating information from several sites

After the data is gathered, it can be organized and analyzed to provide the administrators with a single point for the following functions:

- ▶ View all IBM Power Systems, VIOS, HMCs, NovaLinks, and LPARs across the entire enterprise
- ▶ Organize and show resources by tags
- ▶ See basic health and state
- ▶ Check hardware inventory
- ▶ Aggregated performance views across your Power enterprise
- ▶ Manage thresholds and alerts

When clients purchase a Power E870C or Power E880C server, they are entitled to this new service offering for no extra charge.

Note: As of the time of this writing, the performance and inventory applications are scheduled to be offered in a technology preview in 2016 and to be followed by a full GA offering with more applications in 2017.

3.1.3 Open source cloud automation and configuration tooling for AIX

IBM expanded its commitment to keep key open source cloud management packages updated and to provide timely security fixes to enable clients to use open source skills. IBM Power Systems servers are well-positioned to use the following key packages that were recently provided to enable cloud automation:

- ▶ Chef

Chef is an automation platform for configuration, deployment, and management of VMs. It is based on a client/server architecture in which the server stores the policies that are applied to nodes (known as *Cookbooks*) and their associated metadata. After a Cookbook is created, it can be applied to a client VM being deployed, configuring this VM according to the policies established automatically.

IBM is collaborating with clients in this community to provide useful resources for the use of Chef with AIX systems. Chef-client for AIX is now enhanced with new recipes and the AIX cookbook is available at this website:

<https://supermarket.chef.io/cookbooks/aix/>

► Yum

Yellowdog Updater, Modified (Yum) allows for automatic updates, package, and dependency management on RPM-based distributions.

It is now available with repository access from FTP and https protocols for AIX. Yum also is updated to enable automatic dependency discovery. For more information, see the following AIX Toolbox for Linux Applications website:

<http://www.ibm.com/systems/power/software/aix/linux/toolbox/alpha.html>

► Cloud-init

Cloud-init is a tool that helps with the early initialization of a VM that is being deployed. It uses Openstack metadata information to set a root password, grow file systems, set a default locale, set hostnames, generate ssh private keys, and handle temporary mount points. After Cloud-init is script based, it can be extended to perform other tasks that are specific to a customer environment.

Cloud-init and its dependencies are available and include support for licensed AIX users. For more information, see the following AIX Toolbox for Linux Applications website:

<http://www.ibm.com/systems/power/software/aix/linux/toolbox/alpha.html>

► GitHub

GitHub is a version control system that manages and stores revisions of projects. GitHub provides a web-based interface, access control, and several collaboration features, such as task management and wikis.

In the past, if you wanted to contribute to an open source project, you download the source code, made and noted changes, talked to the owner of the code to explain the changes, and then got the owner's approval to apply your changes to the official code.

With GitHub, a user can clone a project, download and change only the needed files, pull the new code to the platform, and notify the owner that can publish the changes automatically with version control.

Open source projects for AIX can be found at the following repository:

<http://github.org/aixoss>

► Node.js

Node.js is a platform that is built on JavaScript for building fast and scalable applications. Node.js uses an event-driven, non-blocking I/O model that makes it lightweight and efficient, which is ideal for data-intensive and real-time applications that run across distributed devices.

Node.js is available for Linux and AIX platforms and can be downloaded from this website:

<https://nodejs.org/en/download/>

3.2 Hybrid cloud support

Hybrid cloud is quickly becoming the de facto standard of IT. Two-thirds of organizations that blend traditional and cloud infrastructure are gaining advantage from their hybrid cloud. A hybrid cloud model enables building and deploying applications quickly with optimized usage of resources and lower costs.

In addition, the ability to centrally manage private, public, or dedicated cloud resources with a single management tool while securely connecting traditional workloads with cloud-native applications enables clients to respond to their dynamically changing business priorities in a more agile and timely fashion.

3.2.1 Hybrid infrastructure management tools

IBM Power Systems OpenStack-based PowerVC management upwardly integrates into various third-party hybrid cloud orchestration products, including IBM Cloud Orchestrator and VMware vRealize. Clients can manage their private cloud VMs and their public cloud VMs from a single, integrated management tool.

vRealizes uses PowerVC Openstack capabilities to interact with PowerVM and PowerKVM, which allows for management and integration of Power based VMs into a single orchestration platform. The OpenStack-based PowerVC upward integration with VMware vRealize is shown in Figure 3-3.

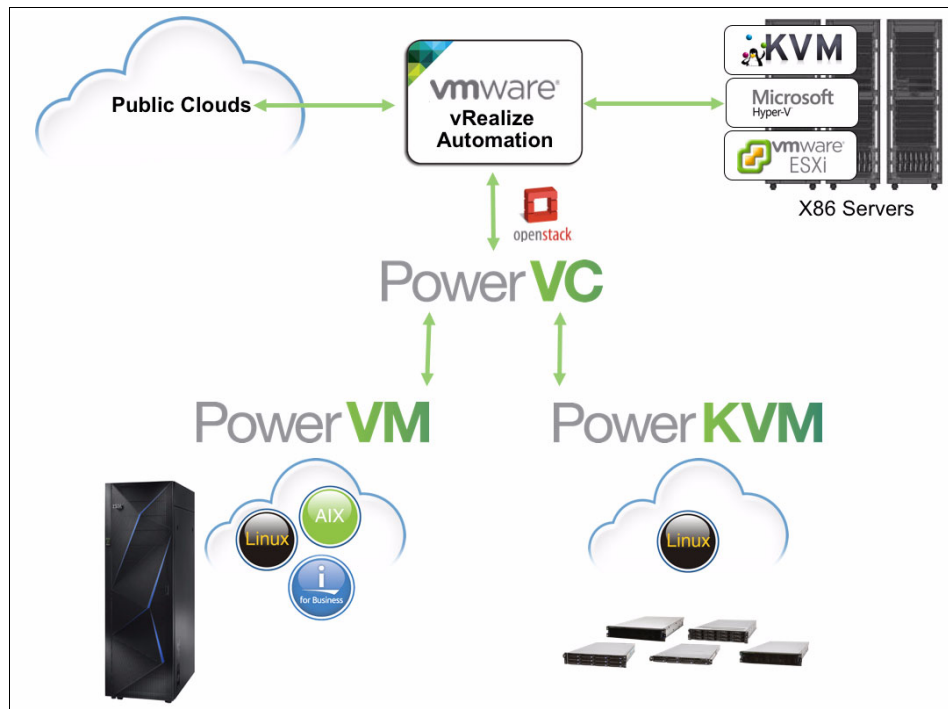


Figure 3-3 OpenStack-based PowerVC integration with VMware vRealize

3.2.2 Securely connecting system of record workloads to cloud native applications

IBM's API Connect and IBM WebSphere Connect provide secure connectivity to cloud-based applications. By using this connectivity, clients can rapidly develop new applications and services, which accelerates their time to value.

