Redbooks
ibm.com/redbooks

IBM Power System E850
Technical Overview and Introduction

Volker Haug

Andrew Laidlaw

Seulgi Yoppy Sung

Power Systems

Redpaper

International Technical Support Organization

IBM Power System E850: Technical Overview and Introduction

June 2015

REDP-5222-00

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

First Edition (June 2015)

This edition applies to the IBM Power System E850 (8408-E8E).

This document was created or updated on April 24, 2017.

Contents

Notices

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

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

The following terms are trademarks of other companies:

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

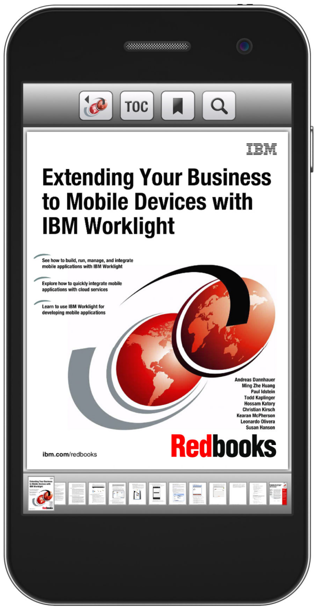
Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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

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

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

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


TOC
IBM
Extending Your Business to Mobile Devices with IBM Worklight
Redbooks
ibm.com/redbooks

It's good to be noticed.

THIS PAGE INTENTIONALLY LEFT BLANK

Preface

This IBM® Redpaper™ publication is a comprehensive guide covering the IBM Power System E850 (8408-E8E) server that supports IBM AIX®, and Linux operating systems. The objective of this paper is to introduce the major innovative Power E850 offerings and their relevant functions:

- The new IBM POWER8™ processor, available at frequencies of 3.02 GHz, 3.35 GHz, and 3.72 GHz
- Significantly strengthened cores and larger caches
- Two integrated memory controllers with improved latency and bandwidth
- Integrated I/O subsystem and hot-pluggable PCIe Gen3 I/O slots
- I/O drawer expansion options offer greater flexibility
- Improved reliability, serviceability, and availability (RAS) functions
- IBM EnergyScale™ technology that provides features such as power trending, power-saving, capping of power, and thermal measurement

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

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

This paper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power E850 system.

This paper does not replace the latest marketing materials and configuration tools. It is intended as an additional source of information that, together with existing sources, 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 IBM International Technical Support Organization (ITSO), Poughkeepsie Center.

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

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

Seulgi Yoppy Sung is a passionate Engineer, supporting multi-platform systems as a System Services Representative for almost three years, including Power System hardware, AIX, high-end and low-end storage DS8000®, and V7000. She is very positive, enthusiastic, and enjoys networking with the POWER8 community.

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, Dean Barclay, Tamikia Barrow, T R Bosworth, David Bruce, Gareth Coates, Tim Damron, Nigel Griffiths, Daniel Henderson, Dan Hurlimann, Roxette Johnson, Carolyn K Jones, Patricia Mackall, Azucena Guadalupe Maldonado Ruano, Michael J Mueller, Brett Murphy, Thoi Nguyen, Mark Olson, Kanisha Patel, Amartey Pearson, Mike Pearson, Audrey Romonosky, Nicole Schwartz, Doug Szerdi, Kristin Thomas, Gerlinde Wochele, Doug Yakesch
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:

  https://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

General description

The IBM Power System E850 (8408-E8E) server uses the latest POWER8 processor technology that is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads.

The Power E850 is optimized for running AIX and Linux workloads. The Power E850 server is four EIA-units tall (4U).

The Power E850 server supports the IBM Solution Editions for the SAP HANA offering.

1.1 Systems overview

The following sections provide detailed information about the Power E850 system.

1.1.1 Power E850 server

The Power System E850 server offers new levels of performance, price/performance, and function with POWER8 technology using a 4-socket server. The Power E850 server impressively provides up to 48 POWER8 cores, up to 4 TB of DDR3 memory on a 4-socket server, up to 51 Gen3 PCIe slots, integrated SAS controllers for integrated disk/SSD/DVD bays, and up to over 1500 SAS bays for disk/SSD with EXP24S I/O drawers.

The Power E850 server is a single enclosure server that uses IBM POWER8 processor technology. It can be configured with two, three, or four processor modules with one machine type-model (8408-E8E).

The minimum processor configuration is two processor modules, however with this configuration, only 7 of the 11 internal PCIe slots function. With three processor modules configured, 9 of the 11 PCIe slots function; and with four processor modules configured, all 11 PCIe slots are available for use.

The processor and memory subsystem in the Power E850 server is contained on a single planar. This CPU planar contains the 4 processor sockets and 32 memory Custom DIMM (CDIMM) slots. The minimum system configuration supported is two processor modules and eight CDIMMs, the rules covering the general offering is a minimum of four CDIMMs per processor socket populated.

The I/O subsystem in Power E850 consists of an IO planar, storage controller card, and a storage backplane. The IO planar supports 10 general-purpose PCI-E Gen3 slots for full high/half length adapters. An additional (11th) slot is defaulted to the LAN adapter. Eight of the 11 slots are Coherent Accelerator Processor Interface (CAPI)-capable. The storage controller card contains two storage controllers that drive four SFF (2.5") and eight 1.8" bays.

There are three versions of the storage controller card:

- High performance dual controller with write cache
- Dual controller without write cache
- Split disk backplane

Up to 11 PCIe Gen3 slots are provided in the system unit. The number of PCIe slots depends on the number of processor modules. Two processor modules provide seven PCIe slots. Three processor modules provide nine PCIe slots. Four processor modules provide 11 PCIe slots. One of the slots must contain an Ethernet LAN adapter. Up to 40 additional PCIe Gen3 slots can be added using PCIe Gen3 I/O Expansion Drawers. A two-processor module system can attach up to two PCIe Gen3 I/O drawers and provide up to 27 PCIe slots. A three-processor module system can attach up to three PCIe Gen3 I/O drawers and provide up to 39 PCIe slots. A four-processor module system can attach up to four PCIe Gen3 I/O drawers and provide up to 51 PCIe slots.

In addition to extensive hardware configuration flexibility, the Power E850 server offers elastic Capacity on Demand for both cores and memory, IBM Active Memory™ Expansion, and Active Memory Mirroring for Hypervisor. The Power E850 supports AIX and Linux operating systems and PowerVM Standard and Enterprise editions.

The Power E850 was designed and built to provide strong reliability, availability, and serviceability characteristics. These include POWER8 chip capabilities, memory protection, multiple SAS storage protection options, hot plug SAS bays, hot plug PCIe slots, redundant and hot plug power supplies and cooling fans, redundant and spare internal cooling fans, hotplug Time of Day battery, and even highly resilient architecture for power regulators.

The Power E850 server supports the IBM Solution Editions for SAP HANA offering.

Figure 1-1 and Figure 1-2 show the Power E850 with front and rear view.

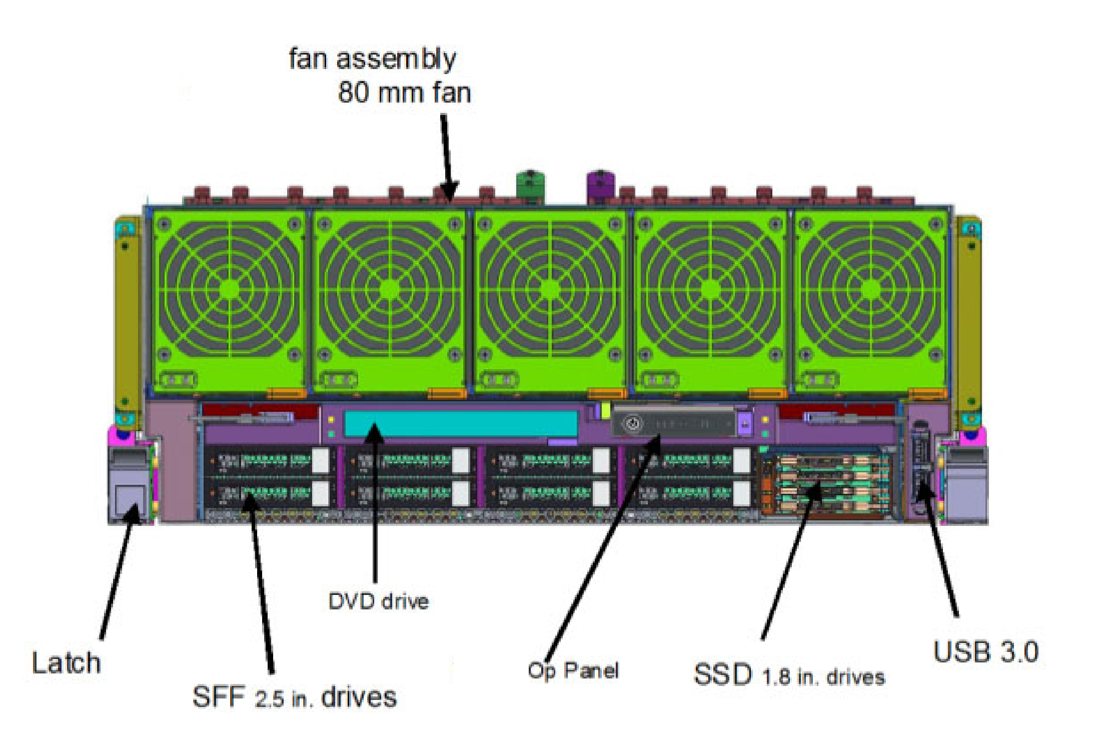


Figure 1-1 Front view of Power E850 server

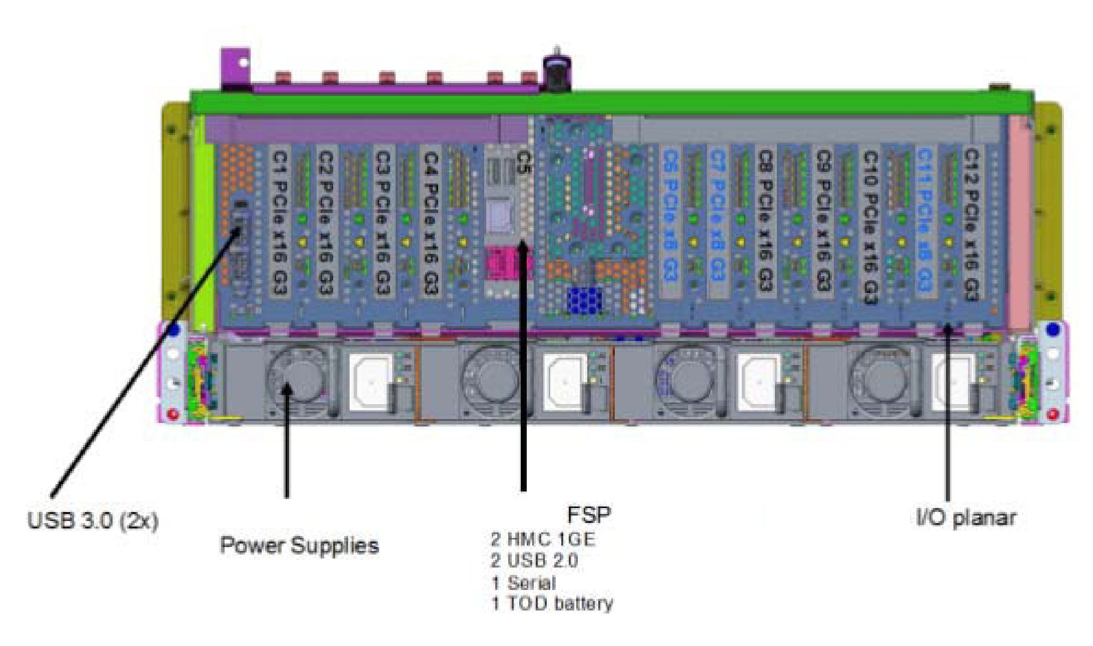


Figure 1-2 Rear view of Power E850 server

1.2 Operating environment

Table 1-1 details the operating environment for the Power E850 server.

Table 1-1 Operating environment for Power E850

Power E850 operating environment		
System	**Power E850**	
	Operating	**Non-operating**
Temperature	Recommended 18 - 27 degrees C (64 - 80 F)	1 - 60 degrees C (34 - 140 F)
	Allowable 5 - 40 degrees C (41 - 104 F)	
Relative humidity	8 - 80%	5 - 80%
Maximum dew point	24 degrees C (80 F)	27 degrees C (75 F)
Operating voltage	200 - 240 V AC 180 - 400 V DC	N/A
Operating frequency	50 - 60 Hz +/- 3 Hz AC	N/A
Maximum power consumption	3500 Watts	N/A
Maximum power source loading	3.57 kVA	N/A
Maximum thermal output	11,940 BTU/hour	N/A
Maximum altitude	3050 m (10,000 ft.)	N/A

Power E850 operating environment		
System	**Power E850**	
Noise level	Declared A-weighted sound power level, LWad (B) [1]	Declared A-weighted sound power level, LWad (B) [1]
	Operating	Heavy workload
Model 8408-E8E with two socket configured.	7.2	N/A
Model 8408-E8E with four socket configured.	7.4	8.1

Notes:

1. Declared level LWad is the upper-limit A-weighted sound power level.

   Notice: Government regulations (such as those prescribed by OSHA or European Community Directives) may govern noise level exposure in the workplace and may apply to you and your server installation. This IBM system is available with an optional acoustical door feature that can help reduce the noise emitted from this system. The actual sound pressure levels in your installation depend upon various factors, including the number of racks in the installation; the size, materials, and configuration of the room where you designate the racks to be installed; the noise levels from other equipment; the room ambient temperature; and employees' location in relation to the equipment. Further, compliance with such government regulations also depends upon various additional factors, including the duration of employees' exposure and whether employees wear hearing protection. IBM recommends that you consult with qualified experts in this field to determine whether you are in compliance with the applicable regulations.

   A 1.0 B increase in the LWad for a product equals a sound that is approximately twice as loud or twice as noisy.

   The Power E850 server must be installed in a rack with a rear door and side panels for EMC compliance. The native HMC Ethernet ports must use shielded Ethernet cables.

1.3 Physical package

Table 1-2 lists the physical dimensions of individual system nodes and of the PCIe I/O expansion drawer.

The system node requires 4U and PCIe I/O expansion drawer requires 4U. Thus, a single-enclosure system with one PCIe I/O expansion drawer requires 8U.

Table 1-2 Physical dimensions of the Power E850 and the PCIe I/O expansion drawer

Dimension	Power E850 system node	PCIe I/O expansion drawer
Width	448 mm (17.6 in.)	482 mm (19 in.)
Depth	776 mm (30.6 in.)	802 mm (31.6 in.)
Height	175 mm (6.9 in.) 4EIA units	175 mm (6.9 in.) 4 EIA units
Weight	69 kg (152 lb)	54.4 kg (120 lb)

To assure installability and serviceability in non IBM industry-standard racks, review the installation planning information for any product-specific installation requirements.

Figure 1-3 shows a picture of a Power E850 system node from the front.

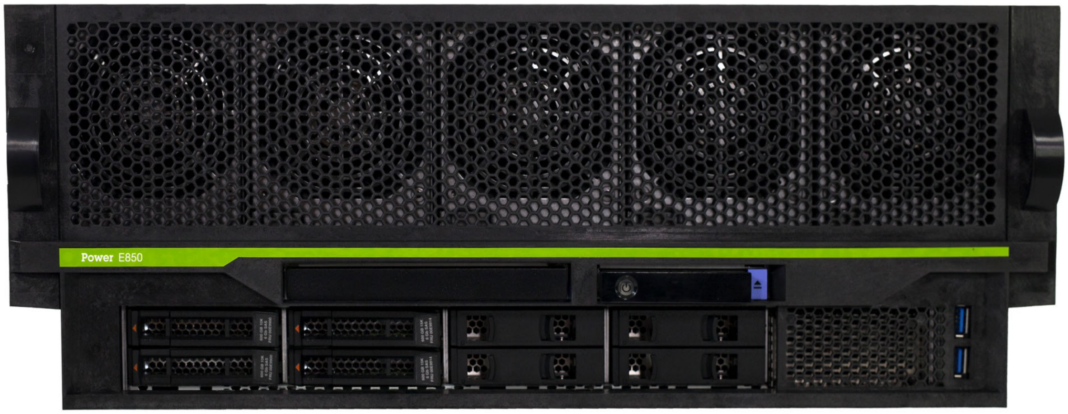


Figure 1-3 Front view of the Power 850 system node

1.4 System features

The Power E850 system nodes contain two to four processor modules with 512 KB L2 cache and 8 MB L3 cache per core, and L4 cache with DDR3 memory.