IBM's Power to Cloud services also can help clients get started with these solutions and design new applications that use IBM Bluemix®. The Bluemix tool enables clients to rapidly build, deploy, and manage their cloud applications, while tapping a growing system of available services and runtime frameworks.

3.2.3 IBM Cloud Starter Pack

To help clients get started with their hybrid cloud infrastructure, the Power 870C or Power E880C offering includes one year of a C812L-M POWER8 Linux bare metal system in the IBM Cloud (SoftLayer®) that features the following configuration:

- ▶ 10-core POWER8 processor 3.49 GHz
- ▶ 256 GB RAM
- ▶ Two 4 TB SATA HDD
- ▶ Ubuntu Linux

3.2.4 Flexible capacity on-demand

With the purchase of a new Power 870C or Power E880C server, clients can convert previously purchased capacity to SoftLayer Linux on Power bare metal monthly server usage.

Allowed types of capacity are Mobile Processor activations and Elastic COD Processor Days.

This feature allows for preserving investments after a customer can convert excess unused capacity into PowerLinux bare metal servers that are running on public cloud.

The feature can be used for the following purposes:

- ▶ Offload workloads to the public cloud to make room for critical private cloud workloads
- ▶ Provide more processing capacity for applications during peaks
- ▶ Use the public cloud as a DR site for the Linux on Power VMs
- ▶ Move development VMs to public cloud

3.3 Geographically Dispersed Resiliency for Power

Geographically Dispersed Resiliency for Power (GDR for Power) is a tool that allows for the remote restart of failed VMs in a disaster recovery scenario.

Most of the disaster recovery scenarios for virtual environments are based on two approaches, each with its own benefits and points of attention: Clustered VMs and Virtual Machine Remote Restart. These disaster recovery models are shown in Figure 3-4.

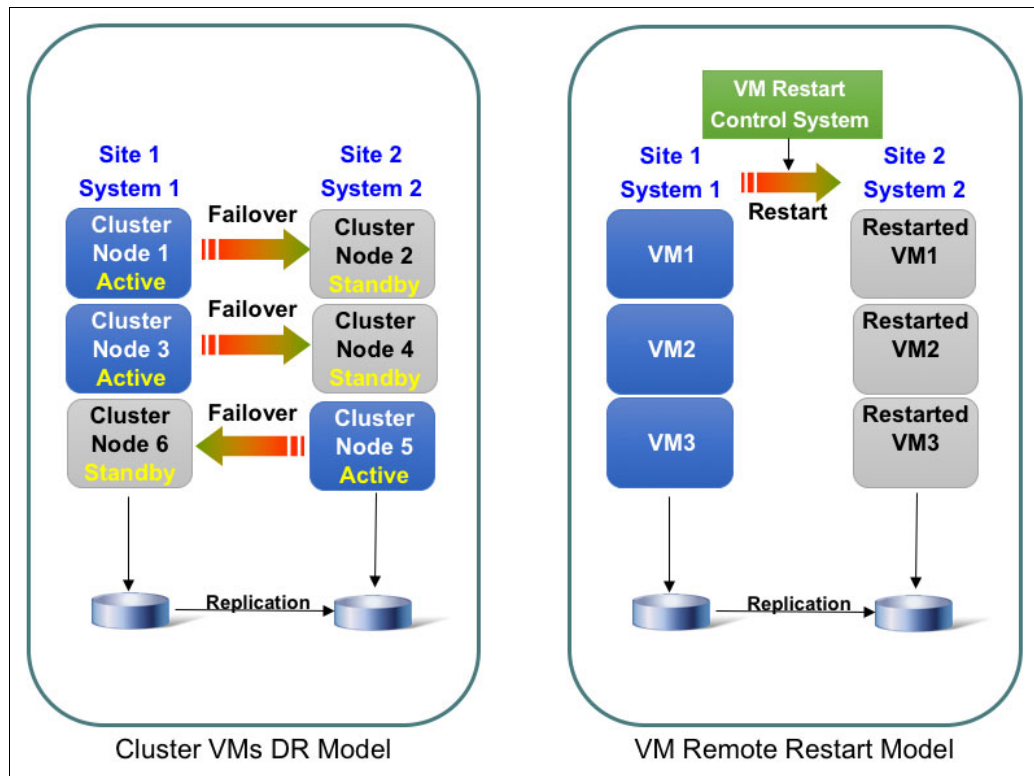


Figure 3-4 Disaster recovery models comparison for virtualized environments

Although critical applications can benefit from the clustered VMs model because of the faster RTO, its implementation might be more complex for many VMs. However, non-critical environments might benefit from the Virtual Machine Remote Restart model because of the simpler implementation and maintenance of the environment. Some aspects of both models are listed in Table 3-1.

Table 3-1 Disaster recovery models comparison for virtualized environments

	Clustered VMs	Virtual Machine Remote Restart
Complexity	Individual VM configuration	Host configuration
Failover Time	Faster (seconds to minutes)	Slower (minutes)
Cost	Higher because of extra software licenses and labor	Lower
Maintenance	Each new VM requires a new cluster configuration	After the host is configured, new VMs are automatically covered

By integrating with Power Enterprise Pool, GDR for Power allows for core and memory activations movement between machines during a DR scenario. This feature reduces the investment on the DR site.

3.4 IBM Power to Cloud Rewards Program

IBM Power to Cloud Rewards Program helps clients design, build, and deliver clouds platforms on IBM Power Systems servers. The IBM Power to Cloud Rewards Program transforms the successful IBM PowerCare program to a new, points-based reward system and helps accelerate the transformation of IT infrastructure to private and hybrid cloud.

The IBM Power to Cloud Rewards Program enables clients to earn reward points on purchases of IBM Power Systems servers, including the Power E870C and Power E880C.


IBM Power to Cloud Reward points can be used for various services that are focused on helping the transition from traditional IT platforms to private and hybrid cloud platforms. Power to Cloud Reward Program services offerings use the proven expertise of IBM Systems Lab Services consultants.

The following services are offered under the IBM Power to Cloud Rewards Program:

- ▶ Design for Cloud Provisioning and Automation
- ▶ Build Infrastructure as a Service for Private Cloud
- ▶ Build Cloud Capacity Pools across Data Centers
- ▶ Deliver with Automation for DevOps
- ▶ Design for Hybrid Cloud Workshop
- ▶ Deliver with Database as a Service
- ▶ Build and Provision for Mobility and Automation
- ▶ Design for Private Cloud Monitoring and Capacity Planning

For clients that are looking for a hybrid cloud solution, workshops services are available to provide instruction about how to produce best in class applications by using API Connect and Bluemix with IBM Power Systems.

For more information about the full offering list and program details (including points redemption value for specific services offers), and scope, contact your IBM Systems Lab Services representative.



Reliability, availability, serviceability, and manageability

This chapter provides information about IBM Power Systems reliability, availability, and serviceability (RAS) design and features.

This chapter includes the following topics:

- ▶ 4.1, “RAS enhancements of POWER8 processor-based servers” on page 116
- ▶ 4.2, “Reliability” on page 118
- ▶ 4.3, “Processor/Memory availability” on page 120
- ▶ 4.4, “Enterprise systems availability” on page 126
- ▶ 4.5, “Availability effects of a solution architecture” on page 127
- ▶ 4.6, “Serviceability” on page 129
- ▶ 4.7, “Manageability” on page 139
- ▶ 4.8, “Selected POWER8 RAS capabilities by operating system” on page 151

4.1 RAS enhancements of POWER8 processor-based servers

IBM Power Systems RAS features the following elements:

- ▶ **Reliability:** Indicates how infrequently a defect or fault in a server occurs.
- ▶ **Availability:** Indicates how infrequently the functioning of a system or application is affected 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.

The following features were included in the entire 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 IBM Power Systems, which uses a higher RAS even for scale-out equipment:

- ▶ **Processor Enhancements Integration**

POWER8 processor chips are implemented by 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 dynamically substituted for a failing data lane during system operation.

A POWER8 processor module includes improved performance compared to POWER7+, including support of a maximum of 12 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 more 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+. The OCC also can 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 E870C and Power E880C servers also support a PCIe switch to provide more integrated I/O capacity.

- ▶ **Memory Subsystem**

Custom DIMMs (CDIMMS) are used, which contain a spare DRAM module per port (per nine DRAMs for x8 DIMMs). These ports can be used to avoid replacing memory. This feature is in addition to the ability to correct a single DRAM fault within an error-correcting code (ECC) word (and then an extra bit fault) to avoid unplanned outages.

The Power E870C and Power E880C systems include the option of mirroring the memory that is used by the Hypervisor. This feature reduces the risk of system outage that is linked to memory faults, as the Hypervisor memory is stored in two distinct memory CDIMMs.

- **Power Distribution and Temperature Monitoring**

All systems use voltage converters that transform the voltage level that is provided by the power supply to the voltage level that is needed for the various components within the system. The Power E870C and Power E880C systems contain two converters for each voltage level that is provided to any specific processor or memory DIMM.

Converters that are used for processor voltage levels are configured for redundancy so that when one is detected as failing, it is called out for repair while the system continues to run with the redundant voltage converter.

The converters that are used for memory are configured with a form of sparing. In this configuration, the system continues operation with another converter without generating a service event or the need to take any sort of outage for repair when a converter fails.

The processor module integrates a new OCC. This OCC is used to handle Power Management and Thermal Monitoring without the need for a separate controller, as was required in POWER7+. The OCC also can be programmed to run other RAS-related functions independent of any host processor.

The E880C and E870C systems use triple redundant ambient temperature sensors.

4.1.1 POWER8 overview

The POWER8 processor-based servers are available in two different classes:

- **Scale-out systems:** For environments that consist 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 that require systems with increased availability. In such environments, mission-critical applications can use 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 use all of the advanced RAS characteristics of the POWER8 processor through the whole portfolio, which offers 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 IBM 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 effect 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 by using the designed ED/FI capabilities.

Fault isolation and recovery of the POWER8 processor and memory subsystems use a dedicated service processor and are meant to be largely independent of any operating system or application deployed.

The Power E870C and Power E880C are enterprise class systems that support the highest levels of RAS. For more information about the features of enterprise class systems, see 4.4, “Enterprise systems availability” on page 126.

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 improves overall system reliability with each successive generation of system offerings. Reliability can be improved in primarily in the following ways:

- ▶ Reducing the number of components
- ▶ Using higher reliability grade parts
- ▶ Reducing the stress on the components

In the POWER8 systems, elements of these methods 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; for example, integrating processor cores on a single POWER chip. The POWER8 chip features more cores per processor module. 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, as shown in Figure 4-1 on page 119.