1.4.1 Power E850 system features

The following features are available on the Power E850:

- The Power E850 server supports 16 - 48 processor cores with two, three, or four POWER8 processor modules:
  - 8-core 3.72 GHz Processor Module (#EPV2)
  - 10-core 3.35 GHz Processor Module (#EPV6)
  - 12-core 3.02 GHz Processor Module (#EPV4)
- 128 GB to 4 TB high-performance DDR3 memory with L4 cache:
  - 16 GB CDIMM Memory (#EM86)
  - 32 GB CDIMM Memory (#EM87)
  - 64 GB CDIMM Memory (#EM88)
  - 128 GB CDIMM Memort (#EM8S)
- Optional Active Memory Expansion (#4798)
- Choice of three storage backplane features with different integrated SAS RAID controller options. All backplane options have eight SFF-3 SAS bays, four 1.8-inch SSD bays, and one DVD bay. All backplane options offer IBM Easy Tier® function for mixed HDD/SSD arrays:
  - Dual SAS Controller Backplane, without write cache (#EPVP)
  - Dual SAS Controller Backplane, with write cache (#EPVN)

  - Split Disk Backplane (two single SAS controllers), without write cache (#EPVQ)
- Up to 11 hot-swap PCIe Gen3 slots in the system unit:
  - Three x8 Gen3 full height, half length adapter slots.
  - Four, six, or eight x16 Gen3 full-height, half-length adapter slots.
  - With two processor modules, there are seven PCIe slots; with three modules, there are nine PCIe slots; and with four modules, there are 11 PCIe slots in the system unit.
  - One x8 PCIe slot is used for a LAN adapter.
- The PCIe Gen3 I/O Expansion Drawer (#EMX0) expands the number of full-high, hot-swap Gen3 slots:
  - Up to two PCIe3 drawers with two processor modules (maximum 27 slots on the server)
  - Up to three PCIe3 drawers with three processor modules (maximum 39 slots on the server)
  - Up to four PCIe3 drawers with four processor modules (maximum 51 slots on the server)
- Up to 64 EXP24S SFF Gen2-bay Drawers (#5887) can be attached, providing up to 1536 SAS bays for disk or SSD.
- System unit I/O (integrated I/O)
  - HMC ports: Two 1 GbE RJ45
  - USB ports: Four USB 3.0 (two front and two rear) for general use and USB 2.0 (rear) for limited use
  - System (serial) port: One RJ45
- Hot-plug, redundant power supplies
  - 4-AC Power Supply 1400 W (200-240 V) (#EB2M)
  - 4-DC Power Supply 1400 W (180-400 V) (#EB2N)
- The Power E850 has been designed and built to provide strong reliability, availability, and serviceability characteristics. These include POWER8 chip capabilities, memory protection, multiple SAS storage protection options, hot-plug SAS bays, hot-plug PCIe slots, redundant and hot-plug power supplies and cooling fans, and even highly resilient architecture for power regulators.
- Other features:
  - Second-generation service processor
  - Hot-plug SAS bays and PCIe slots
  - Redundant and hot-plug power supplies and fans
  - Optional Active Memory Mirroring for Hypervisor (#EM81)
  - Highly resilient power regulator architecture
- System unit only 4U in a 19-inch rack-mount hardware
- Primary operating systems:
  - AIX (#2146) (small tier licensing)
  - Linux (#2147): RHEL, SLES, and Ubuntu

1.4.2 Minimum configuration

The minimum Power E850 initial order must include two processor modules, 128 GB of memory, a storage backplane, one HDD or SSD DASD device, a PCIe2 4-port 1 GbE adapter, four power supplies and power cords, an operating system indicator, and a Language Group Specify.

Table 1-3 shows the minimum defined initial order configuration.

Table 1-3 Initial order configuration of Power E850

Feature number	Description
#EPV2 x 2	8-core 3.72 GHz POWER8 Processor module
and	
#EPV9 x 16	Processor core entitlement activations for #EPV2 to equal total processor cores ordered
#EM83 x 8 #EMAA x 128	16 GB CDIMM, 1600 MHz DDR3 Memory and memory activation features. All memory ordered in pairs
#EPVQ x 1	Split Storage Backplane without write cache
#ESDB x 1	300 GB 15 K RPM SAS SFF-3 Disk Drive
or	
#0837 x 1	SAN Load Source Specify
#5899 x 1	PCIe2 4-port 1 GbE Adapter
#EB2M (or EB2N) x 4	AC Power Supply - 1400 W for Server (200-240 V AC)
and	
#6458 x 4	Power cord 4.3 m (14 ft), Drawer to Wall/IBM PDU (250V/10A)
#9300/97xx	Language Group Specify
#2146	Primary Operating System Indicator: AIX
or	
#2147 x 1	Primary Operating System Indicator: Linux

Notes:

- A machine type/model rack, if wanted, should be ordered as the primary rack.
- A minimum number of 16 processor activations must be ordered per system.
- A minimum of four memory features per processor module is required.
- At least 50% of available memory must be activated through a combination of features EMAA and EMAB.
- The language group is auto-selected based on geographic rules.

1.4.3 Power supply features

The following are the key power supply features:

- (#EB2M) AC Power Supply - 1400 W for Server (200-240 VAC)
  - One 200 - 240 V, 1400-watt AC power supply
  - The power supply is configured in a one-plus-one or two-plus-two configuration to provide redundancy. Supported in rack models only.
  - To be operational, a minimum of four power supplies in the server are required. If there is a power supply failure, any of the power supplies can be exchanged without interrupting the operation of the system.
- (#EB2N) DC Power Supply - 1400 W for Server (180 - 400 VDC)
  - One 180 - 400 V, 1400-watt DC power supply
  - The power supply is configured in a one-plus-one or two-plus-two configuration to provide redundancy. Supported in rack models only.
  - To be operational, a minimum of four power supplies in the server are required. If there is a power supply failure, any of the power supplies can be exchanged without interrupting the operation of the system.
- (#EMXA) AC Power Supply Conduit for PCIe3 Expansion Drawer
  - Provides two 320-C14 inlet electrical connections for two separately ordered AC power cords with C13 connector plugs. Conduit provides electrical power connection between two power supplies in the front of a PCIe Gen3 I/O Expansion Drawer (#EMX0) and two power cords that connect on the rear of the PCIe Gen3 I/O Expansion Drawer.
  - A maximum of two per I/O Exp Drawer #EMX0

1.4.4 Processor card features

The following are the key processor card features:

- The Power E850 server supports 16 - 48 processor cores, as a combination of two, three, or four processor modules:
  - 8-core 3.72 GHz (#EPV2)
  - 10-core 3.35 GHz (#EPV6)
  - 12-core 3.02 GHz (#EPV4)
- A minimum of two and a maximum of four processor modules are required for each system. The third and fourth module can be added to a system later through an MES order, but requires scheduled downtime to install. All processor modules in one server must be the same frequency (same processor module feature number). They cannot be mixed.
- Permanent processor core activations are required for the first two processor modules in the configuration and are optional for the third and fourth modules. Specifically:
  - Two, three, or four 8-core, 3.72 GHz processor modules (#EPV2) require 16 processor core activations (#EPV9) at a minimum.
  - Two, three, or four 10-core, 3.35 GHz processor modules (#EPV6) require 20 processor core activations (#EPVH) at a minimum.
  - Two, three, or four 12-core, 3.02 GHz processor modules (#EPV4) require 24 processor core activations (#EPVD) at a minimum.

- Temporary Capacity on Demand (CoD) capabilities are optionally used for processor cores that are not permanently activated:
  - 90 Days Elastic CoD Processor Core Enablement (#EP9T)
  - 1 and 100 Processor Day Elastic CoD billing for #EPV2 (#EPJW, #EPJX)
  - 1 and 100 Processor Day Elastic CoD billing for #EPV4 (#EPK0, #EPK1)
  - 1 and 100 Processor Day Elastic CoD billing for #EPV6 (#EPK3, #EPK4)
  - 100 Processor-minutes Utility CoD billing: for #EPV2 (#EPJY), for #EPV4 (#EPK2), or for #EPV4 (#EPK5)
  - An HMC is required for Elastic CoD and Utility CoD

1.4.5 Summary of processor features

Sixteen to 48 processor cores on two, three, and four POWER8 processor modules on the Power E850 server are supported by 8-core 3.72 GHz Processor Module (#EPV2), 10-core 3.35 GHz Processor Module (#EPV6), and 12-core 3.02 GHz Processor Module (#EPV4).

Table 1-4 summarizes the processor feature codes for the Power E850.

Table 1-4 Summary of processor features for the Power E850

Feature Code	Description
EPVL	3.72 GHz 8-core processor and 256 GB Memory for HANA Solution
EPVM	Power Processor Activation for #EPVL
ELJK	Power IFL Processor Activation
ELJL	Power IFL Processor Activation
ELJM	Power IFL Processor Activation
EP9T	90 Days Elastic CoD Processor Core Enablement
EPJW	Processor-Day Elastic CoD Billing for #EPV2, AIX/Linux
EPJX	100 Processor-Days Elastic CoD Billing for #EPV2, AIX/Linux
EPJY	100 Processor-Minutes Utility CoD Billing for #EPV2, AIX/Linux
EPK0	1 Processor-Day Elastic CoD Billing for #EPV4, AIX/Linux
EPK1	100 Processor-Days Elastic CoD Billing for #EPV4, AIX/Linux
EPK2	100 Processor-Minutes Utility CoD Billing for #EPV4, AIX/Linux
EPK3	1 Processor-Day Elastic CoD Billing for #EPV6, AIX/Linux
EPK4	100 Processor-Days Elastic CoD Billing for #EPV6, AIX/Linux
EPK5	100 Processor-Minutes Utility CoD Billing for #EPV6, AIX/Linux
EPV2	3.72 GHz, 8-core POWER8 Processor Module
EPV4	3.02 GHz, 12-core POWER8 Processor Module
EPV6	3.35 GHz, 10-core POWER8 Processor Module
EPV9	1 Core Processor Activation for #EPV2
EPVD	1 Core Processor Activation for #EPV4
EPVH	1 Core Processor Activation for #EPV6

1.4.6 Memory features

The following are the key memory features:

- The Power E850 supports 128 GB to 2 TB high-performance 1600 MHz DDR3 ECC memory with L4 cache using the following feature codes:
  - 16 GB CDIMM Memory (#EM86)
  - 32 GB CDIMM Memory (#EM87)
  - 64 GB CDIMM Memory (#EM88)
- The Power E850 supports 1024 GB to 4 TB high performance 1600 MHz DDR4 ECC memory with L4 cache using the following feature code:
  - 128 GB CDIMM Memory (#EM8S)
- As the customer memory requirements increase, the system capabilities are increased as follows:
  - With two processor modules installed, 16 CDIMM slots are available, minimum memory is 128 GB and maximum is 2 TB.
  - With three modules, 24 CDIMM slots are available, minimum memory is 192 GB and maximum is 3 TB.
  - With four modules, 32 CDIMM slots are available, minimum memory is 256 GB and maximum is 4 TB. Four CDIMMs are available per socket. The more CDIMM slots that are filled, the larger the available bandwidth available to the server.
  - Permanent memory activations are required for at least 50% of the physically installed memory. For example, for a system with two processor modules installed, no less than 128 GB must be activated. Use 1 GB activation (#EMAA) and 100 GB activation (#EMAB) features to order permanent memory activations.
  - DDR4 CDIMMS cannot be mixed with DDR3 memory CDIMMS. If #EM8S is used, then it must be the only memory used on the server.
- Memory is ordered in pairs of the same memory feature. Both CDIMMs of a CDIMM pair must be installed in the slots supporting one processor. Different size pairs can be mixed on the same processor module's slots. For optimal performance, all pairs would be the same size. Also, for optimal performance, generally the amount of memory per processor module would be the same or about the same.
- Temporary CoD for memory is available for memory capacity that is not permanently activated:
  - 90 Days Elastic CoD Memory Enablement (#EM9T).
  - 1, 100, and 999 GB-Day billing for Elastic CoD memory (#EMJA, #EMJB, #EMJC).
  - An HMC is required for Elastic CoD.

Table 1-5 lists memory features that are available for the Power E850.

Table 1-5 Summary of memory features

Feature Code	Description
4798	Active Memory Expansion Enablement
ELJP	Power IFL Memory Activation
EM81	Active Memory Mirroring
EM86	16 GB CDIMM, 1600 MHZ, DDR3 Memory

Feature Code	Description
EM87	32 GB CDIMM, 1600 MHZ, DDR3 Memory
EM88	64 GB CDIMM, 1600 MHZ, DDR3 Memory
EM8s	128 GB CDIMM, 1600 MHZ, DDR4 Memory
EM9T	90 Days Elastic CoD Memory Enablement
EMAA	1 GB Memory Activation
EMAB	Quantity of 100 1 GB Memory Activations (#EMAA)
EMAJ	256 GB Memory Activation for #EPVL
EMB9	Five Hundred and Twelve Memory Activations for IFL
EMJA	1 GB-Day billing for Elastic CoD memory
EMJB	100 GB-Day billing for Elastic CoD memory
EMJC	999 GB-Day billing for Elastic CoD memory
EPVL	3.72 GHz 8-core processor and 256 GB Memory for HANA Solution

1.4.7 System node PCIe slots

The Power E850 server has up to 11 PCIe hot-plug Gen3 slots, providing excellent configuration flexibility and expandability. Eight adapter slots are x16 Gen3, and three adapter slots are x8 Gen3. All adapter slots are full height, half length.

Table 1-6 shows that the number of slots supported varies by the number of processor modules.

Table 1-6 Number of I/O slots supported by number of processor modules

Processor modules	2 socket	3 socket	4 socket
x16 Gen3 slots (CAPI capable)	4	6	8
x8 Gen3 slots	3	3	3

Note:

- At least one PCIe Ethernet adapter is required on the server by IBM to ensure proper manufacture and test of the server. One of the x8 PCIe slots is used for this required adapter, which is identified as the C11 slot.
- Blind swap cassettes (BSCs) are not used for adapters in the system unit.
- All PCIe slots in the system unit are SR-IOV capable.

Statement of Direction:

IBM plans to enhance the Power 850 to support a maximum of 4 TB of memory.

1.5 Disk and media features

In this section, the key disk and media features are described.

1.5.1 SAS bays and storage backplane options

Clients have a choice of three storage features with eight SFF disk bays, four x 1.8-inch disk bays, and one DVD bay:

- The Dual Controller Disk Backplane with write cache supports RAID 0, 1, 5, 6, 10, 5T2, 6T2, and 10T2 (#EPVN). The pair of controllers handles all 12 integrated SAS bays and DVD bay.
- The Dual Controller Disk Backplane without write cache supports RAID 0, 1, 5, 6, 10, 5T2, 6T2, and 10T2 (#EPVP). The pair of controllers handles all 12 integrated SAS bays and DVD bay.
- The Split Disk Backplane (two single controllers without write cache supports RAID 0, 1, 5, 6, 10, and 10T2 (#EPVQ). This is the default configuration in e-config. Each one of the two controllers handles four SFF-3 and two 1.8-inch SAS bays. One of the controllers handles the DVD bay.

Each of the three backplane options provides SFF-3 SAS bays in the system unit. These 2.5-inch or small form factor (SFF) SAS bays can contain SAS drives (HDD or SSD) mounted on a Gen3 tray or carrier. Thus the drives are designated for SFF-1, or SFF-2 bays do not fit in an SFF-3 bay. All SFF-3 bays support concurrent maintenance or hot-plug capability. All three of these backplane options support HDDs or SSDs or a mixture of HDDs and SSDs in the SFF-3 bays. If mixing HDDs and SSDs, they must be in separate arrays (unless using the Easy Tier function).

For more information about disk controller options, see 1.6, “I/O drawers” on page 16.

1.5.2 Storage Backplane Integrated Easy Tier function

The Easy Tier function is provided with all three storage backplanes (#EPVN, #EPVP, #EPVQ). Conceptually, this function is like the Easy Tier function found in the IBM Storage products such as the DS8000, IBM Storwize® V7000, or SAN Volume Controller, but implemented just within the integrated Power Systems SAS controllers and the integrated SAS bays. Hot data is automatically moved to SSDs, and cold data is automatically moved to disk (HDD) in an AIX, Linux, or VIOS environment. No user application coding is required.

Clients commonly have this hot/cold characteristic for their data. It is typical for 10% - 20% of the data to be accessed 80% - 90% of the time. This is called the *hot data*. If you can get the hot data onto SSDs, it can dramatically improve the performance of I/O-bound applications. And by keeping the cold data on HDDs, the total cost per gigabyte of the solution can be minimized. You can end up with high I/O performance at a reasonable price. By avoiding the need for lots of HDD arms for performance, you can reduce the number of I/O drawers, maintenance, rack/floor space, and energy.

For more information about Easy Tier, see 1.6, “I/O drawers” on page 16.

1.5.3 DVD and boot devices

A slimline media bay that can hold a SATA DVD-RAM is included in the feature EPVN, EPVP, and EPVQ backplanes. The DVD drive is run by one of the integrated SAS controllers in the storage backplane, and a separate PCIe controller is not required.

1.5.4 Storage features

Table 1-7 shows storage features supported for Power E850 server, and the EXP24S drawer.

Table 1-7 Supported storage features supported for Power E850 and the EXP24S drawer

Description	Feature	CCIN
Storage features supported internal to server:		
Hard disk drives		
600 GB 10K RPM SAS SFF-3 Disk Drive AIX/Linux	ESD5	
1.2 TB 10K RPM SAS SFF-3 Disk Drive AIX/Linux	ESD9	
300 GB 15K RPM SAS SFF-3 Disk Drive AIX/Linux	ESDB	
600 GB 15K RPM SAS SFF-3 Disk Drive - 5xx Block AIX/Linux	ESDF	
300 GB 10K RPM SAS SFF-3 Disk Drive AIX/Linux	ESDR	
146 GB 15K RPM SAS SFF-3 Disk Drive AIX/Linux	ESDT	
600 GB 10K RPM 4K SAS SFF-3 Disk Drive AIX/Linux	ESF5	
300 GB 15K RPM SAS SFF-3 4K Block - 4096 Disk Drive	ESFB	
600 GB 15K RPM SAS SFF-3 4K Block - 4096 Disk Drive	ESFF	
1.2 TB 10K RPM 4K SAS SFF-3 Disk Drive AIX/Linux	ESF9	
1.8 TB 10K RPM 4K SAS SFF-3 Disk Drive AIX/Linux	ESFV	
Solid-state devices (SFF)		
775 GB SFF-3 SSD for AIX/Linux	ESON	
387 GB SFF-3 4k SSD AIX/Linux	ES0U	
387 GB SFF-3 SSD for AIX/Linux	ES0L	
775 GB SFF-3 4k SSD for AIX/Linux	ES0W	
387 GB SFF-3 SSD 5xx eMLC4 for AIX/Linux	ES7K	
775 GB SFF-3 SSD 5xx eMLC4 for AIX/Linux	ES7P	
1.9 TB Read Intensive SAS 4k SFF-3 SSD for AIX/Linux	ES8J	
387 GB SFF-3 SSD 4k eMLC4 for AIX/Linux	ES8N	
775 GB SFF-3 SSD 4k eMLC4 for AIX/Linux	ES8Q	
1.55TB SFF-3 SSD 4k eMLC4 for AIX/Linux	ES8V	
Solid-state devices (1.8 inch)		
387GB 1.8'' SAS 5xx SSD eMLC4 for AIX/Linux	**ES1C**	**5B32**

Description	Feature	CCIN
387GB 1.8'' SAS 4k SSD eMLC4 for AIX/Linux	ES2V	5B30
775GB 1.8'' SAS 5xx SSD eMLC4 for AIX/Linux	ES2X	5B33
775GB 1.8'' SAS 4k SSD eMLC4 for AIX/Linux	ES4K	5B31
Storage features supported in EXP24S expansion drawer:		
Hard disk drives		
600 GB 15K RPM SAS SFF-2 Disk Drive - 5xx Block AIX/Linux	ESDP	
600 GB 10K RPM 4K SAS SFF-2 Disk Drive AIX/Linux	ESEV	
300 GB 15K RPM SAS SFF-2 4K Block - 4096 Disk Drive	ESEZ	
1.2 TB 10K RPM 4K SAS SFF-2 Disk Drive AIX/Linux	ESF3	
600 GB 15K RPM SAS SFF-2 4K Block - 4096 Disk Drive	ESFP	
1.8 TB 10K RPM 4K SAS SFF-2 Disk Drive AIX/Linux	ESFT	
900 GB 10k RPM SAS SFF-2 Disk Drive AIX/Linux	1752	
146 GB 15k RPM SAS SFF-2 Disk Drive AIX/Linux	1917	
300 GB 10k RPM SAS SFF-2 Disk Drive AIX/Linux	1925	
300 GB 15k RPM SAS SFF-2 Disk Drive AIX/Linux	1953	
1.2 TB 10K RPM SAS SFF-2 Disk Drive AIX/Linux	ESD3	
600 GB 10k RPM SAS SFF-2 Disk Drive AIX/Linux	1964	
Solid-state devices (SFF)		
775 GB SFF-2 SSD for AIX/Linux	ES0G	
775 GB SFF-2 4k SSD for AIX/Linux	ES0S	
387 GB SFF-2 SSD for AIX/Linux	ES19	
387 GB SFF-2 SSD for AIX/Linux with eMLC	ES0C (Support only)	
387 GB SFF-2 4K SSD for AIX/Linux	ES0Q	
387 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	ES78	
775 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	ES7E	
1.9 TB Read Intensive SAS 4k SFF-2 SSD for AIX/Linux	ES80	
387 GB SFF-2 SSD 4k eMLC4 for AIX/Linux	ES85	
775 GB SFF-2 SSD 4k eMLC4 for AIX/Linux	ES8C	
1.55 TB SFF-2 SSD 4k eMLC4 for AIX/Linux	ES8F	

1.6 I/O drawers

The following sections describe the available I/O drawer options for this server.

1.6.1 PCIe Gen3 I/O expansion drawer

PCIe Gen3 I/O expansion drawers (#EMX0) can be attached to the Power E850 to expand the number of full-high, hot-swap Gen3 slots available to the server. The maximum number of PCIe Gen3 I/O drawers depends on the number of processor modules physically installed. The maximum is independent of the number of processor core activations.

- Up to two PCIe3 drawers with two processor modules
- Up to three PCIe3 drawers with three processor modules
- Up to four PCIe3 drawers with four processor modules

Each PCIe Gen3 I/O expansion drawer adds a total of 10 PCIe Gen3 adapter slots to the server's capabilities.

Figure 1-4 and Figure 1-5 show the front and rear of PCIe Gen3 I/O expansion drawer.

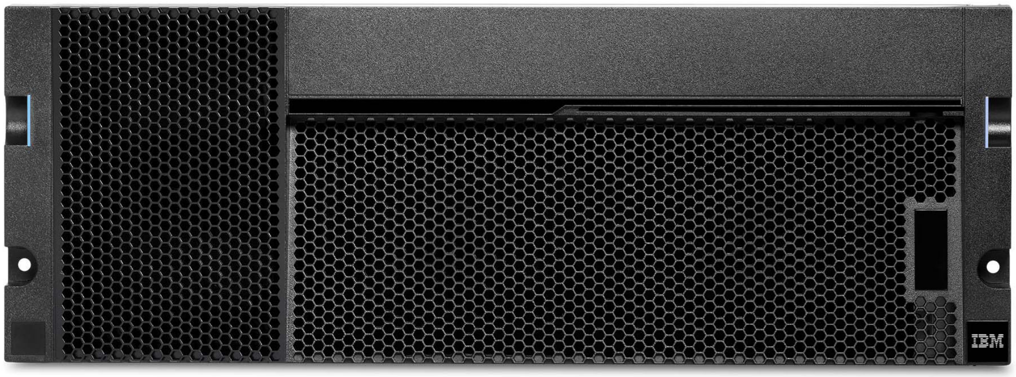


Figure 1-4 Front of the PCIe Gen3 I/O expansion drawer

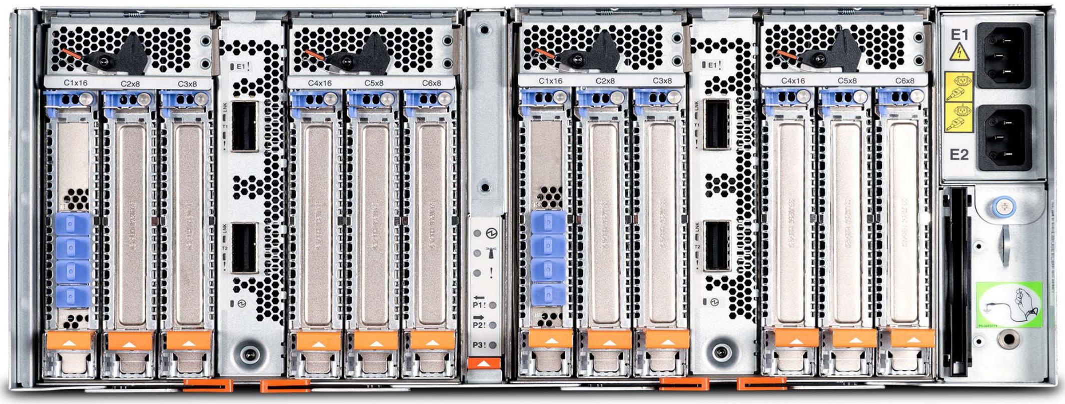


Figure 1-5 Rear of the PCIe Gen3 I/O expansion drawer

Table 1-8 shows PCIe adapters supported by the Power E850, either internally or in a PCIe Gen3 I/O expansion drawer.

Table 1-8 Feature of PCIe adapters supported internally or in a PCIe Gen3 I/O expansion drawer

Description	Feature
PCIe2 4-port 1GbE Adapter	5899
PCIe2 4-port10Gb+1GbE) SR+RJ45 Adapter	EN0S
PCIe2 4-port10Gb+1GbE) Copper SFP+RJ45 Adapter	EN0U
PCIe2 2-port 10/1GbE BaseT RJ45 Adapter	EN0W
PCIe3 4-port 10GbE SR Adapter	EN15
PCIe3 4-port 10GbE SFP+ Copper Adapter	EN17
PCIe3 2-port 10GbE NIC&RoCE SR Adapter	EC2N
PCIe3 2-port 10GbE NIC&RoCE SFP+ Copper Adapter	EC38
PCIe3 2-port 40GbE NIC RoCE QSFP+ Adapter	EC3B
PCIe3 Optical Cable Adapter for 4U CEC	EJ08
PCIe2 16Gb 2-port Fibre Channel Adapter	EN0A
PCIe2 8Gb 2-Port Fibre Channel Adapter	EN0G
PCIe2 8Gb 4-port Fibre Channel Adapter	5729
PCIe2 4-port 10GbE&1GbE SR&RJ45 Adapter	5744
10 Gigabit Ethernet-SR PCI Express Adapter	5769
PCIe2 4-port10Gb FCoE & 1GbE) SR&RJ45	EN0H
PCIe2 4-port10Gb FCoE & 1GbE) SFP+Copper&RJ45	EN0K
PCIe2 4-port(10Gb FCoE & 1GbE) LR&RJ45 Adapter	EN0M
PCIe2 4-port10Gb+1GbE) SR+RJ45 Adapter	EN0S
PCIe2 4-port10Gb+1GbE) Copper SFP+RJ45 Adapter	EN0U
PCIe2 2-port 10GbE SFN6122F Adapters	EC2J
PCIe2 2-port 10GbE RoCE SFP+ Adapter	EC28
PCIe2 2-port 10GbE SR Adapter	5287 (10Gb FCoE)
PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6Gb x8	EJ0L
PCIe3 RAID SAS Adapter Quad-port 6Gb x8	EJ0J
PCIe3 SAS Tape/DVD Adapter Quad-port 6Gb x8	EJ10
PCIe3 12GB Cache RAID PLUS SAS Adapter Quad-port 6Gb x8	57B1
PCIe 380 MB Cache Dual - x4 3Gb SAS RAID Adapter	5805
PCIe Dual-x4 SAS Adapter	5901
8 Gigabit PCI Express Dual Port Fibre Channel Adapter	5735
4 Gigabit PCI Express Dual Port Fibre Channel Adapter	5774

Description	Feature
PCIe Crypto Coprocessor No BSC 4765-001	EJ27
PCIe Crypto Coprocessor Gen3 BSC 4765-001	EJ28
4-Port 10/100/1000 Base-TX PCI Express Adapter	5717
2-Port 10/100/1000 Base-TX Ethernet PCI Express Adapter	5767
2-Port Gigabit Ethernet-SX PCI Express Adapter	5768
2 Port Async EIA-232 PCIe Adapter	EN27
4 Port Async EIA-232 PCIe Adapter	5785
PCIe2 3D Graphics Adapter x1	EC42
IBM POWER® GXT145 PCI Express Graphics Accelerator	5748
PCIe2 4-port USB 3.0 Adapter	EC46
PCIe3 1.6TB NVMe Flash Adapter	EC55
PCIe3 3.2TB NVMe Flash Adapter	EC57

Note:

- Features EC38 and EC2N are not available in the following countries: Abu Dhabi, Algeria, Bahrain, Comoros, Djibouti, Dubai, Iraq, Kuwait, Lebanon, Libya, Malaysia, Morocco, Oman, Pakistan Qatar, Saudi Arabia, Somalia, Tunisia, United Arab Emirates, and Yemen.

For more details about connecting PCIe Gen3 I/O expansion drawers to the Power E850 servers, see 2.11.1, “PCIe Gen3 I/O expansion drawer” on page 76.

1.6.2 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). Mode setting is done by IBM Manufacturing. If you need to change the mode after installation, ask your IBM support representative to refer to the following site:

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 using SAS YO or X cables.

Figure 1-6 shows the EXP24S SFF Gen2-bay drawer.

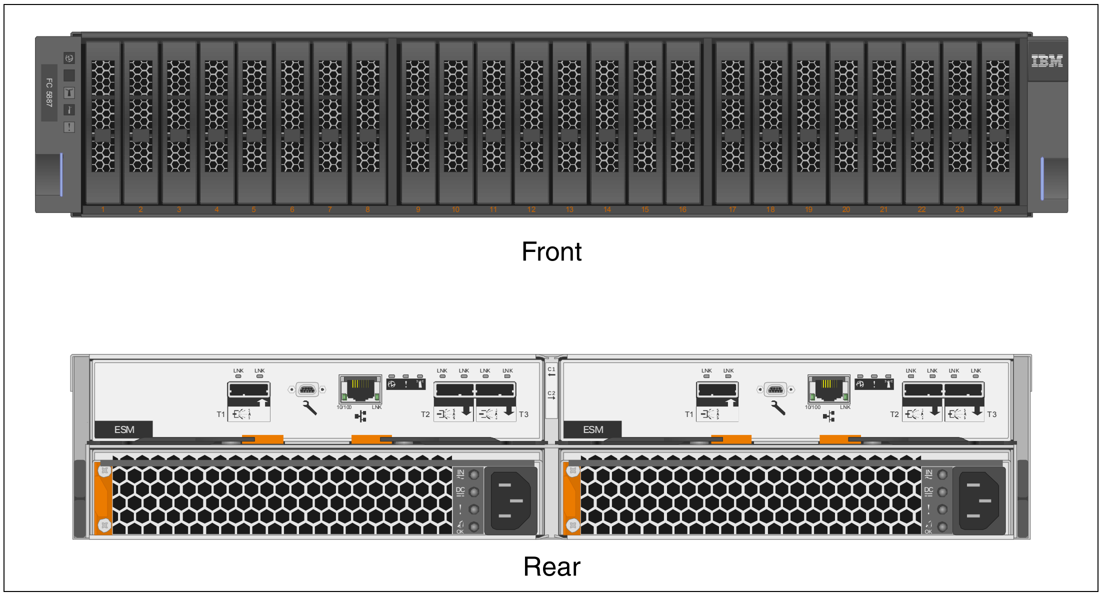


Figure 1-6 EXP24S SFF Gen2-bay drawer

For more information about connecting the EXP24S Gen2-bay drawer to the Power E850 server, see 2.12.1, “EXP24S SFF Gen2-bay Drawer” on page 82.

1.7 Build to order

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

This method is the only configuration method for the Power E850 servers.

1.8 IBM Solution Edition for SAP HANA

Power Systems Solution Editions for SAP HANA provide starting configurations that allow customers to quickly deploy HANA and grow with their business needs while meeting SAP design criteria. Power Systems servers are ideally suited to provide excellent SAP HANA performance and system reliability. The SAP HANA code running on the POWER architecture exploits the exceptional POWER8 memory bandwidth, SIMD parallelization, and simultaneous multithreading (SMT).

These solution edition offerings on POWER8 are available on the following servers.

Table 1-9 Server lists of Solution edition offerings for SAP HANA

Server	Processor	Usage
E824	24 cores at 3.52 GHz	Suitable for HANA database growing up to 0.5 TB
E850	32 cores at 3.72 GHz	Suitable for HANA database growing up to 1 TB

Server	Processor	Usage
E870	40 cores at 4.19 GHz	Suitable for HANA database starting at 0.5 TB with expected growth beyond 1 TB
E870	80 cores at 4.19 GHz	Suitable for HANA database starting at 1 TB with expected growth beyond 2 TB

Note: Effective SAP database size (containing business data) is typically half the memory capacity of the server.

Power E850 system configuration

The Power E850 initial order for SAP HANA Solution Edition with 32 cores at 3.72 GHz processor and 1 TB memory must include a minimum of the following items.

Table 1-10 Features required for an initial order for SAP HANA

Feature number	Description
EPVL x 4	3.72 GHz 8-core processor and 256 GB Memory for HANA Solution
EPVM x 32	Power Processor Activation for #EPVL
EMAJ x 4	256 GB Memory Activation for #EPVL
EM88 x 16	64 GB (4 x 16 GB) CDIMMs, 1600 MHz, 4 Gb DDR3 DRAM
EPVN x 1	Dual Controller Disk Backplane with write cache
EN0S x 2	PCIe2 4-Port (10Gb+1GbE) SR+RJ45 Adapter
5729 x 2	8 GB 4-port Fibre Channel Adapter
0837 x 1	SAN Load Source Specify
EB2M x 4	AC Power Supply - 1400W for Server (200-240 V ac)
6458 x 4	Power Cord 4.3 m (14-ft), Drawer to Wall/IBM PDU (250V/10A)
9300/97xx	Language Group Specify
2147 x 1	Primary Operating System Indicator

Note:

- The minimum requirement of SLES 11 (5639-S11) is 4-socket based licenses.
- The required SLES 11 minimum level is SP3, plus the kernel update for SAP HANA on Power.
- SAN boot is required for the SAP HANA solution running on SLES 11 SP3.
- Five units of STG Lab Services (6911-300) (#0003 x 5) is defaulted as applicable.
- Order configurations may be increased over this minimum configuration to meet customer requirements.

1.9 Management options

This section discusses the supported management interfaces for the servers.

The Power E850 platforms support two main service environments:

- Attachment to one or more HMCs. This environment is the common configuration for servers supporting logical partitions with dedicated or virtual I/O. In this case, all servers have at least one logical partition.
- No HMC. There are two service strategies for non-HMC systems:
  - Full-system partition with PowerVM: A single partition owns all the server resources and only one operating system can be installed.
  - Partitioned system with PowerVM: In this configuration, the system can have more than one partition and can be running more than one operating system. In this environment, partitions are managed by the Integrated Virtualization Manager (IVM), which provides some of the functions provided by the HMC.

Both the HMC and IVM provide the following functions necessary to manage the system:

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

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

1.9.1 HMC models

In 2015, a new HMC model was announced, machine type 7042-CR9. Hardware features on the CR9 model include a new x86 processor, dual hard drives (RAID 1), and redundant power supply. HMC code V8.3.0 is required for Power E850 servers.

> **Note:** A single HMC can manage multiple Power Systems Servers. HMCs supported on POWER8 hardware are 7042-CR5 through 7042-CR9.

The IBM Power E850 can be managed by the IVM.

Several HMC models are supported to manage POWER8 based systems.

Table 1-11 HMC models supporting POWER8 processor technology-based servers

MTM	Availability	Description
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

Note: When PowerVC is enabled, 4 GB of RAM is recommended. To use the Enhanced login mode, 8 GB is recommended. HMC 7042-CR5 ships with a default of 2 GB RAM.

HMC base Licensed Machine Code Version 8 Revision 8.3.0 or later is required to support the Power E850 (8408-E8E).

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

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

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

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

For further information about managing the Power E850 servers from an HMC, see 2.13, “Hardware Management Console” on page 93.

1.10 System racks

The Power E850 server and its I/O drawers are designed to fit a standard 19-inch rack such as IBM feature 7014-T00, 7014-T42, 7014-B42, 7965-94Y, #0553, and #0551.