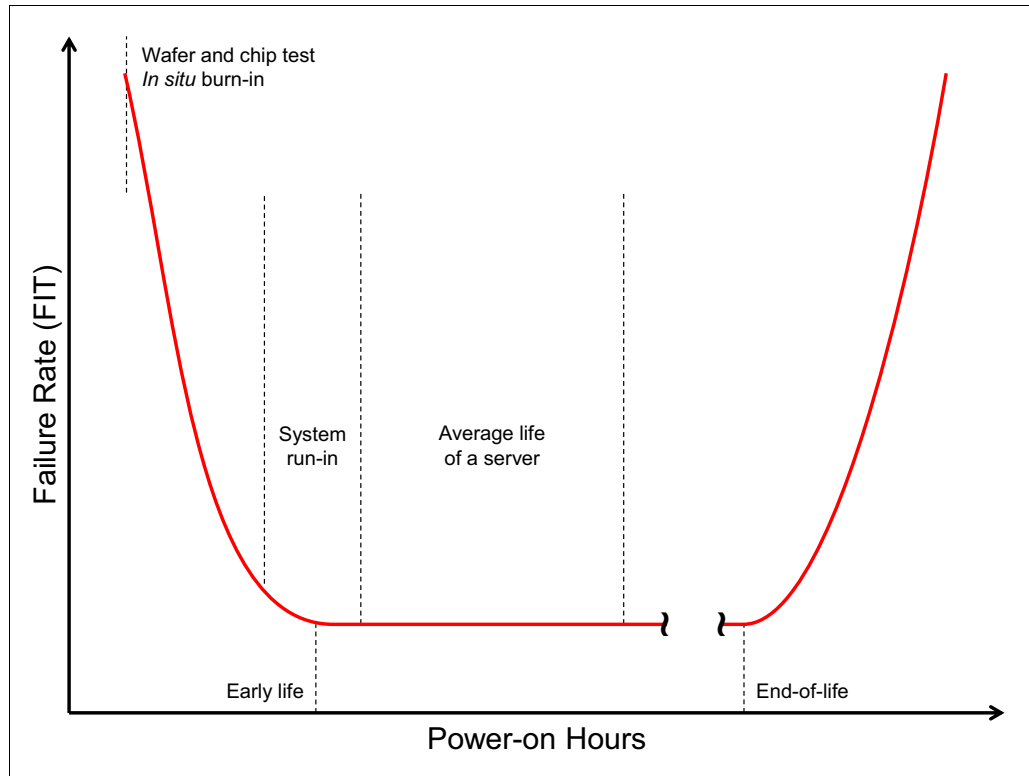


Figure 4-1 Failure rate bathtub curve

4.2.2 Component placement

Packaging delivers 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

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 those faults. 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. 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 often are reported to the dedicated service processor that is 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 errors

Intermittent or soft errors often are 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 also is the risk that a correctable error is predictive of a fault that continues to worsen over time, which results 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 cause 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 issue in turn often requires that the failing hardware is replaced in the system. The resulting service action also can temporarily affect 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 that are built into the hardware. These features include the use 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 errors

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 stop only 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. The software layers then must be responsible for determining how to minimize the effect of faults.

The advanced RAS features that are built in to POWER8 processor-based systems handle certain uncorrectable errors in ways that minimize the effect of the faults, even keeping an entire system up and running after experiencing such a failure.

Depending on the fault, such recovery can 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 affected 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 also feature SEC/DED ECC protection.

SEC/DED capabilities also are 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 by using the Power On Reset Engine).

4.3.4 Processor instruction retry and other try again techniques

Within the processor core, soft error events can 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 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 that are internal to the memory controller are detected. This feature 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 that is 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 stop the failing application.

Successful migration requires that there is sufficient available spare capacity to reduce the overall processing capacity within the system by one processor core. In highly virtualized environments, the requirements of partitions often can be reduced to accomplish this task without any further effect 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. When you use this Partition Priority feature of PowerVM, 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 that is running the highest priority workload) if a partition must be terminated for alternative processor recovery to complete.

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 by using the Hardware Management Console (HMC).

4.3.6 Core contained checkstops and other PowerVM error recovery

PowerVM can handle certain other hardware faults without stopping 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 only a partition, although potentially PowerVM if the failing instruction were in a critical area of PowerVM code).

Processor designs without Processor Instruction Retry often 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. This correction is done by purging and deleting a single cache line. Such purge and delete operations are contained within the hardware, 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 included bad data.

Nothing within the system terminates simply 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, the partition operating system might be responsible for terminating itself or only the program that uses the marked page. If data is owned by the Hypervisor, the hypervisor might choose to terminate, which resulted 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, other functions are available in addition to processor cores.

POWER8 processors feature 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 that is using these features, a freeze event might be handled without an application or partition outage.

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 using the techniques that are described in this chapter. 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 where 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 that communicate with buffer modules on memory DIMMS through memory channels and access the DRAM memory modules on DIMMs, as shown in Figure 4-2.

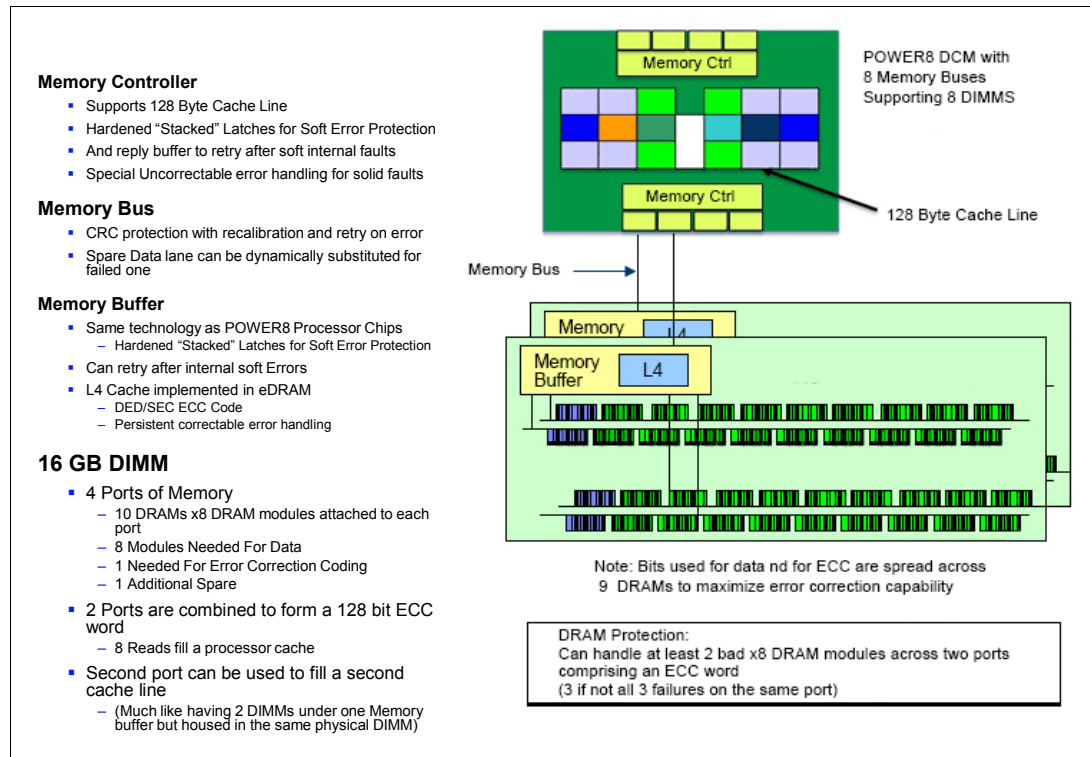


Figure 4-2 Memory protection features

The memory buffer chip is made by the same 22 nm technology that is used to make the POWER8 processor chip. The memory buffer chip also 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 and 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 that uses 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. For example, the 16 GB DIMM features one rank that is composed of four ports of x8 DRAM modules and each port contains 10 DRAM modules.