However, for initial system orders, the racks must be ordered as machine type 7014-T00 36U (1.8 meter), also support #0551 for MES. The 42U (2.0-meter) rack (#0553) is available to order only on Miscellaneous Equipment Specification (MES) upgrade orders.

Notes: 7014-B42 (42U) support only, cannot be ordered from IBM and 7965-94Y (42U) support only by field merge.

Supported PDUs are either feature 7109 or 7188.

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

1.10.1 IBM 7014 model T00 rack

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

- 36U (EIA units) of usable space.
- Optional removable side panels.
- Optional highly perforated front door.
- Optional side-to-side mounting hardware for joining multiple racks.

- Standard business black or optional white color in OEM format.
- Increased power distribution and weight capacity.
- Supports both AC and DC configurations.
- The rack height is increased to 1926 mm (75.8 in.) if a power distribution panel is fixed to the top of the rack.
- The #6068 feature provides a cost effective plain front door.
- Weights are as follows:
  - T00 base empty rack: 244 kg (535 lb.)
  - T00 full rack: 816 kg (1795 lb.)
  - Maximum Weight of Drawers is 572 kg (1260 lb.)
  - Maximum Weight of Drawers in a zone 4 earthquake environment is 490 kg (1080 lb.). This equates to 13.6 kg (30 lb.)/EIA.

1.10.2 IBM 7014 model T42 rack

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

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

Notes: The Power E850 server with 7014-T42 rack includes one standard PDU.

1.10.3 IBM 7965 model 94Y rack

The 2.0-meter (79-inch) model 7965-94Y is compatible with past and present IBM Power Systems servers and provides an excellent 19-inch rack enclosure for your data center. Its 600 mm (23.6 in.) width combined with its 1100 mm (43.3 in.) depth plus its 42 EIA enclosure capacity provides great footprint efficiency for your systems and allows it to be easily on standard 24-inch floor tiles.

The IBM 42U Slim Rack has a lockable perforated front steel door, providing ventilation, physical security, and visibility of indicator lights in the installed equipment within. In the rear, either a lockable perforated rear steel door (#EC02) or a lockable Rear Door Heat Exchanger (RDHX)(1164-95X) is used. Lockable optional side panels (#EC03) increase the rack’s aesthetics, help control airflow through the rack, and provide physical security.

Multiple 42U Slim Racks can be bolted together to create a rack suite (indicate feature code #EC04). Up to six optional 1U PDUs can be placed vertically in the sides of the rack. Additional PDUs can be located horizontally, but they each use 1U of space in this position.

- A minimum of one of the following is required:
  - Feature #ER2B allows you to reserve 2U of space at the bottom of the rack.
  - Feature #ER2T allows you to reserve 2U of space at the top of the rack.

Note: The Power E850 server requires field support to insert in this rack.

1.10.4 Feature code #0551 rack

The 1.8-meter rack (#0551) is a 36U (EIA units) rack. The rack that is delivered as #0551 is the same rack delivered when you order the 7014-T00 rack. The included features might differ. Several features that are delivered as part of the 7014-T00 must be ordered separately with the #0551. The #0551 is initial orders of Power E850 server.

1.10.5 Feature code #0553 rack

The 2.0-meter rack (#0553) is a 42U (EIA units) rack. The rack that is delivered as #0553 is the same rack delivered when you order the 7014-T42. The included features might differ. Several features that are delivered as part of the 7014-T42 or must be ordered separately with the #0553. The #0553 is only available with MES orders of Power E850 server.

1.10.6 The AC power distribution unit and rack content

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

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

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

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 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.

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

http://www.ibm.com/support/knowledgecenter/8408-E8E/p8had/p8had_rpower.htm

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

1.10.7 Rack-mounting rules

Consider the following primary rules when you mount the system into a rack.

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

Three to four service personnel are required to manually remove or insert a system node drawer into a rack, given its dimensions, and weight and content. To avoid the need for this many people to assemble at a client site for a service action a lift tool can be very useful. Similarly, if the client has chosen to install this Customer Set Up (CSU) system, similar lifting considerations apply.

The Power E850 has a maximum weight of 69 kg (152 lbs). However, by temporarily removing the power supplies, fans, and RAID assembly, the weight is easily reduced to a maximum of 55 kg (121 lbs).

When lowering the Power E850 onto its rails in the rack, the server must be tilted on one end about 15 degrees so that the pins on the server enclosure fit onto the rails. This equates to lifting one end of the server about 4 cm (1.6 in.). This can be done by using a tip plate on a lift tool or manually adjusting the load on a lift tool or tilting during the manual lift.

2

Architecture and technical overview

This chapter describes the overall system architecture for the IBM Power System E850 (8408-E8E). The bandwidths that are provided throughout the section are theoretical maximums that are used for reference.

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

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

2.1 Physical system design

The Power E850 is configured as a 4U (4 EIA) rack-mount server, designed to fit into an industry standard 19-inch rack. The server capabilities can be extended by adding I/O expansion drawers to the system. These include the 4U (4 EIA) PCIe Gen3 I/O expansion drawer and the 2U (2 EIA) EXP24S I/O drawer for added local storage capacity.

Table 2-1 lists the dimensions of the system components for installation planning.

Table 2-1 Dimensions of the system components for installation planning

Dimension	Power E850 system node	PCIe Gen3 I/O expansion drawer
Width	448 mm (17.6 in.)	482 mm (19 in.)
Depth	776 mm (30.6 in.)	802 mm (31.6 in.)
Height	175 mm (6.9 in.), 4 EIA units	175 mm (6.9 in.), 4 EIA units
Weight	69.0 kg (152 lb)	54.4 kg (120 lb)

We recommend that a Power E850 server be installed into a rack that has been certified and tested to support the system, such as the IBM 7014-T00 rack.

The front of the system node contains the front fans (five) for the system. It also provides access to the internal storage bays, the operator panel, and the optional DVD-RAM drive. Figure 2-1 shows a front view of the Power E850 server.

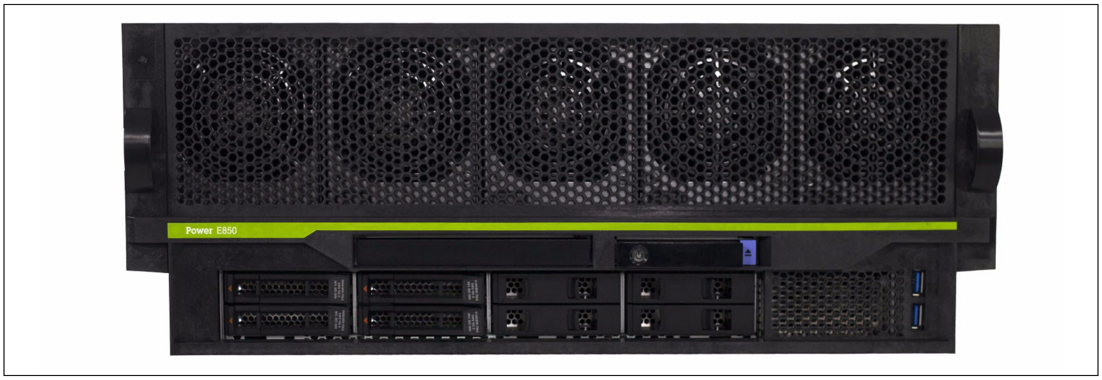


Figure 2-1 Front view of the Power E850 server

The rear of the system provides access to the internally installed PCIe adapter cards, the service processor connections, and the power supply units. Figure 2-2 shows a rear view of the Power E850 server.

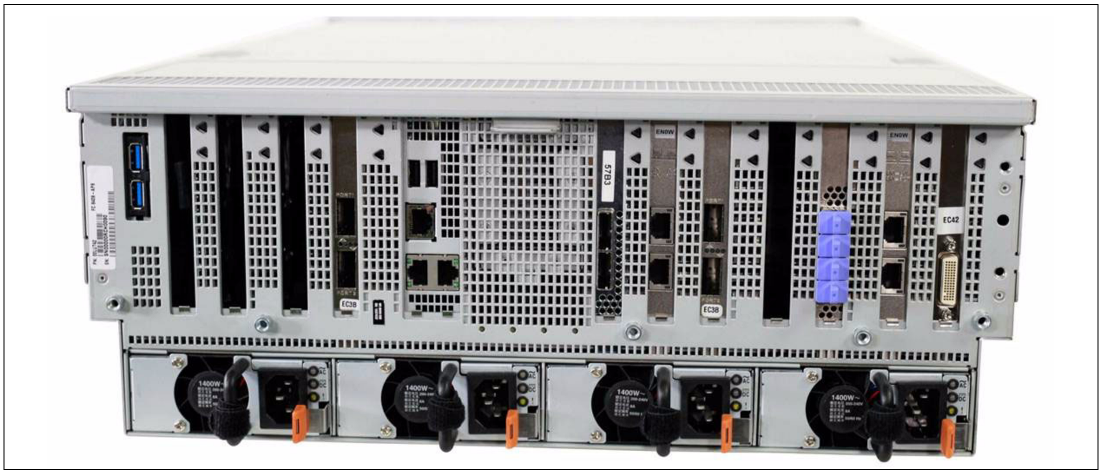


Figure 2-2 Rear view of the Power E850 server

Figure 2-3 shows the internal layout of major components of the Power E850 server within the 4U chassis.

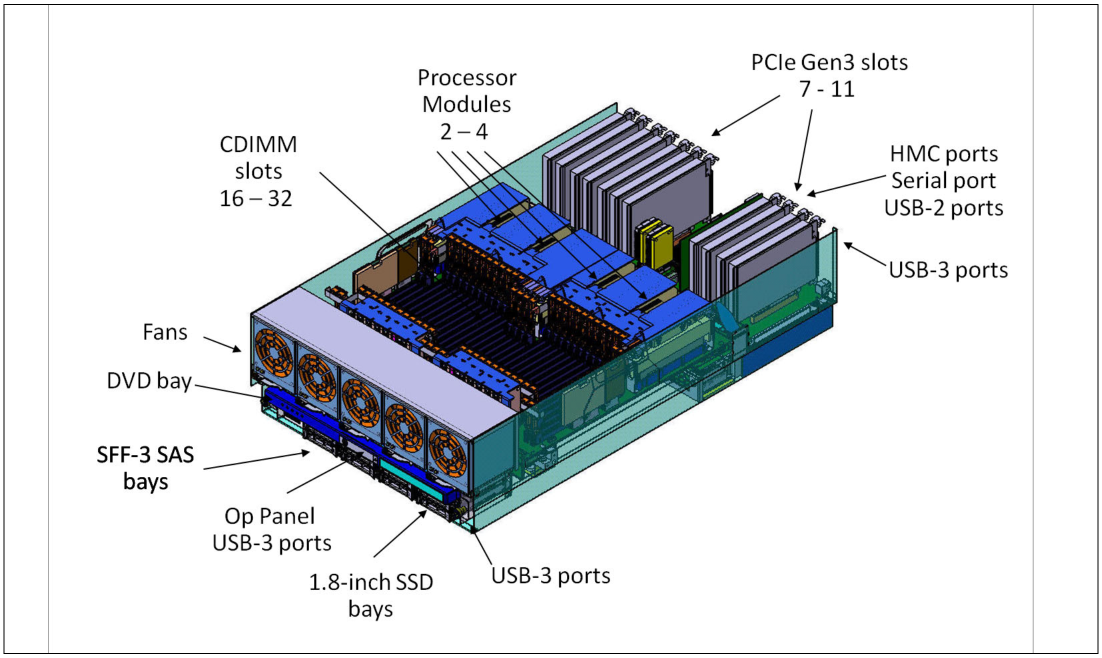


Figure 2-3 Layout view of the Power E850 showing component placement

2.2 Logical system design

This section contains logical diagrams of the Power E850 server in different configurations. It also covers some factors to consider when configuring a new or upgraded Power E850 server.

The Power E850 can be configured with two, three, or four processor modules installed. The number of memory CDIMM slots and the number of PCIe adapter slots that can be utilized varies based on the number of processor modules installed. Table 2-2 shows the possible options.

Table 2-2 Usable memory CDIMM slots and PCIe adapter slots in the Power E850

Processor modules	Memory CDIMM slots	PCIe adapter slots in system node
Two (2)	16	Four x16 and three x8 (seven total)
Three (3)	24	Six x16 and three x8 (nine total)
Four (4)	32	Eight x16 and three x8 (eleven total)

The number of available PCIe adapters supported can be increased by adding PCIe Gen3 I/O expansion drawers to the server. The Power E850 can support up to one PCIe I/O expansion drawer per processor module.

The following factors may influence the placement of PCIe adapter cards in the Power E850 server:

- All PCIe slots in the system node are SR-IOV capable.
- All PCIe x16 slots in the system node have dedicated bandwidth (16 lanes of PCIe Gen 3 bandwidth).
- One of the PCIe x8 slots (C11) has dedicated bandwidth (eight lanes of PCIe Gen3 bandwidth). The other two PCIe x8 slots (C6 and C7) share bandwidth (via PEX unit, and also shared with the USB 3.0 controller).
- All PCIe x16 slots in the system node are CAPI capable.

Figure 2-4 shows the logical design of the Power E850 server with two processor modules installed.

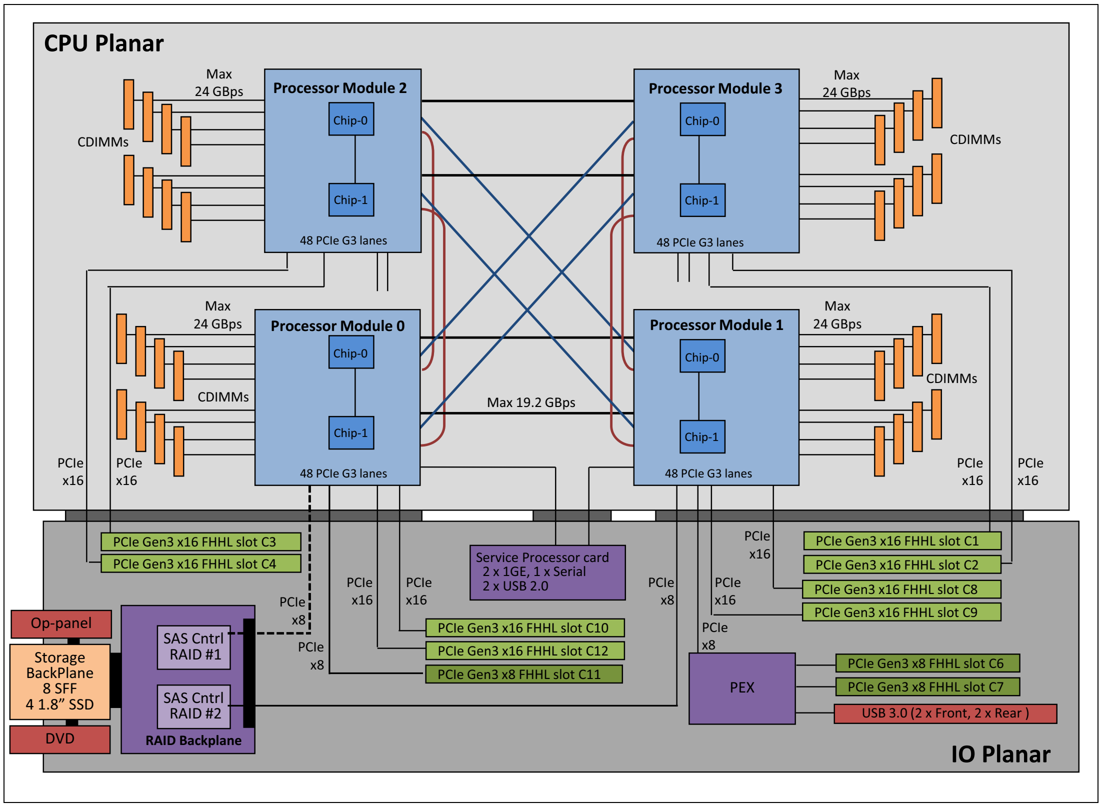


Figure 2-4 Logical diagram for a Power E850 with two processor modules installed

Figure 2-5 shows the logical design of the Power E850 server with three processor modules installed.

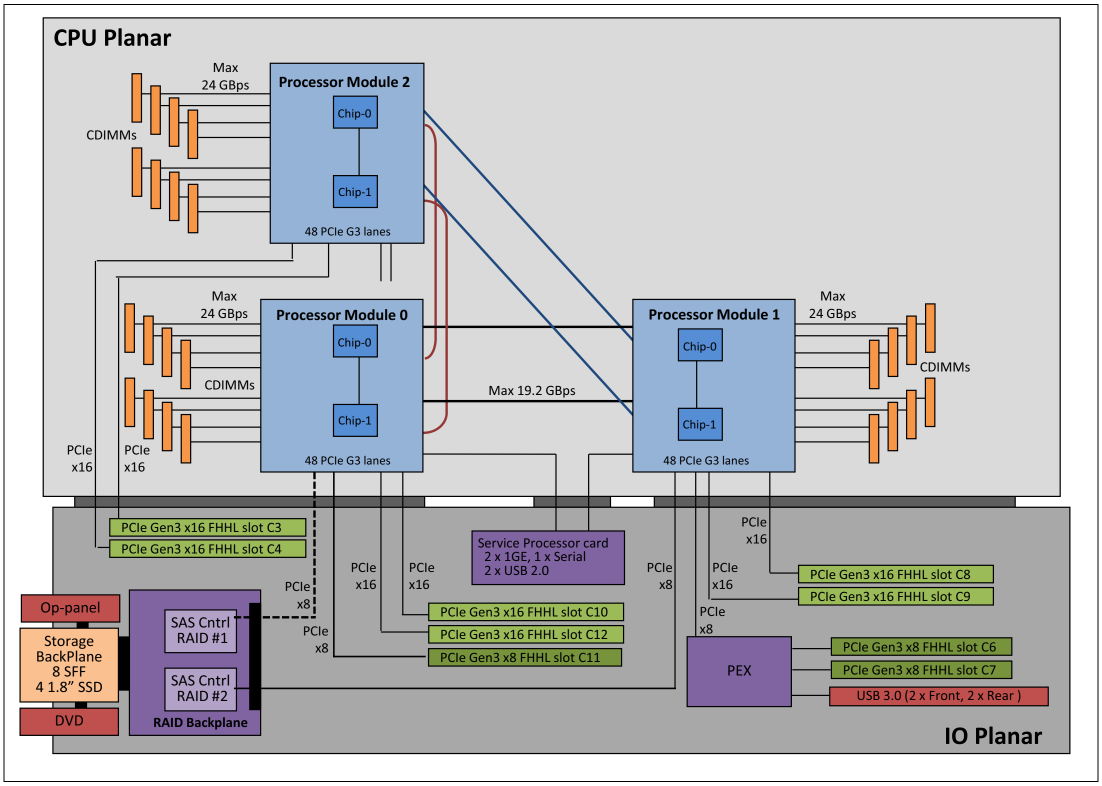


Figure 2-5 Logical diagram for a Power E850 with three processor modules installed

Figure 2-6 shows the logical design of the Power E850 server with four processor modules installed.