For each such port, there are eight DRAM modules worth of data (64 bits) and another DRAM module's worth of error correction and other such data. There also is 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 process is also known as *Chipkill correction*. Then, it can correct at least one other bit within the ECC word.

The extra 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. This substitution avoids 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. Then, still correct another DRAM module that is failing within the DIMM.

Other DIMMs are 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 10 x8 DRAM modules.

A 64 GB DIMM also 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 implements scrubbing 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 affected is marked with a special uncorrectable error code and handled as described in 4.3.7, “Cache uncorrectable error handling” on page 122.

4.3.11 I/O subsystem availability and Enhanced Error Handling

Multi-path I/O and VIOS for I/O adapters and RAID for storage devices must 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, IBM 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 often 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 or firmware to respond with a defined error signature whenever an attempt is made to access the device. If necessary, a special uncorrectable error code can be used to mark a section of data as bad when the freeze is first started.

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. PCIe busses that are generated directly from a host bridge can 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

In addition to all of the standard RAS features that were described in this chapter, Enterprise class systems allow for increased RAS and availability by including several features and redundant components.

The following main features are exclusive to Enterprise class systems:

- **Redundant Service Processor**

The service processor is an essential component of a system. It is responsible for the initial power load, setup, monitoring, control, and management. The control units that are present on enterprise class systems house two redundant service processors. If there is a failure in either of the service processors, the second processor ensures continued operation of the system until a replacement is scheduled. Even a system with a single system node included dual service processors in the system control unit.

- **Redundant System Clock Cards**

Another component that is crucial to system operations is the system clock cards. These cards are responsible for providing synchronized clock signals for the entire system. The control units that are present on enterprise class systems house two redundant system clock cards. If there is a failure in any of the clock cards, the second card ensures continued operation of the system until a replacement is scheduled. Even a system with a single system node included 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 effect on server performance that is caused by a failed processor. An inactive processor is activated if a failing processor reaches a predetermined error threshold, which helps to maintain performance and improve system availability.

Dynamic processor sparing happens dynamically and automatically when dynamic logical partitioning (DLPAR) is used and the failing processor is detected before failure. Dynamic processor sparing does not require purchasing an activation code. Instead, 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 effect on server performance that is 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 be in memory CDIMMs. If there is a failure of CDIMM, the hypervisor can become inoperative. The Active memory mirroring for hypervisor allows for the memory blocks that are 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 effects of a solution architecture

Any solution should not rely on only the hardware platform. Despite IBM Power Systems being far superior RAS than other comparable systems, it is advisable to design a redundant architecture that surrounds the application to allow for easier maintenance tasks and greater flexibility.

By working in a redundant architecture, some tasks that require that a specific application is brought offline can now be executed with the application running, which allows for greater availability.

When determining a highly available architecture that fits your needs, consider the following questions:

- ▶ Will I need to move my workloads off 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 that is running under PowerVM and IBM i, AIX, and Linux support a spectrum of clustering solutions. These solutions meet requirements for application availability regarding server outages and data center disaster management, reliable data backups, and so forth. These offerings include distributed applications with IBM DB2® PureScale, HA solutions that use clustering technology with IBM PowerHA SystemMirror®, and disaster management across geographies with PowerHA SystemMirror Enterprise Edition.

For more information, see the following resources:

- ▶ *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 by a single VIOS. However, if a single VIOS is used and that VIOS stops for any reason (hardware or software caused), all of the partitions that use that VIOS stop.

The use of Redundant VIOS servers mitigates this risk. Maintaining the redundancy of adapters within each VIOS (in addition to having redundant VIOS) avoids most faults that keep a VIOS from running. Therefore, multiple paths to networks and SANs are advised. A partition that is accessing data from two distinct Virtual I/O Servers, each one with multiple network and SAN adapters to provide connectivity, is shown in Figure 4-3.

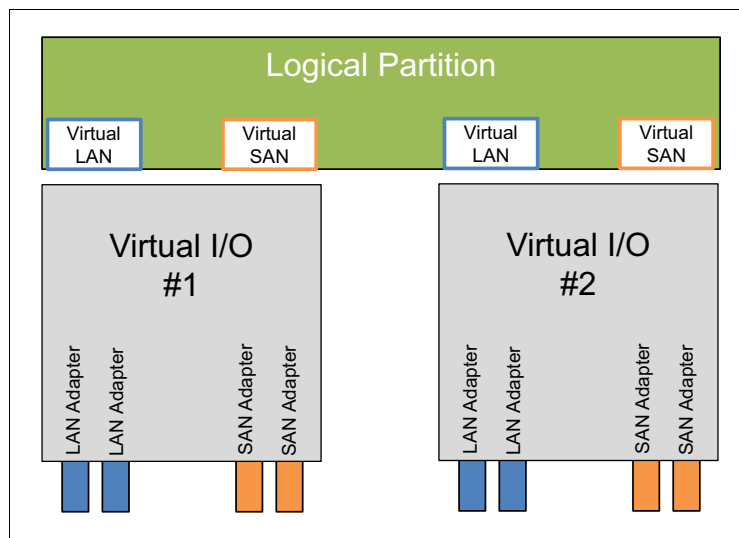


Figure 4-3 Partition with dual redundant Virtual I/O Servers

Because each VIOS can be considered as an AIX based partition, each VIOS also must access a boot image, have paging space, and so on, 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 that is locally attached to a server (DASD devices or SSDs). However accessed, the rootvgs should use mirrored or RAID drives with redundant access to the devices for best availability.

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, you can move a server's logical partitions to another server before the failure occurs if a server indicates a potential failure. Partition mobility can help avoid unplanned downtime.