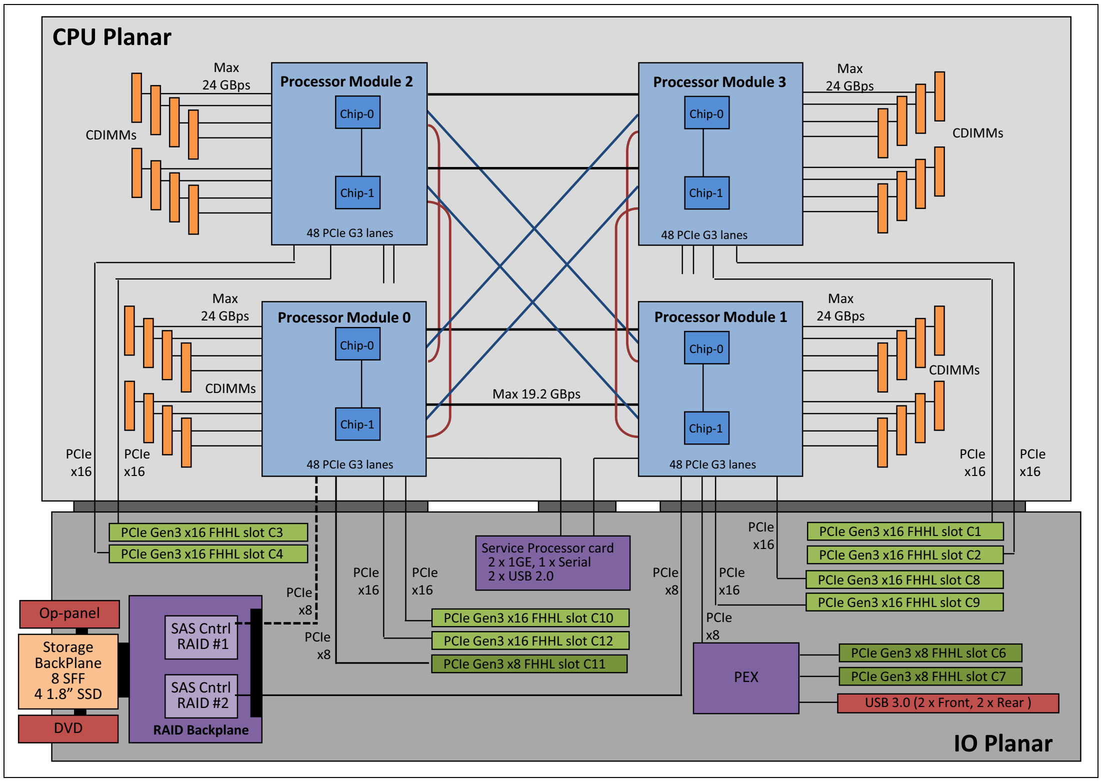

CPU Planar
Processor Module 2
Processor Module 3
Processor Module 0
Processor Module 1
Chip-0
Chip-1
48 PCIe G3 lanes
Max
24 GBps
CDIMMs
Max 19.2 GBps
PCIe
x16
PCIe
x8
PCIe Gen3 x16 FHHL slot C3
PCIe Gen3 x16 FHHL slot C4
PCIe Gen3 x16 FHHL slot C1
PCIe Gen3 x16 FHHL slot C2
PCIe Gen3 x16 FHHL slot C8
PCIe Gen3 x16 FHHL slot C9
PCIe Gen3 x16 FHHL slot C10
PCIe Gen3 x16 FHHL slot C12
PCIe Gen3 x8 FHHL slot C11
PCIe Gen3 x8 FHHL slot C6
PCIe Gen3 x8 FHHL slot C7
USB 3.0 (2 x Front, 2 x Rear)
Service Processor card
2 x 1GE, 1 x Serial
2 x USB 2.0
PEX
Op-panel
Storage
BackPlane
8 SFF
4 1.8" SSD
DVD
SAS Cntrl
RAID #1
SAS Cntrl
RAID #2
RAID Backplane
IO Planar

Figure 2-6 Logical diagram for a Power E850 server with four processor modules installed

2.3 The IBM POWER8 processor

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

2.3.1 POWER8 processor overview

The POWER8 processor is manufactured by using the IBM 22 nm Silicon-On-Insulator (SOI) technology. Each chip is 649 mm^2 and contains 4.2 billion transistors. As shown in Figure 2-7, the chip contains 12 cores, two memory controllers, and an interconnection system that connects all components within the chip. On some systems, only 6, 8, 10, or 12 cores per processor may be available to the server. Each core has 512 KB of L2 cache, and all cores share 96 MB of L3 embedded DRAM (eDRAM). The interconnect also extends through module and board technology to other POWER8 processors in addition to DDR3 memory and various I/O devices.

POWER8 systems use memory buffer chips to interface between the POWER8 processor and DDR3 or DDR4[1] memory. Each buffer chip also includes an L4 cache to reduce the latency of local memory accesses.

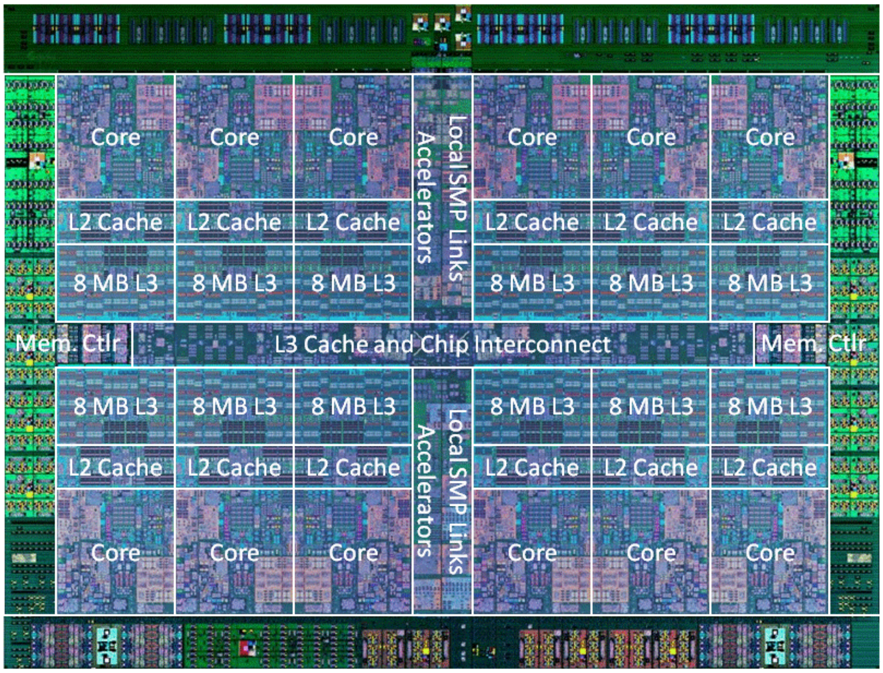


Figure 2-7 The POWER8 processor chip

The POWER8 processor is designed for system offerings from single-socket servers to multi-socket Enterprise servers such as the Power E850. It incorporates a triple-scope broadcast coherence protocol over local and global SMP links to provide superior scaling attributes. Multiple-scope coherence protocols reduce the amount of SMP link bandwidth that is required by attempting operations on a limited scope (single chip or multi-chip group) when possible. If the operation cannot complete coherently, the operation is reissued using a larger scope to complete the operation.

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

- Support is provided for DDR3 and DDR4[1] memory through memory buffer chips that offload the memory support from the POWER8 memory controller.
- Each memory CDIMM has 16 MB of L4 cache within the memory buffer chip that reduces the memory latency for local access to memory behind the buffer chip; the operation of the

[1] At the time of the publication, the available POWER8 processor-based systems use DDR3 memory. The memory subsystem design allows for the use of DDR4 memory or other memory technology without any architectural changes to the processor.

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

- Shared L3 cache, allowing cores to utilize shared L3 cache and unused L3 cache from other cores when needed using the L3 cache interconnects.
- Hardware transactional memory.
- On-chip accelerators, including on-chip encryption, compression, and random number generation accelerators.
- Coherent Accelerator Processor Interface, which allows accelerators plugged into a PCIe slot to access the processor bus using a low latency, high-speed protocol interface.
- Adaptive power management.

2.3.2 The POWER8 processor module

There are two versions of the POWER8 processor chip. Both chips use the same building blocks. The scale-out systems and the Power E850 use a 6-core version of POWER8. The 6-core chip is installed in pairs in a Dual Chip Module (DCM) that plugs into a socket in the system board of the systems. Functionally, it works as a single chip module (SCM).

Figure 2-8 shows a graphic representation of the 6-core processor. Two 6-core processors are combined within a DCM in a configuration that is only available on the scale-out systems and the Power E850. The Power E870 and Power E880 servers use the 12-core SCM.

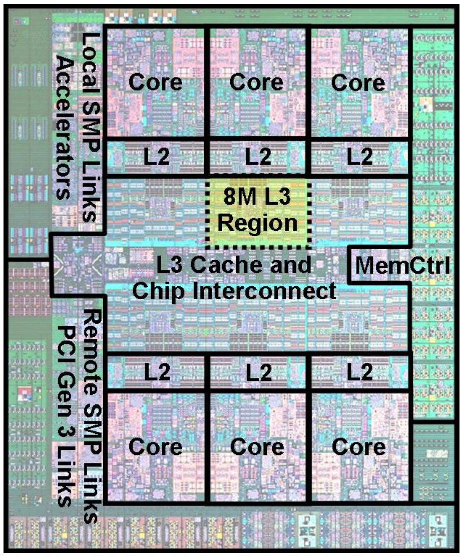


Figure 2-8 6-core POWER8 processor chip

Table 2-3 summarizes the technology characteristics of the POWER8 processor.

Table 2-3 Summary of POWER8 processor technology

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

2.3.3 POWER8 processor core

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

- Multi-threaded design, which is capable of up to eight-way simultaneous multithreading (SMT)
- 32 KB, eight-way set-associative L1 instruction cache
- 64 KB, eight-way set-associative L1 data cache
- Enhanced prefetch, with instruction speculation awareness and data prefetch depth awareness
- Enhanced branch prediction, using both local and global prediction tables with a selector table to choose the best predictor
- Improved out-of-order execution
- Two symmetric fixed-point execution units
- Two symmetric load/store units and two load units, all four of which can also run simple fixed-point instructions
- An integrated, multi-pipeline vector-scalar floating point unit for running both scalar and SIMD-type instructions, including the Vector Multimedia eXtension (VMX) instruction set and the improved Vector Scalar eXtension (VSX) instruction set, and capable of up to sixteen floating point operations per cycle (eight double precision or sixteen single precision)
- In-core Advanced Encryption Standard (AES) encryption capability
- Hardware data prefetching with 16 independent data streams and software control
- Hardware decimal floating point (DFP) capability

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

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

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

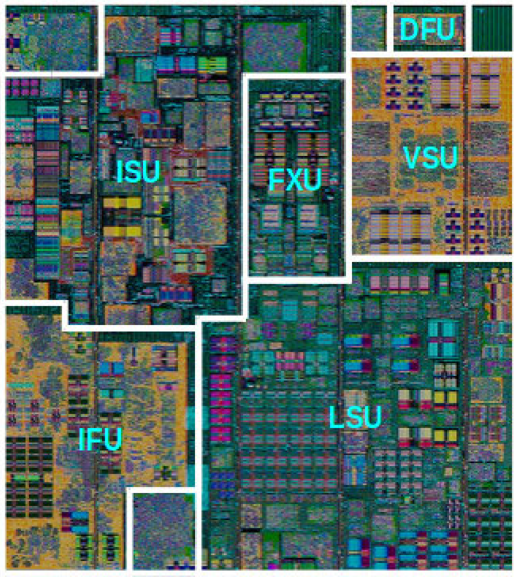


Figure 2-9 POWER8 processor core

2.3.4 Simultaneous multithreading

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

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

Table 2-4 shows a comparison between the different POWER processors in terms of SMT capabilities that are supported by each processor architecture.

Table 2-4 SMT levels that are supported by POWER processors

Technology	Cores/system	Maximum SMT mode	Maximum hardware threads per system
IBM POWER4	32	Single Thread (ST)	32
IBM POWER5	64	SMT2	128
IBM POWER6	64	SMT2	128

Technology	Cores/system	Maximum SMT mode	Maximum hardware threads per system
IBM POWER7	256	SMT4	1024
IBM POWER8	192[a]	SMT8	1536[a]

[a] The Power E880 server supports up to 192 installed cores running SMT8, for a total of 1536 threads. The Power E850 supports a maximum of 48 installed cores capable of running SMT8 for a maximum of 384 threads per system.

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

2.3.5 Memory access

On the Power E850, each POWER8 processor has two memory controllers, each connected to four memory channels. Each memory channel operates at 1600 MHz and connects to a CDIMM. Each CDIMM on a POWER8 system has a memory buffer that is responsible for many functions that were previously on the memory controller itself, such as scheduling logic and energy management. The memory buffer also has 16 MB of level 4 (L4) cache. This provides Enterprise level RAS features, as well as flexibility to make use of other memory technologies in the future without changes to the processor architecture.

On the Power E850, each memory channel can address up to 128 GB of memory. Therefore, a server with four processor modules installed can address up to 4 TB of memory. A server with three processor modules installed can address up to 3 TB of memory and a server with two processor modules installed can address up to 2 TB of memory.

Figure 2-10 gives a simple overview of the POWER8 processor memory access structure in the Power E850.

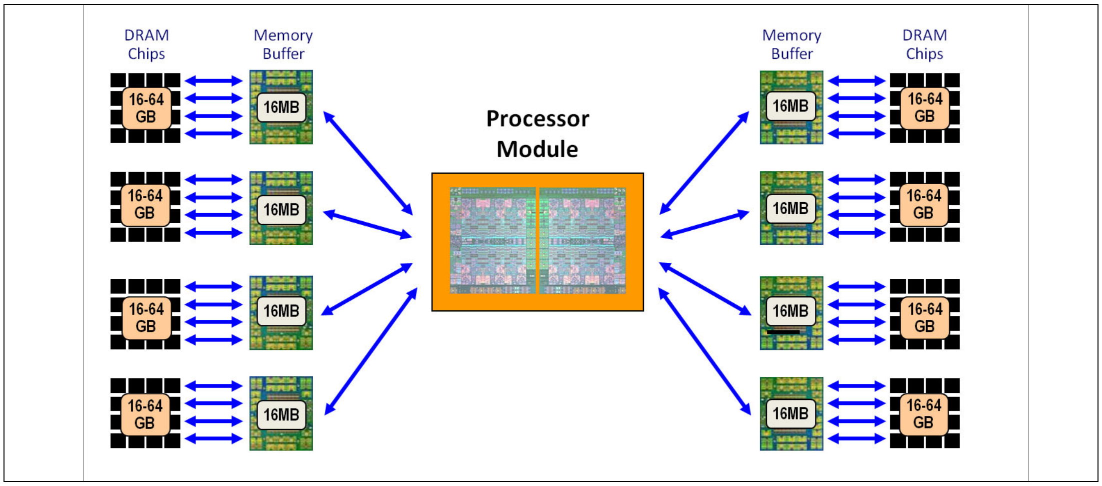


Figure 2-10 Overview of POWER8 memory access structure

2.3.6 On-chip L3 cache innovation and Intelligent Cache

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

The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a fast 8 MB local region of L3 cache (FLR-L3), but also has access to other L3 cache regions as shared L3 cache. Additionally, each core can negotiate to use the FLR-L3 cache that is associated with another core, depending on the reference patterns. Data can also be cloned and stored in more than one core's FLR-L3 cache, again depending on the reference patterns. This Intelligent Cache management enables the POWER8 processor to optimize the access to L3 cache lines and minimize overall cache latencies.

Figure 2-7 on page 34 and Figure 2-8 on page 35 show the on-chip L3 cache, and highlight one fast 8 MB L3 region closest to a processor core.

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

- Latency improvement

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

- Bandwidth improvement

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

- No off-chip driver or receivers

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

- Small physical footprint

  The performance of eDRAM when implemented on-chip is similar to conventional SRAM but requires far less physical space. IBM on-chip eDRAM uses only a third of the components that conventional SRAM uses, which has a minimum of six transistors to implement a 1-bit memory cell.

- Low energy consumption

  The on-chip eDRAM uses only 20% of the standby power of SRAM.

2.3.7 Level 4 cache and memory buffer

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

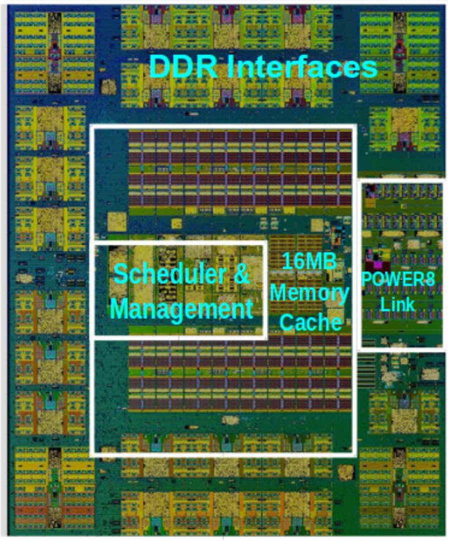


Figure 2-11 Memory buffer chip

Table 2-5 shows a comparison of the different levels of cache in the POWER7, POWER7+, and POWER8 processors.

Table 2-5 POWER8 cache hierarchy

Cache	POWER7	POWER7+	POWER8
L1 instruction cache: Capacity/associativity	32 KB/4-way	32 KB/4-way	32 KB/8-way
L1 data cache: Capacity/associativity bandwidth	32 KB/8-way Two 16 B reads or one 16 B write per cycle	32 KB/8-way Two 16 B reads or one 16 B write per cycle	64 KB/8-way Two 16 B reads or one 16 B write per cycle
L2 cache: Capacity/associativity bandwidth	256 KB/8-way Private 32 B reads and 16 B writes per cycle	256 KB/8-way Private 32 B reads and 16 B writes per cycle	512 KB/8-way Private 32 B reads and 16 B writes per cycle
L3 cache: Capacity/associativity bandwidth	On-Chip 4 MB/core/8-way 16 B reads and 16 B writes per cycle	On-Chip 10 MB/core/8-way 16 B reads and 16 B writes per cycle	On-Chip 8 MB/core/8-way 32 B reads and 32 B writes per cycle
L4 cache: Capacity/associativity bandwidth	N/A	N/A	Off-Chip 16 MB/buffer chip/16-way Up to 8 buffer chips per socket

For more information about the POWER8 memory subsystem, see 2.4, “Memory subsystem” on page 44.

2.3.8 Hardware transactional memory

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

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

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

2.3.9 Coherent Accelerator Processor Interface

The Coherent Accelerator Interface Architecture (CAIA) defines a coherent accelerator interface structure for attaching peripherals to Power Systems. This allows accelerators to work coherently with system memory, removing overhead from the main system processor and reducing the overall memory requirements for the accelerator.

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

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

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

Figure 2-12 on page 42 shows a high-level view of how an accelerator communicates with the POWER8 processor through CAPI. The POWER8 processor provides a Coherent Attached Processor Proxy (CAPP), which is responsible for extending the coherence in the processor communications to an external device. The coherency protocol is tunneled over standard PCIe Gen3 connections, effectively making the accelerator part of the coherency domain.

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

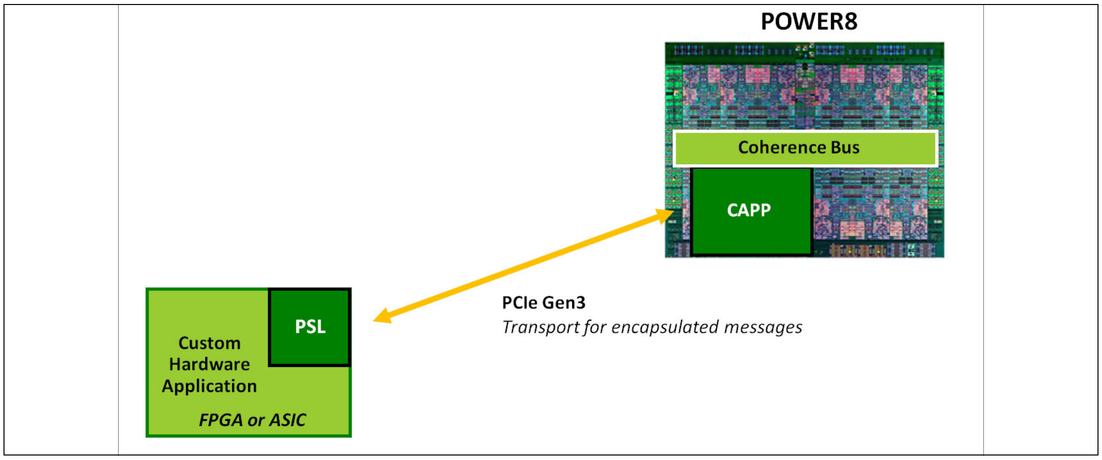


Figure 2-12 CAPI accelerator that is attached to the POWER8 processor

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

2.3.10 Power management and system performance

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

Static Power Saver

Static Power Saver lowers the processor frequency and voltage a fixed amount, reducing the power 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.

Dynamic Power Saver: favor performance

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

Dynamic Power Saver: favor power

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

Dynamic Power Saver: Tunable Parameters

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

Idle Power Saver

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

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

http://www.ibm.com//common/ssi/ecm/po/en/pow03125usen/index.html

For more information, see 2.15, “Energy management” on page 101.

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

Table 2-6 shows comparable characteristics between the generations of POWER7, POWER7+, and POWER8 processors.

Table 2-6 Comparison of technologies for the POWER8 processor and the prior generations

Characteristics	POWER7	POWER7+	POWER8
Technology	45 nm	32 nm	22 nm
Die size	567 mm^2	567 mm^2	649 mm^2

Characteristics	POWER7	POWER7+	POWER8
Number of transistors	1.2 billion	2.1 billion	4.2 billion
Maximum cores	8	8	12
Maximum SMT threads per core	4 threads	4 threads	8 threads
Maximum frequency	4.25 GHz	4.4 GHz	4.35 GHz
L1 Cache	32 KB per core	32 KB per core	32 KB per core
L2 Cache	256 KB per core	256 KB per core	512 KB per core
L3 Cache	4 MB or 8 MB of FLR-L3 cache per core with each core having access to the full 32 MB of L3 cache, on-chip eDRAM	10 MB of FLR-L3 cache per core with each core having access to the full 80 MB of L3 cache, on-chip eDRAM	8 MB of FLR-L3 cache per core with each core having access to the full 96 MB of L3 cache, on-chip eDRAM
L4 Cache	N/A	N/A	Off-chip: 16 MB per buffer chip, up to 8 buffer chips per socket
Memory support	DDR3	DDR3	DDR3 and DDR4
I/O bus	GX++	GX++	PCIe Gen3

2.4 Memory subsystem

The Power E850 can have two, three, or four processor modules installed per server. Each processor module enables eight DDR3 CDIMM slots capable of supporting 16 GB, 32 GB, and 64 GB CDIMMs running at speeds of 1600 MHz or 128 GB DDR4 CDIMS running at speeds of 1600 MHz. A server with four processor modules can support up to 4 TB of installed memory, a server with three processor modules installed can support up to 3 TB of installed memory, while a server with two processor modules installed can support up to 2 TB of installed memory.

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

The Power E850 server supports an optional feature called *Active Memory Expansion* (#4798). This allows the effective maximum memory capacity to be much larger than the true physical memory maximum. Sophisticated compression and decompression of memory content using the POWER8 processor along with a dedicated coprocessor present on each POWER8 processor can provide memory expansion up to 100% or more. This ratio depends on the workload type and its memory usage. As an example, a server with 256 GB of RAM physically installed can effectively be expanded over 512 GB of RAM. 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. The processor resource used to expand memory is part of the processor entitlement assigned to the partition enabling Active Memory Expansion.

Active Memory Expansion is compatible with all AIX partitions. Each individual AIX partition can have Active Memory Expansion enabled or disabled. Control parameters set the amount of expansion wanted in each partition to find a balance of memory expansion and processor utilization. A partition needs to be rebooted to turn on Active Memory Expansion.

A planning tool is included within AIX that allows you to sample actual workloads and estimate the level of expansion and processor usage expected. This can be run on any Power Systems server running PowerVM as a hypervisor. A one-time, 60-day trial of Active Memory Expansion is available on each server to confirm the estimated results. You can request the trial activation code on the IBM Power Systems Capacity on Demand website:

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

To activate Active Memory Expansion on a Power E850 server, the chargeable feature code #4798 must be ordered, either as part of the initial system order or as an MES upgrade. A software key is provided which is applied to the server. There is no need to reboot the system.

2.5 Memory Custom DIMMs

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

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

By adopting this architecture for the memory DIMMs, several decisions and processes regarding memory optimizations are run internally into the CDIMM. This saves bandwidth and allows for faster processor-to-memory communications. This also allows for a more robust RAS. For more information, see Chapter 4, “Reliability, availability, and serviceability” on page 147.

2.5.1 CDIMM design

The CDIMMs exist in two different form factors, a 152 SDRAM design named the *Tall CDIMM* and an 80 SDRAM design named the *Short CDIMM*. Each design is composed of multiple 4 GB SDRAM devices depending on its total capacity. The CDIMMs for the Power E850 server are short CDIMMs. Tall CDIMMs from other Enterprise Systems such as the Power E870 and Power E880 are not compatible with the Power E850 server.

The Power E850 supports CDIMMs in 16 GB, 32 GB, 64 GB or 128 GB capacities. Each CDIMM incorporates a 16 MB Memory Buffer, also known as *L4 cache*, which 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 E850 server with four processor modules has 512 MB of L4 Cache. The L4 Cache performs several functions that have direct impact on performance and bring a series of benefits for the Power E850 server:

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

2.5.2 Memory placement rules

For the Power E850, each memory feature code provides a single CDIMM. These memory features must be ordered in pairs of the same memory feature. Both CDIMMs of a CDIMM pair must be installed in CDIMM slots supporting one processor. Different size pairs can be mixed on the same processor, however for optimal performance we recommend that all CDIMM pairs connected to a processor are of the same capacity. We would also recommend that the number of memory CDIMM pairs is the same on each processor module of a system.

System performance improves as more CDIMM pairs match. System performance also improves as more CDIMM slots are filled, as this increases the memory bandwidth available. Therefore, if 256 GB of memory is required, using sixteen 16 GB CDIMMs would offer better performance than using eight 32 GB CDIMMs. This allows memory access in a consistent manner and typically results in the best possible performance for your configuration. You should account for any plans for future memory upgrades when you decide which memory feature size to use at the time of the initial system order.

A minimum of four CDIMM slots must be populated for each installed processor module. For a Power E850 with two processor modules, eight CDIMM slots must be populated. For a Power E850 with three processor modules installed, 12 CDIMM slots must be populated, and for a Power E850 with four processor modules installed, 16 CDIMM slots must be populated.

All the memory CDIMMs are capable of capacity upgrade on demand and must have a minimum of 50% of its physical capacity activated. For example, if a Power E850 has 512 GB of memory installed, a minimum of 256 GB would need to be activated. A minimum of 128 GB of memory must be activated on each system. For more information about Capacity on Demand and activation requirements, see 2.6, “Capacity on Demand” on page 54.

For the Power E850, the following memory options are orderable:

- 16 GB CDIMM, 1600 MHz DDR3 DRAM (#EM86)
- 32 GB CDIMM, 1600 MHz DDR3 DRAM (#EM87)
- 64 GB CDIMM, 1600 MHz DDR3 DRAM (#EM88)
- 128 GB CDIMM, 1600 MHz DDR4 DRAM (#EM8S)

Note that DDR4 DRAMs provide the same memory throughput as DDR3.

Table 2-7 summarizes the minimum and maximum CDIMM and memory requirements for the Power E850 server.

Table 2-7 Minimum and maximum memory requirements

Number of processor modules installed	Minimum number of CDIMMs	Minimum memory capacity (16 GB CDIMMs)	Maximum number of CDIMMs	Maximum memory capacity (128 GB CDIMMs)
Two	8	128 GB	16	2048 GB (2 TB)
Three	12	192 GB	24	3072 GB (3 TB)
Four	16	256 GB	32	4096 GB (4 TB)

The basic rules for memory placement follow:

- Each feature code provides a single memory CDIMM.
- Memory CDIMMs must be ordered and installed in matching pairs.
- For each installed processor, there must be at least four CDIMMs populated.
- There are a maximum of eight memory CDIMMs per installed processor.
- At least 50% of the installed memory must be activated via memory activation features.

- DDR4 memory modules cannot be mixed with DDR3, therefore if a DDR4 feature is used, all the memory on the system must be DDR4

2.5.3 CDIMM plugging order

Each processor module has two memory controllers. Each of these memory controllers connects to 4 CDIMM slots within the system. The physical connections and location codes of the memory slots are shown in Figure 2-13.

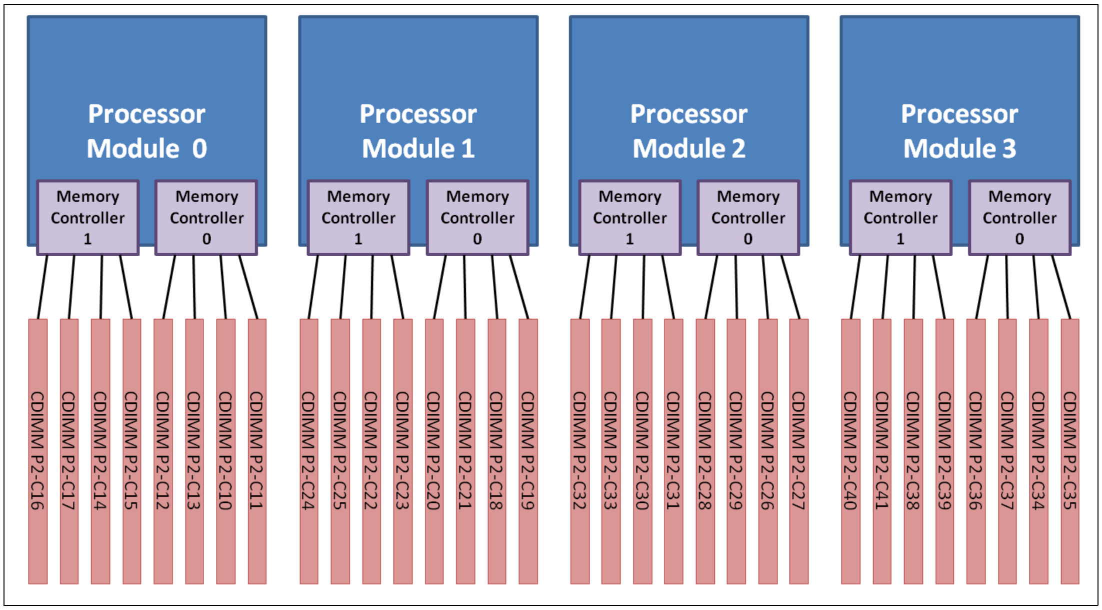


Figure 2-13 Memory slot physical connections and location codes for the Power E850

Table 2-8 shows the recommended plugging order for CDIMMs in a Power E850 server with two processor modules installed to ensure optimal performance and CDIMM size flexibility. Each number represents two matching capacity CDIMMs in two adjacent CDIMM slots.

Table 2-8 Memory plugging order for Power E850 with two processor modules

Processor Module 0				**Processor Module 1**			
Memory Controller 1		Memory Controller 0		Memory Controller 1		Memory Controller 0	
C16 C17	C14 C15	C12 C13	C10 C11	C24 C25	C22 C23	C20 C21	C18 C19
1	5	3	7	2	6	4	8

- First CDIMM pair is identical and plugged into P2-C16 and P2-C17
- Second CDIMM pair is identical and plugged into P2-C24 and P2-C25
- Third CDIMM pair is identical and plugged into P2-C12 and P2-C13
- Fourth CDIMM pair is identical and plugged into P2-C20 and P2-C21
- Fifth CDIMM pair is identical and plugged into P2-C14 and P2-C15
- Sixth CDIMM pair is identical and plugged into P2-C22 and P2-C23
- Seventh CDIMM pair is identical and plugged into P2-C10 and P2-C11
- Eighth CDIMM pair is identical and plugged into P2-C18 and P2-C19

Table 2-9 on page 48 shows the recommended plugging order for CDIMMs in a Power E850 server with three processor modules installed to ensure optimal performance and CDIMM size flexibility. Each number represents two matching capacity CDIMMs in two adjacent CDIMM slots.

Table 2-9 Memory plugging order for Power E850 with three processor modules

Processor Module 0				**Processor Module 1**			
Memory Controller 1		Memory Controller 0		Memory Controller 1		Memory Controller 0	
C16 C17	C14 C15	C12 C13	C10 C11	C24 C25	C22 C23	C20 C21	C18 C19
1	7	4	10	2	8	5	11

Processor Module 0			
Memory Controller 1		Memory Controller 0	
C32 C33	C30 C31	C28 C29	C26 C27
3	9	6	12

- First CDIMM pair is identical and plugged into P2-C16 and P2-C17
- Second CDIMM pair is identical and plugged into P2-C24 and P2-C25
- Third CDIMM pair is identical and plugged into P2-C32 and P2-C33
- Fourth CDIMM pair is identical and plugged into P2-C12 and P2-C13
- Fifth CDIMM pair is identical and plugged into P2-C20 and P2-C21
- Sixth CDIMM pair is identical and plugged into P2-C28 and P2-C29
- Seventh CDIMM pair is identical and plugged into P2-C14 and P2-C15
- Eighth CDIMM pair is identical and plugged into P2-C22 and P2-C23
- Ninth CDIMM pair is identical and plugged into P2-C30 and P2-C31
- Tenth CDIMM pair is identical and plugged into P2-C10 and P2-C11
- Eleventh CDIMM pair is identical and plugged into P2-C18 and P2-C19
- Twelfth CDIMM pair is identical and plugged into P2-C26 and P2-C27

Table 2-10 shows the recommended plugging order for CDIMMs in a Power E850 server with four processor modules installed to ensure optimal performance and CDIMM size flexibility. Each number represents two matching capacity CDIMMs in two adjacent CDIMM slots.