- Take advantage of server optimization:
 - Consolidation: You can consolidate workloads that run on several small, under-used servers onto a single large server.
 - Deconsolidation: You can move workloads from server-to-server to optimize resource use and workload performance within your computing environment. With 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:

```
migr1par -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. It also is supported in compliance with all operating systems that are compatible with POWER8 technology.

The VIOS partition cannot be migrated.

For more information about Live Partition Mobility and how to implement it, see the links to updated information in the Abstract at *IBM PowerVM Live Partition Mobility (Obsolete - See Abstract for Information)*, SG24-7460:

<http://www.redbooks.ibm.com/abstracts/sg247460.html>

4.6 Serviceability

The purpose of serviceability is to repair the system while attempting to minimize or eliminate service cost (within budget objectives) and maintain application availability and high customer satisfaction. Serviceability includes system installation, miscellaneous equipment specification (MES) (system upgrades or downgrades), and system maintenance or 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 are available on the system and delivered through the HMC.
- ▶ Step-by-step service procedures are 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 threat to system availability, those errors 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, which extend 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 IBM 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 to re-create the problem or run an extended tracing or diagnostics program.

For many 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. 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 and 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 the following 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 start 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 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 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

When enabled, the auto-restart (reboot) option 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, set and reset service indicators (Light Path for low-end servers), and access all service processor functions without powering 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.

- Managing 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 created 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.

By 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 into IBM servers, re-creating diagnostic tests for failures or requiring user intervention is unnecessary. Solid and intermittent errors are 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 multitier 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. This determination 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; however, on the POWER8 processor, 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 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 offloaded to the HMC when it is available.

Run time

All IBM Power Systems servers can monitor critical system components during run time. They also 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. These routines enable quick and accurate predefined responses to 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 start 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 that is on the system node 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 that is 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, a Call Home service request is started and the pertinent failure data with service parts information and part locations is sent to the IBM service organization if the system is correctly configured. 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, see “Service Event Manager” on page 149.

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 the following types of basic data:

- ▶ 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. 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, which includes 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 those functions 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. Also, 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, which provides a redundant error-reporting path in case errors are 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, starts a Call Home for service. This methodology ensures that all platform errors are reported through at least one functional path, which results 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 data that is collected automatically at the time of a failure or manually later. Although the data that is collected depends on the invocation method, it includes information, such as firmware levels, operating system levels, other fault isolation register values, recoverable error threshold register values, and system status.

The data is formatted and prepared for transmission back to IBM to assist the service support organization with preparing a service action plan for the IBM SSR or for more 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 can be offloaded to the optional HMC. Specific management console information is included as part of the information that can be sent to IBM Support for analysis. If more 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 on the FSP or in the operating system, depending on the type of memory dump that was started and whether the operating system is operational.

4.6.6 Notifying

After a IBM Power Systems server detects, diagnoses, and reports an error to an appropriate aggregation point, it notifies 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 replacement part.

Client Notify

When an event is important enough to report but does not indicate the need for a repair action or 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, including 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. The use of 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 147.

Vital product data and inventory management

IBM 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, which enables the IBM SSRs to help keep 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 that require 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 cannot 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. 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. The IBM SSR also can be guided 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 visible only while the system is in standby power. There are two gold caps implemented. The gold cap is used to illuminate the amber LEDs after power is removed from the system. One cap is inside the drawer to identify DIMMs, processors, and VRMs. The other cap is in the RAID assembly.

IBM Knowledge Center

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 on the Internet; however, a CD-ROM based version also is available.

The purpose of IBM Knowledge Center is to provide client-related product information and softcopy information to diagnose and fix any problems that might occur with the system. Because the information is electronically maintained, updates or new capabilities can be used by service representatives immediately.

The IBM Knowledge Center is available this website:

<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.

The following service labels are available:

- ▶ **Location diagrams**

These 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**

These labels contain procedures that often are 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) that describe how to remove and replace certain serviceable hardware components.

- ▶ **Numbered arrows**

These 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 mechanisms can engage or disengage. Arrows often improve the ease of serviceability.

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, which indicate advancement through the system power-on and initialization processes. The operator panel also is used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, which enable 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 concurrently maintainable when properly configured. Concurrent maintenance is facilitated by the redundant design for the power supplies, fans, and physical storage.

In addition to these 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 user engineers and IBM SSR providers who are familiar with the task and those engineers and providers 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.

The following primary service interfaces are used, depending on the state of the system and its operating environment:

- ▶ Light Path (see “Light Path” on page 137 and “Service labels” on page 137)
- ▶ 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 includes 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 node's state. The system node 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, which ensures the connection to the management console for manageability purposes and for accepting ASMI Secure Sockets Layer (SSL) network connections. The service processor can view and manage the machine-wide settings by using the ASMI. It also 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 points:

- ▶ 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.
- ▶ The Ethernet ports have the following default IP address:
 - 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 with which you manage the operation of the server, such as auto-power restart. You also can 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, a standard Ethernet cable or a crossed cable) or through an Ethernet network. ASMI can also be accessed from an ASCII terminal, but this option 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 use the service processor's default settings to operate your server. If the default settings are used, accessing the ASMI is not necessary. To access ASMI, use one of the following methods:

- **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)**.

- **Web browser**

At the time of this 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** option, 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**

On the HMC or when properly configured on a remote system, the ASMI web interface can be started from the HMC command line. Open a window on the HMC or access the HMC with a terminal emulation and run the following command:

```
asmmenu --ip <ip address>
```

On the HMC, 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.

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 the following ways:

- By using the normal operational front view
- By pulling it out to access the switches and viewing the LCD display

The operator panel includes the following 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).

When installed, online diagnostic tests 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 can access 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 following modes are available:

- ▶ 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 more actions by the user or a diagnostic application before testing can be done.

- ▶ Maintenance mode

This mode enables checking 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 an IBM 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 (**PWRDWSYS**) 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 review 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 often 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 errors are repeated errors and consolidating them into one error.

In addition, you can use the Manage Serviceable Events task to start 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 IBM Power Systems and its associated I/O adapters.

Firmware entitlement

With the HMC Version V8R8.1.0.0, 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 installed in the machine by manufacturing. The key is valid for the regular warranty period plus some extra time.

The IBM 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. The following 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, 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 following IBM CoD Home website:

<http://www-912.ibm.com/pod/pod>

For more information about IBM entitled Software Support, see this 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 are 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 a release level.

The management console is used for system firmware updates. By using the management console, you can use 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 specific 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 was available for approximately one 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 features a new function that is needed in the environment.
- ▶ A scheduled maintenance action causes a platform reboot, which also provides an opportunity to upgrade to a new firmware release.

Updating and upgrading 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 described in the firmware section of Fix Central, which is available at the following website:

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

You also might 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 from the HMC or a running partition.

IBM Power Systems includes 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 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 following POWER8 section on the IBM Support Portal:

http://www.ibm.com/support/knowledgecenter/TI0003M/p8ei8/p8ei8_update_firmware.htm

For POWER8 based IBM 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 on the temporary side of system firmware. It also is installed into memory after the managed system is powered off and then powered on.
- ▶ **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.

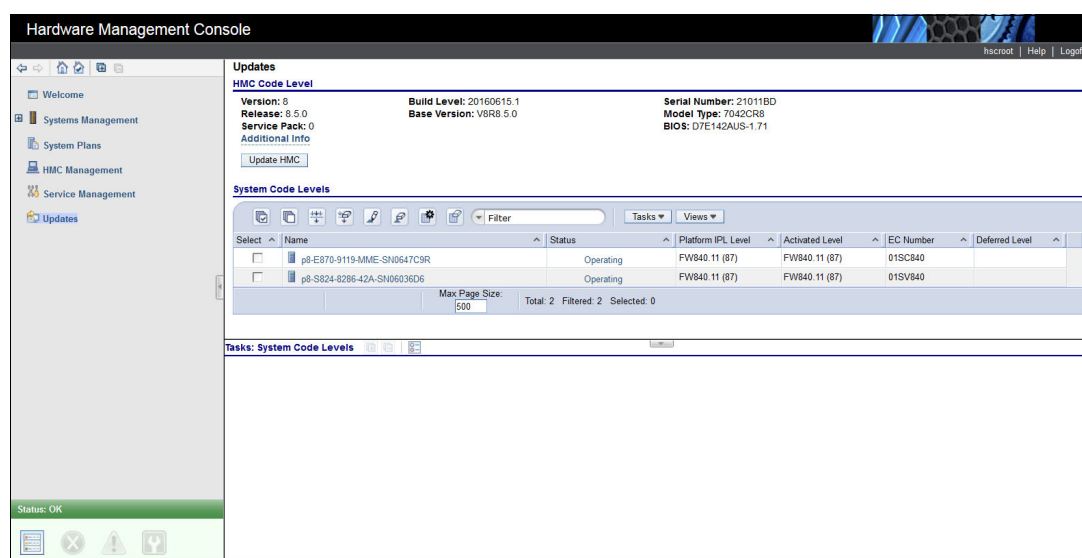


Figure 4-4 HMC System Firmware window

IBM provides the CFM function on the Power E870C and Power E880C 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. Any deferred fixes 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.

The file-naming convention for system firmware is listed in Table 4-1.

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. However, a patch in this area required a system reboot for activation because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during operation.

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 affect to any applications that are 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 extra 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 the following 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. The installation and use of ESA for problem reporting enables IBM to provide better support and service for your IBM server.

For more information about how Electronic Services can work for you, see this website (an IBM ID is required):

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

Electronic Services features the following benefits:

- ▶ 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 benefit 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 secure in monitoring, reporting, and storing the data at IBM. The ESA tool securely transmits through the Internet (HTTPS or VPN) or modem. It 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, which decreases 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 following Electronic Support website:

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

My Systems provides valuable reports of installed hardware and software by 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, which provides advanced search of the technical support knowledge base. By 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/electronicssupport>

Service Event Manager

The Service Event Manager allows the user to decide which of the Serviceable Events are called home with the ESA. Certain events can be locked. 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.

Consider the following points:

- 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`

The basic configuration of the SEM can be done by using the HMC GUI. After you select the Service Event Manager (see Figure 4-5), you must add the HMC console.

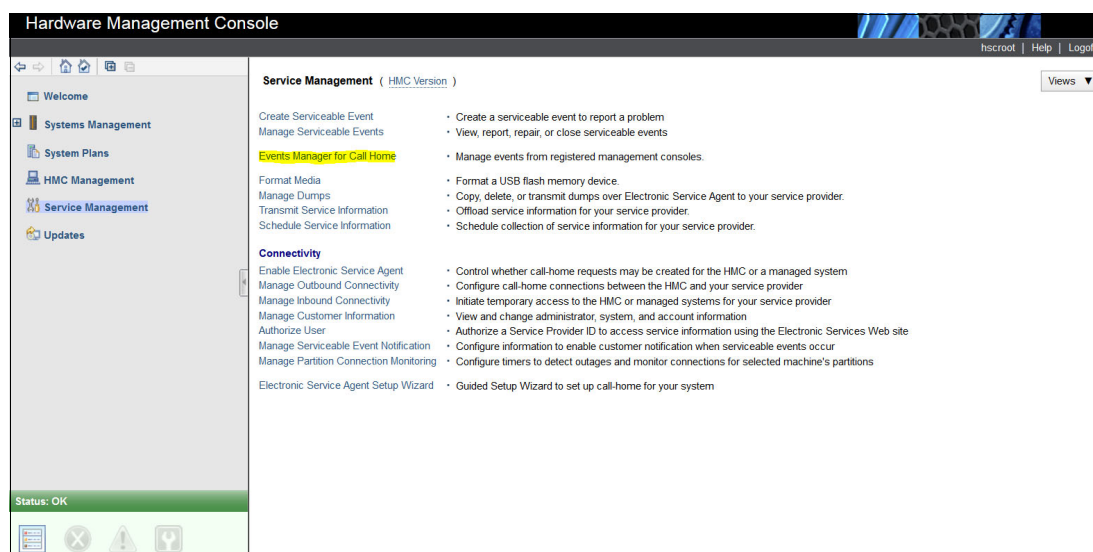


Figure 4-5 HMC selection for Service Event Manager

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 Manage Consoles

Event Criteria

Approval state

Status

Originating HMC

Refresh

Serviceable Events

Last Reported Time	Call Home Intended	Approval State	PMH #	Status	Failing MTMS	Reference Code	Originating HMC	Problem #	
No items to display									