Table 2-10 Memory plugging order for Power E850 with four processor modules

Processor Module 0				**Processor Module 1**			
Memory Controller 1		Memory Controller 0		Memory Controller 1		Memory Controller 0	
C16 C17	C14 C15	C12 C13	C10 C11	C24 C25	C22 C23	C20 C21	C18 C19
1	9	5	13	2	10	6	14

Processor Module 0				**Processor Module 1**			
Memory Controller 1		Memory Controller 0		Memory Controller 1		Memory Controller 0	
C32 C33	C30 C31	C28 C29	C26 C27	C40 C41	C38 C39	C36 C37	C34 C35
3	11	7	15	4	12	8	16

- First CDIMM pair is identical and plugged into P2-C16 and P2-C17
- Second CDIMM pair is identical and plugged into P2-C24 and P2-C25
- Third CDIMM pair is identical and plugged into P2-C32 and P2-C33

- Fourth CDIMM pair is identical and plugged into P2-C40 and P2-C41
- Fifth CDIMM pair is identical and plugged into P2-C12 and P2-C13
- Sixth CDIMM pair is identical and plugged into P2-C20 and P2-C21
- Seventh CDIMM pair is identical and plugged into P2-C28 and P2-C29
- Eighth CDIMM pair is identical and plugged into P2-C36 and P2-C37
- Ninth CDIMM pair is identical and plugged into P2-C14 and P2-C15
- Tenth CDIMM pair is identical and plugged into P2-C22 and P2-C23
- Eleventh CDIMM pair is identical and plugged into P2-C30 and P2-C31
- Twelfth CDIMM pair is identical and plugged into P2-C38 and P2-C39
- Thirteenth CDIMM pair is identical and plugged into P2-C10 and P2-C11
- Fourteenth CDIMM pair is identical and plugged into P2-C18 and P2-C19
- Fifteenth CDIMM pair is identical and plugged into P2-C26 and P2-C27
- Sixteenth CDIMM pair is identical and plugged into P2-C34 and P2-C35

2.5.4 Memory activations

Several types of Capacity on Demand capability are available for processors and memory on the Power E850 server. All the memory CDIMMs in a Power E850 system are capable of capacity upgrade on demand and must have a minimum of 50% of their physical capacity activated. A minimum of 128 GB of memory must be activated on each system. The remaining installed memory capacity can be activated as Capacity on Demand, either permanently or temporarily.

Initial activations of memory resources must meet the minimum requirement of 50% of all installed memory, or 128 GB, whichever is higher. The maximum initial order for activations is the entire installed memory capacity.

Initial memory activations can be ordered in quantities of 1 GB (#EMAA) and 100 GB (#EMAB), and this memory capacity remains permanently active.

Any memory resources that are installed but not activated as part of the initial order are available for future use of the Capacity on Demand features of the Power E850.

For more information about Capacity on Demand and activation requirements, see 2.6, “Capacity on Demand” on page 54.

2.5.5 Memory throughput

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

> **Warning:** All bandwidth figures listed in the section are theoretical maximums, and are based on the nominal clock speeds of the processors listed. Because clock speeds can vary depending on power-saving mode and server load, these bandwidths are provided for information only.

Cache bandwidths

Table 2-11 on page 50 shows the maximum bandwidth estimates for a single core on the Power E850 system.

Table 2-11 Power E850 single core bandwidth estimates

Single core	Power E850	Power E850	Power E850
	1 core @ 3.02 GHz	**1 core @ 3.35 GHz**	**1 core @ 3.72 GHz**
L1 (data) cache	145.25 GBps	161.23 GBps	178.75 GBps
L2 cache	145.25 GBps	161.23 GBps	178.75 GBps
L3 cache	193.66 GBps	214.98 GBps	238.33 GBps

The bandwidth figures for the caches are calculated as follows:

- L1 cache: In one clock cycle, two 16-byte load operations and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 3.026 GHz Core: (2 x 16 B + 1 x 16 B) x 3.026 GHz = 145.25 GBps
  - 3.359 GHz Core: (2 x 16 B + 1 x 16 B) x 3.359 GHz = 161.23 GBps
  - 3.724 GHz Core: (2 x 16 B + 1 x 16 B) x 3.724 GHz = 178.75 GBps
- L2 cache: In one clock cycle, one 32-byte load operation and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 3.026 GHz Core: (1 x 32 B + 1 x 16 B) x 3.026 GHz = 145.25 GBps
  - 3.359 GHz Core: (1 x 32 B + 1 x 16 B) x 3.359 GHz = 161.23 GBps
  - 3.724 GHz Core: (1 x 32 B + 1 x 16 B) x 3.724 GHz = 178.75 GBps
- L3 cache: In one clock cycle, one 32-byte load operation and one 32-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:
  - 3.026 GHz Core: (1 x 32 B + 1 x 32 B) x 3.026 GHz = 193.66 GBps
  - 3.359 GHz Core: (1 x 32 B + 1 x 32 B) x 3.359 GHz = 214.98 GBps
  - 3.724 GHz Core: (1 x 32 B + 1 x 32 B) x 3.724 GHz = 238.33 GBps

Memory bandwidths

Each processor module in the Power E850 server has two memory controllers, each of which controls four memory channels, each of which can be connected to a CDIMM. These high-speed memory channels run at 8 GHz, and can support two byte read operations and one byte write operation concurrently. This is independent of the processor clock speed.

Note: DDR4 memory (#EM8S) is accessed at the same speed as DDR3 memory - 1600 MHz.

So a single processor module can support a memory bandwidth of:

8 GHz x 8 channels x (2 x read byte + 1 x write byte) = 192 GBps

This calculation assumes that all memory CDIMM slots are populated. If fewer memory CDIMM slots are populated, this figure is lower.

The theoretical maximum memory bandwidth of a server is dependent on the number of processor modules installed. These figures are listed in Table 2-12 on page 51.

Table 2-12 Maximum theoretical memory bandwidths

Processor modules installed	Two processor modules installed	Three processor modules installed	Four processor modules installed
Maximum theoretical memory bandwidth	384 GBps	576 GBps	768 GBps

Server summaries

The following tables summarize the cache and memory bandwidths for a number of different Power E850 server configurations.

Table 2-13 shows the theoretical maximum bandwidths for two, three, or four installed 3.02 GHz (12 core) processor modules.

Table 2-13 Theoretical maximum bandwidths for 3.02 GHz system configurations

System bandwidth	Power E850	Power E850	Power E850
	2 processor modules @ 3.02 GHz (24 cores)	3 processor modules @ 3.02 GHz (36 cores)	4 processor modules @ 3.02 GHz (48 cores)
L1 (data) cache	3486 GBps	5229 GBps	6972 GBps
L2 cache	3486 GBps	5229 GBps	6972 GBps
L3 cache	4648 GBps	6972 GBps	9296 GBps
L4 cache/memory	384 GBps	576 GBps	768 GBps

Table 2-14 shows the theoretical maximum bandwidths for two, three, or four installed 3.35 GHz (12 core) processor modules.

Table 2-14 Theoretical maximum bandwidths for 3.35 GHz system configurations

System bandwidth	Power E850	Power E850	Power E850
	2 processor modules @ 3.35 GHz (20 cores)	3 processor modules @ 3.35 GHz (30 cores)	4 processor modules @ 3.35 GHz (40 cores)
L1 (data) cache	3225 GBps	4837 GBps	6450 GBps
L2 cache	3225 GBps	4837 GBps	6450 GBps
L3 cache	4300 GBps	6450 GBps	8600 GBps
L4 cache/memory	384 GBps	576 GBps	768 GBps

Table 2-15 shows the theoretical maximum bandwidths for two, three, or four installed 3.72 GHz (12 core) processor modules.

Table 2-15 Theoretical maximum bandwidths for 3.72 GHz system configurations

System bandwidth	Power E850	Power E850	Power E850
	2 processor modules @ 3.72 GHz (16 cores)	3 processor modules @ 3.72 GHz (24 cores)	4 processor modules @ 3.72 GHz (32 cores)
L1 (data) cache	2860 GBps	4290 GBps	5720 GBps
L2 cache	2860 GBps	4290 GBps	5720 GBps
L3 cache	3813 GBps	5720 GBps	7927 GBps

System bandwidth	Power E850	Power E850	Power E850
	2 processor modules @ 3.72 GHz (16 cores)	3 processor modules @ 3.72 GHz (24 cores)	4 processor modules @ 3.72 GHz (32 cores)
L4 cache/memory	384 GBps	576 GBps	768 GBps

2.5.6 Active Memory Mirroring

The Power E850 server 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.

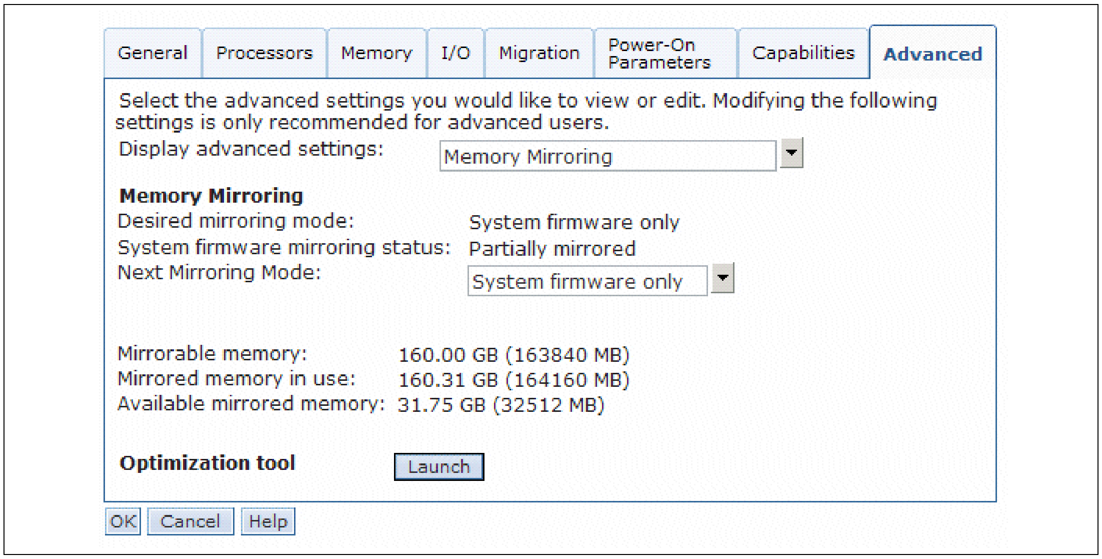
Active Memory Mirroring (AMM) is a chargeable feature on the Power E850. It can be ordered as part of the initial order, or as an MES upgrade later by using feature code #EM81. Once licensed, it can be enabled, disabled, or re-enabled depending on the user’s requirements.

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

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

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

It is possible to check whether the Active Memory Mirroring option is enabled and change its status through Hardware Management Console (HMC), under the Advanced tab on the System Properties panel (Figure 2-14).



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

Display advanced settings: Memory Mirroring

Memory Mirroring

Desired mirroring mode: System firmware only
System firmware mirroring status: Partially mirrored
Next Mirroring Mode: System firmware only

Mirrorable memory: 160.00 GB (163840 MB)
Mirrored memory in use: 160.31 GB (164160 MB)
Available mirrored memory: 31.75 GB (32512 MB)

Optimization tool

Figure 2-14 System Properties panel on an HMC

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

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

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

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

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

2.5.7 Memory Error Correction and Recovery

There are many features within the Power E850 memory subsystem that are designed to reduce the risk of errors, or to minimize the impact of any errors that do occur. This ensures that those errors do not have an impact on critical enterprise data.

Each memory chip has error detection and correction circuitry built in, which is designed so that the failure of any one specific memory module within an ECC word can be corrected without any other fault.

In addition, a spare dynamic random access memory (DRAM) per rank on each memory CDIMM provides for dynamic DRAM device replacement during runtime operation. Also, dynamic lane sparing on the memory link allows for replacement of a faulty data lane without impacting performance or throughput.

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

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

For more details about Memory RAS, see 4.3.10, “Memory protection” on page 156.

2.5.8 Special Uncorrectable Error handling

Special Uncorrectable Error (SUE) handling prevents an uncorrectable error in memory or cache from immediately causing the system to terminate. Rather, the system tags the data and determines whether it will ever be used again. If the error is irrelevant, it does not force a checkstop. If the data is used, termination can be limited to the program/kernel or hypervisor owning the data, or can freeze the I/O adapters controlled by an I/O hub controller if data is to be transferred to an I/O device.

2.6 Capacity on Demand

Several types of Capacity on Demand (CoD) offerings are optionally available on the Power E850 server to help meet changing resource requirements in an on-demand environment, by using resources that are installed on the system but that are not activated.

Hardware Management Consoles (HMCs): The Power E850 does not require an HMC for management, however most Capacity on Demand capabilities require an HMC to manage them. Capacity Upgrade on Demand (CUoD) does not require an HMC because permanent activations can be enabled through the ASMI menus.

2.6.1 Capacity Upgrade on Demand

A Power E850 server includes a number of active processor cores and memory units. It can also include inactive processor cores and memory units. Active processor cores or memory units are processor cores or memory units that are already available for use on your server when it comes from the manufacturer. Inactive processor cores or memory units are processor cores or memory units that are installed in your server, but not available for use until you activate them. Inactive processor cores and memory units can be permanently activated by purchasing an activation feature called *Capacity Upgrade on Demand* (CUoD) and entering the provided activation code on your server.

With the CUoD offering, you can purchase additional processor or memory capacity in advance at a low cost, and dynamically activate them when needed, without requiring that you restart your server or interrupt your business. All the processor or memory activations are restricted to the single server they are licensed to, and cannot be transferred to another system.

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

By using Capacity Upgrade on Demand the company could start with a system with enough installed resources to support the whole project lifecycle but only with enough active resources necessary for the initial project phases. More resources could be activated as the project continues, adjusting the hardware platform with the project needs. This would allow the company to reduce the initial investment in hardware and only acquire software licenses that are needed on each project phase, reducing the Total Cost of Ownership and Total Cost of Acquisition of the solution. Figure 2-15 shows a comparison between two scenarios: A fully activated system versus a system with CUoD resources being activated along with the project timeline.

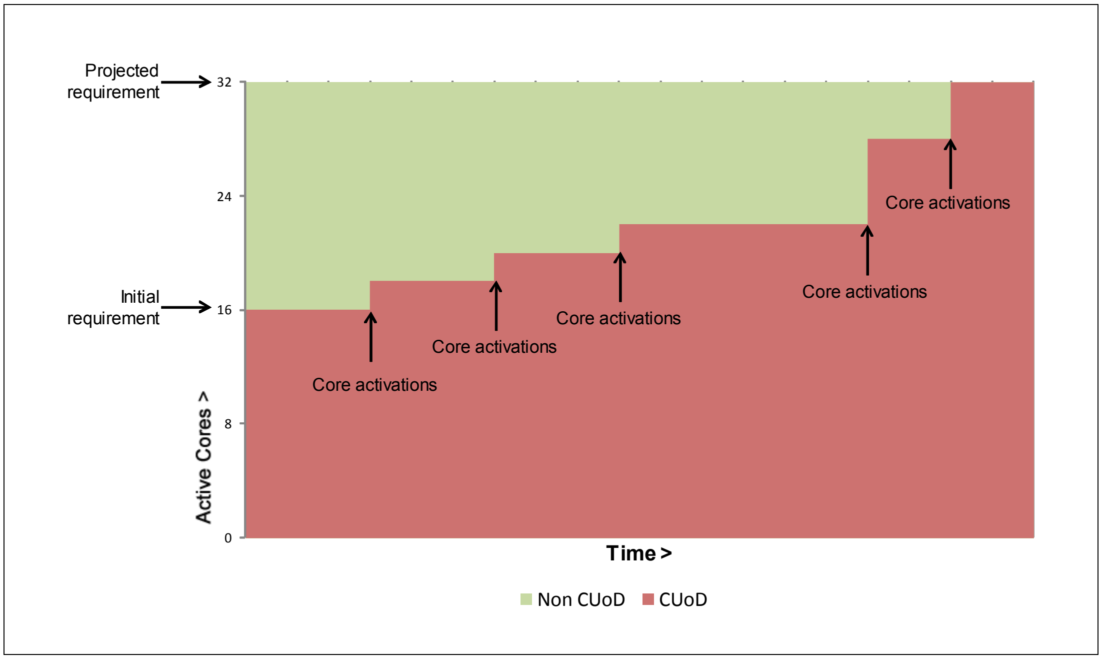


Figure 2-15 An example of Capacity Upgrade on Demand usage

Table 2-16 lists the processor activation features for the Power E850 server. Each feature code activates a single processor core. You cannot activate more processor cores than are physically installed in the server.

Table 2-16 Permanent processor activation codes for Power E850

Processor module	Processor module feature code	Permanent activation feature code
12 core 3.02 GHz	#EPV4	#EPVD
10 core 3.35 GHz	#EPV6	#EPVH
8 core 3.72 GHz	#EPV2	#EPV9

Permanent activations for memory features in the Power E850 can be ordered by using the following feature codes:

- #EMAA for 1 GB activation of memory
- #EMAB for 100 GB activation of memory

These feature codes are the same regardless of the capacity of the memory CDIMMs installed in the system. You cannot activate more memory capacity than is physically installed in the server.

2.6.2 Elastic Capacity on Demand

> **Note:** Some websites or documents still refer to Elastic Capacity on Demand (Elastic CoD) 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 enables processors or memory to be temporarily activated in one day increments as needed. These activations cover a period of 24 hours from the time the activation is enabled on the system. When you order an Elastic CoD feature, you receive an enablement code that allows a system operator to make requests for additional processor and memory capacity in increments of one processor day or 1 GB memory day. The system monitors the amount and duration of the activations. Both prepaid and postpay options are available.