[Learn more →](#) OK Cancel

Figure 4-6 Initial SEM window

The following configurable options are available:

- ▶ **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 the **All consoles** option to filter the list.
- ▶ **Serviceable Events**
The Serviceable Events table shows the list of events that are 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 yet approved.

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

The IBM Power Systems RAS capabilities are listed by operating system in Table 4-2. The HMC is an optional feature on scale-out IBM Power Systems servers.

Table 4-2 Selected RAS features by operating system

RAS feature	AIX V7.2 TL0 V7.1 TL4 V7.1 TL3 SP4 V6.1 TL9 SP3	IBM i V7R1 TR10 V7R2 TR4 V7R3	Linux RHEL6.5 RHEL7.1 SLES11SP3 SLES12 Ubuntu 16.04
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 V7.2 TL0 V7.1 TL4 V7.1 TL3 SP4 V6.1 TL9 SP3	IBM i V7R1 TR10 V7R2 TR4 V7R3	Linux RHEL6.5 RHEL7.1 SLES11SP3 SLES12 Ubuntu 16.04
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, with limited support in a PowerKVM environment

c. For POWER8 systems, IBM i requires IBM i 7.1 TR9 and IBM i 7.2 TR1, or later.



Related publications

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

IBM Redbooks

The following IBM Redbooks publications provide more information about the topic in this document. Note that some publications that are referenced in this list might be available in softcopy only:

- ▶ *IBM Power Systems S812L and S822L Technical Overview and Introduction*, REDP-5098
- ▶ *IBM Power Systems S812LC Technical Overview and Introduction*, REDP-5284
- ▶ *IBM Power Systems S821LC Technical Overview and Introduction*, REDP-5406
- ▶ *IBM Power System S822 Technical Overview and Introduction*, REDP-5102
- ▶ *IBM Power Systems S822LC Technical Overview and Introduction*, REDP-5283
- ▶ *IBM Power System S822LC for Big Data Technical Overview and Introduction*, REDP-5407
- ▶ *IBM Power Systems S822LC for High Performance Computing Technical Overview and Introduction*, REDP-5405
- ▶ *IBM Power Systems S814 and S824 Technical Overview and Introduction*, REDP-5097
- ▶ *IBM Power Systems E850 Technical Overview and Introduction*, REDP-5222
- ▶ *IBM Power Systems E850C Technical Overview and Introduction*, REDP-5413
- ▶ *IBM Power Systems E870 and E880 Technical Overview and Introduction*, REDP-5137
- ▶ *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491
- ▶ *IBM Power Systems SR-IOV Technical Overview and Introduction*, REDP-5065
- ▶ *IBM PowerVC Version 1.3.1 Introduction and Configuration*, SG24-8199
- ▶ *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

The following websites are also relevant as further information sources:

- ▶ *Active Memory Expansion: Overview and Usage Guide* documentation:
<http://www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=POW03037USEN>
- ▶ *IBM EnergyScale for POWER8 Processor-Based Systems* white paper:
<http://public.dhe.ibm.com/common/ssi/ecm/po/en/pow03125usen/POW03125USEN.PDF>
- ▶ IBM POWER8 systems facts and features
<http://www.ibm.com/systems/power/hardware/reports/factsfeatures.html>
- ▶ IBM Power Systems S812L server specifications
<http://www.ibm.com/systems/power/hardware/s812l-s822l/index.html>
- ▶ IBM Power Systems S814 server specifications
<http://www.ibm.com/systems/power/hardware/s814/index.html>
- ▶ IBM Power Systems S821LC server specifications
<http://www.ibm.com/systems/power/hardware/s821lc/index.html>
- ▶ IBM Power Systems S822 server specifications
<http://www.ibm.com/systems/power/hardware/s822/index.html>
- ▶ IBM Power Systems S822L server specifications
<http://www.ibm.com/systems/power/hardware/s812l-s822l/index.html>
- ▶ IBM Power Systems S822LC for Big Data server specifications
<http://www.ibm.com/systems/power/hardware/s822lc-big-data/index.html>
- ▶ IBM Power System S822LC for Commercial Computing server specifications
<http://www.ibm.com/systems/power/hardware/s822lc-commercial/index.html>
- ▶ IBM Power Systems S822LC for High Performance Computing server specifications
<http://www.ibm.com/systems/power/hardware/s822lc-hpc/index.html>
- ▶ IBM Power Systems S824 server specifications
<http://www.ibm.com/systems/power/hardware/s824/index.html>
- ▶ IBM Power Systems S824L server specifications:
<http://www.ibm.com/systems/power/hardware/s824l/index.html>
- ▶ IBM Power Systems E850 server specifications:
<http://www.ibm.com/systems/power/hardware/e850/index.html>

- ▶ IBM Power Systems E870 server specifications:
<http://www.ibm.com/systems/power/hardware/e870/index.html>
- ▶ IBM Power Systems E870C server specifications:
<http://www.ibm.com/systems/power/hardware/enterprise-cloud/index.html>
- ▶ IBM Power Systems E880 server specifications:
<http://www.ibm.com/systems/power/hardware/e880/index.html>
- ▶ IBM Power Systems E880C server specifications:
<http://www.ibm.com/systems/power/hardware/enterprise-cloud/index.html>
- ▶ *POWER8 Processor-Based Systems RAS - Introduction to Power Systems Reliability, Availability, and Serviceability*:
<https://ibm.biz/BdsRu4>
- ▶ IBM Fix Central website:
<http://www.ibm.com/support/fixcentral/>
- ▶ IBM Knowledge Center:
<http://www.ibm.com/support/knowledgecenter/>
- ▶ IBM Power Systems website:
<http://www.ibm.com/systems/power/>
- ▶ IBM Power Systems Hardware IBM Knowledge Center:
<http://www.ibm.com/support/knowledgecenter/api/redirect/powersys/v3r1m5/index.jsp>
- ▶ IBM Storage website:
<http://www.ibm.com/systems/storage/>
- ▶ IBM System Planning Tool website:
<http://www.ibm.com/systems/support/tools/systemplanningtool/>
- ▶ Chef-client for AIX is now enhanced with new recipes and the AIX cookbook:
<https://supermarket.chef.io/cookbooks/aix/>
- ▶ For more information on Yum and Cloud-init, see the AIX Toolbox for Linux Applications at:
<http://www.ibm.com/systems/power/software/aix/linux/toolbox/alpha.html>
- ▶ Open source projects for AIX can be found at the following repository:
<http://github.org/aioxss>
- ▶ Node.js is available for Linux and AIX platforms and can be downloaded from:
<https://nodejs.org/en/download/>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



REDP-5413-00

ISBN 0738455636

Printed in U.S.A.

Get connected