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

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

Before using temporary capacity on your server, you must enable your server. To enable, an enablement feature (MES only) must be ordered and the required contracts must be in place. The feature codes are #EP9T for processor enablement and #EM9T for memory enablement.

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

- Enablement

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

  A *processor enablement* code (#EP9T) allows you to request up to 90 processor days of temporary capacity for each inactive processor core within the server. For instance, if you have 32 processor cores installed, and 16 are permanently activated, you would have 16 inactive processor cores. You would therefore receive enablement for (16 x 90) = 1440 processor days of elastic CoD. If the 90 processor-day limit is reached, place an order for another processor enablement code to reset the number of days that you can request back to 90 per inactive processor core.

  A *memory enablement* code (#EM9T) lets you request up to 90 memory days of temporary capacity for each GB of inactive memory within the server. For instance, if you have 256 GB of memory installed in the system, and 156 GB is permanently activated, you would have 100 GB of inactive memory. You would therefore receive an enablement code for (100 x 90) = 9000 GB days of elastic CoD. If you reach the limit of 90 memory days, place an order for another memory enablement code to reset the number of allowable days you can request back to 90.

- Activation requests

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

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

- Billing

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

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

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

HMC requirement: Elastic Capacity on Demand requires that an HMC is used for management of the Power E850 server.

2.6.3 Utility Capacity on Demand

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

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

If additional workload requires a higher level of performance, the system automatically allows the additional Utility CoD processors to be used, and the system automatically and continuously monitors and charges for the performance needed above the base (permanently activated) 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 Enterprise Edition to be active on the Power E850 system.

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

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

HMC requirement: Utility Capacity on Demand requires that an HMC is used for management of the Power E850 server.

2.6.4 Trial Capacity on Demand

A *standard request* for Trial Capacity on Demand (Trial CoD) requires you to complete a form including contact information and vital product data (VPD) from your Power E850 server with inactive CoD resources.

A standard request activates eight processors or 64 GB of memory (or both eight processors 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 E850 server 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.

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

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

2.6.5 Software licensing and Capacity on Demand

Although Capacity on Demand orders are placed using a hardware feature code, some CoD offerings include entitlement to certain IBM Systems Software and operating systems for the temporarily enabled capacity. Any other software products, including IBM applications and middleware, are not included in this cost. Check with your software vendor to find out how they charge for temporarily activated resources.

Capacity Upgrade on Demand permanently activates processor cores on the server. As such, this is seen as new permanent capacity, and licenses need to be purchased for any operating system and systems software running on this capacity. The activation code provided is for the hardware resources only.

Elastic CoD, Utility CoD, and Trial CoD activations include incremental licensing for the following IBM Systems Software and operating systems:

- AIX
- PowerVM
- PowerVC
- IBM PowerVP™
- IBM Cloud Manager with OpenStack
- IBM PowerHA®
- IBM PowerSC™
- Cluster Systems Management (CSM)
- IBM General Parallel File System (GPFS™)

Other IBM Systems Software or operating systems might also be included in the activations. Linux operating systems licenses may already cover the additional capacity, dependent on the licensing metric used. Check with your distributor for more details.

> **Note:** CoD does not ship any software or provide the base licensing entitlement. The software has to be initially installed and licensed on the server before temporary CoD provides the incremental licensing to cover the additional processor cores, which have been temporarily activated.

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*:

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

2.6.6 Integrated Facility for Linux

When running Linux workloads on a Power E850 server, it is possible to reduce the overall cost of processor and memory activations by ordering the Integrated Facility for Linux (IFL) package (#ELJN). This chargeable option provides four processor activations, 32 GB of memory activations and four licenses of PowerVM for Linux. This package is at a lower overall cost than ordering the same number of processor, memory, and PowerVM activations separately. However, the activated resources may only be used to run Linux server-based and VIO server-based workloads.

Any combination of Linux distributions supported by PowerVM can be used in logical partitions that utilize these resources. Any partitions running AIX cannot use these resources, and shared processor pools need to be set up on the system to prevent this. The Linux partitions can use any other resources permanently activated on the system, or that are temporarily activated using Elastic CoD or Utility CoD.

IFL packages can be ordered as part of the initial system order, or as an MES upgrade. These activations are permanent, and can count towards the minimum activations required for the system.

It is possible to use only IFL packages for permanent activations, in which case the entire system will be restricted to running only Linux workloads.

When ordering an Integrated Facility for Linux package (#ELJN), the following no-cost features are added to the system:

- 1 x four processor activations (#ELJK, #ELJL, or #ELJN depending on processor module)
- 1 x 32 GB memory activation (#ELJP)
- 1 x four PowerVM for Linux entitlements (#ELJQ)

2.7 System buses

This section provides additional information related to the internal buses.

2.7.1 PCI Express Generation 3

The internal I/O subsystem in the Power E850 is connected directly to the PCIe controllers on the POWER8 processor modules. Each POWER8 processor module has four buses, two of which have 16 lanes (x16) and two of which have eight lanes (x8). This gives a total of 48 lanes of PCIe connectivity per processor node. Each lane runs at 1 GBps in each direction, giving a total maximum PCIe bandwidth per processor module of 96 GBps.

The first and second processor modules in a system use all of their available PCIe connectivity. The third and fourth processor modules utilize only the two x16 connections at this time. The maximum PCIe bandwidth of a Power E850 server with four processor modules installed is therefore 320 GBps.

2.7.2 PCIe logical connectivity

Figure 2-16 shows how the internal PCIe Gen3 adapter slots are connected logically to the processor modules in the system. Each processor module supports two PCIe Gen3 x16 slots. Processor module 0 also supports a single PCIe Gen3 x8 slot, which is used for the default LAN adapter. The other x8 connection is used by the RAID adapter in the system. Processor module 1 supports two x8 slots connected through a PCIe Gen3 switch, along with the internal USB 3.0 host adapter for the four USB ports on the server (two rear, two front). These adapters therefore share the bandwidth of the x8 connection from the processor module.

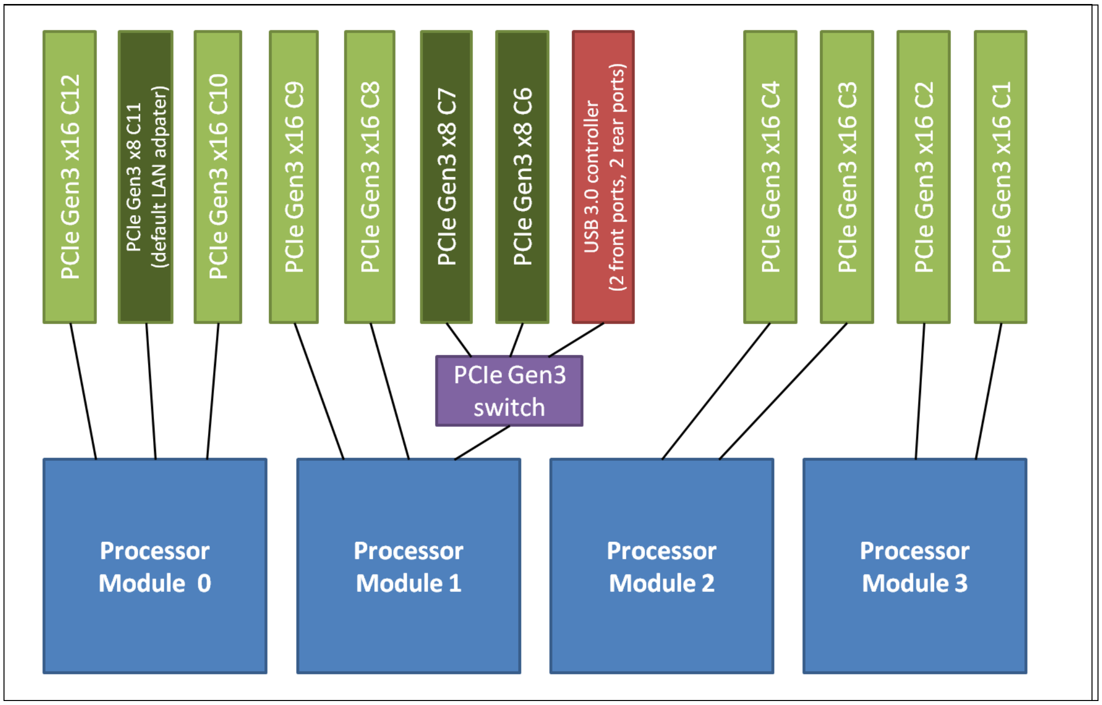


Figure 2-16 PCIe connectivity for the Power E850 server

A Power E850 server with fewer than four processor modules installed will not support all of the PCIe slots inside the system enclosure. The third and fourth processor module each support two of the PCIe Gen3 x16 slots. As such, the maximum number of adapters supported and the total PCIe bandwidth available varies by configuration.

All PCIe Gen3 slots in the server are hot-plug compatible for concurrent maintenance and repair. These procedures should be initiated through the HMC or ASMI menus.

Where the number of PCIe adapters in a server is important, the Power E850 supports the external PCIe Gen3 I/O expansion drawer (#EMX0). This contains two Fan-Out modules, each supports six PCIe adapter slots (four x8 slots and two x16 slots). These Fan-Out modules connect via optical cables to a PCIe Optical Cable Adapter (#EJ08) card, which is placed in a PCIe Gen3 x16 adapter slot in the server.

Each processor module in the Power E850 server supports up to two PCIe Optical Cable Adapters, and therefore up to two Fan-Out modules. This means that each processor module can support up to one full PCIe Gen3 I/O expansion drawer, giving a total of 12 PCIe Gen3 slots. All of the adapters connected to a Fan-Out module share the bandwidth of the single x16 slot in the server. For more information about the PCIe Gen3 I/O expansion drawer, see 2.11.1, “PCIe Gen3 I/O expansion drawer” on page 76.

Table 2-17 summarizes the maximum numbers of adapters and bandwidths for different configurations of the Power E850 server.

Table 2-17 Maximum PCIe adapters and bandwidth supported on the Power E850

Processor Modules	Maximum PCIe bandwidth	PCIe slots in server enclosure	Maximum I/O expansion drawers	Maximum PCIe slots supported
Two	192 GBps	7	2	27
Three	256 GBps	9	3	39
Four	320 GBps	11	4	51

2.8 Internal I/O connections

The internal I/O subsystem resides on the I/O planar, which supports all of the PCIe Gen3 x16 and x8 slots. All PCIe slots are hot-pluggable and enabled with enhanced error handling (EEH). In the unlikely event of a problem, EEH-enabled adapters respond to a special data packet that is generated from the affected PCIe slot hardware by calling system firmware, which examines the affected bus, allows the device driver to reset it, and continues without a system reboot. For more information about RAS on the I/O buses, see 4.3.11, “I/O subsystem availability and Enhanced Error Handling” on page 157.

All of the PCIe slots within the system enclosure support full-height half-length PCIe adapter cards. These can be PCIe Gen1, Gen2, or Gen3 adapters. The server also supports full-height full-length audiotapes in the I/O expansion drawer. For more information, see 2.11.1, “PCIe Gen3 I/O expansion drawer” on page 76.

Table 2-18 lists the slot configuration of the Power E850 server.

Table 2-18 Slot configuration and capabilities

Slot	Location code	Slot type	CAPI capable[a]	SRIOV capable
Slot 1[b]	P1-C1	PCIe Gen3 x16	Yes	Yes
Slot 2[bc]	P1-C2	PCIe Gen3 x16	Yes	Yes
Slot 3[d]	P1-C3	PCIe Gen3 x16	Yes	Yes
Slot 4[dc]	P1-C4	PCIe Gen3 x16	Yes	Yes
FSP slot[e]	N/A[e]	N/A[e]	N/A[e]	N/A[e]
Slot 5	P1-C6	PCIe Gen3 x8	No	Yes
Slot 6	P1-C7	PCIe Gen3 x8	No	Yes
Slot 7[c]	P1-C8	PCIe Gen3 x16	Yes	Yes
Slot 8	P1-C9	PCIe Gen3 x16	Yes	Yes
Slot 9[c]	P1-C10	PCIe Gen3 x16	Yes	Yes
Slot 10[f]	P1-C11	PCIe Gen3 x8	No	Yes
Slot 11	P1-C12	PCIe Gen3 x16	Yes	Yes

[a] At the time of writing, there are no supported CAPI adapters for the Power E850 server.
[b] Slots 1 and 2 are only active when the fourth processor module is installed.
[c] This slot is capable of supporting 75 W adapters.

d. Slots 3 and 4 are only active when the third processor module is installed.

e. The space for slot P1-C5 is used for the FSP card and connections, so it is not a usable PCIe slot.

f. Slot 10 (P1-C11) is used by the default LAN adapter card, which is required for manufacturing and testing.

Figure 2-16 on page 60 shows the physical and logical placement of the slots.

Table 2-19 shows the priorities for the PCIe adapter slots in the Power E850.

Table 2-19 Adapter slot priorities for the Power E850

Configuration	Two processor modules installed	Three processor modules installed	Four processor modules installed
PCIe slot priority	10, 9, 7, 11, 8, 5, 6	10, 9, 7, 4, 11, 8, 3, 5, 6	10, 9, 7, 4, 2, 11, 8, 3, 1, 5, 6

2.8.1 System ports

The Power E850 server has a number of system ports, which are used for management of the server. These are not accessible from the host operating system, instead they connect to the Flexible System Processor (FSP). The system ports sit within the space allocated for C5 on the backplane of the system enclosure. The following system ports can be found:

- 2 x USB 2.0 ports

  These can be used to connect uninterruptible power supplies (UPSs), as well as for FSP installation and code updates if necessary. These USB 2.0 ports are not available to host operating systems. They are only accessible by the service processor.

- 2 x RJ-45 Ethernet ports

  These are used for management connectivity to the system processor. They can be used to connect to the ASMI menus through a web interface, or can be used for connection to a Hardware Management Console (HMC) if applicable. Each Ethernet port has a unique MAC address, and both can be assigned different Ethernet addresses on two distinct subnets. One Ethernet port is used as the primary connection, and the second Ethernet port can be used for redundancy.

- 1 x serial port

  This RJ-45 port can be converted to a standard serial connection using an optional adapter (#3930). This allows management of the system through a standard serial connection.

2.9 PCI adapters

This section covers the types and functions of the PCI cards supported by the Power E850 server.

2.9.1 PCI Express

PCI Express (PCIe) uses a serial interface and allows for point-to-point interconnections between devices (using a directly wired interface between these connection points). A single PCIe serial link is a dual-simplex connection that uses two pairs of wires, one pair for transmit and one pair for receive, and can transmit only one bit per cycle. These two pairs of wires are called a *lane*. A PCIe link can consist of multiple lanes. In such configurations, the connection is labeled as x1, x2, x8, x12, x16, or x32, where the number is effectively the number of lanes.

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

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

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

- PCIe low profile (LP) cards, which are not used with the Power E850 server.
- PCIe full height cards, which are used in the Power E850 server and the PCIe Gen3 I/O expansion drawer (#EMX0).

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

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

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

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

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

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

> **Note:** PCIe full height and full high cards are used in the Power E850 server and any attached PCIe Gen3 I/O expansion drawer (#EMX0).

2.9.2 LAN adapters

To connect the Power E850 server to a local area network (LAN), you can use the LAN adapters that are supported in the PCIe slots of the system. Table 2-20 lists the available LAN adapters. Information about Fibre Channel over Ethernet (FCoE) adapters can be found in Table 2-24 on page 66.

Table 2-20 Available LAN adapters

Feature code	CCIN	Description	Max per system	OS support
5744	2B44	PCIe2 4-Port 10 GbE&1 GbE SR&RJ45 Adapter	48	Linux
5767	5767	2-Port 10/100/1000 Base-TX Ethernet PCI Express Adapter	50	AIX, Linux

Feature code	CCIN	Description	Max per system	OS support
5768	5768	2-Port Gigabit Ethernet-SX PCI Express Adapter	48	AIX, Linux
5769	5769	10 Gigabit Ethernet-SR PCI Express Adapter	48	AIX, Linux
5899	576F	PCIe2 4-port 1 GbE Adapter	50	AIX, Linux
EC2N		PCIe3 2-port 10 GbE NIC&RoCE SR Adapter	50	AIX, Linux
EC38		PCIe3 2-port 10 GbE NIC&RoCE SFP+ Copper Adapter	50	AIX, Linux
EC3B	57B6	PCIe3 2-Port 40 GbE NIC RoCE QSFP+ Adapter	50	AIX, Linux
EN0S	2CC3	PCIe2 4-Port (10Gb+1 GbE) SR+RJ45 Adapter	50	AIX, Linux
EN0U	2CC3	PCIe2 4-port (10Gb+1 GbE) Copper SFP+RJ45 Adapter	50	AIX, Linux
EN0W	2CC4	PCIe2 2-port 10/1 GbE BaseT RJ45 Adapter	50	AIX, Linux
EN15	2CE3	PCIe3 4-port 10 GbE SR Adapter	50	AIX, Linux
EN17	2CE4	PCIe3 4-port 10 GbE SFP+ Copper Adapter	50	AIX, Linux

2.9.3 Graphics accelerator adapters

Table 2-21 lists the available graphics accelerator adapters. An adapter can be configured to operate in either 8-bit or 24-bit color modes. The adapter supports both analog and digital monitors.

Table 2-21 Available graphics accelerator adapters

Feature Code	Description	Max per system	OS support
5748	POWER GXT145 PCI Express Graphics Accelerator	9	AIX, Linux
EC42	PCIe2 3D Graphics Adapter x1	10	AIX, Linux

2.9.4 SAS adapters

Table 2-22 lists the SAS adapters that are available for the Power E850 server.

Table 2-22 Available SAS adapters

Feature code	CCIN	Description	Max per system	OS support
5901	57B3	PCIe Dual-x4 SAS Adapter	50	AIX, Linux
EJ0J	57B4	PCIe3 RAID SAS Adapter Quad-port 6Gb x8	34	AIX, Linux
EJ0L	57CE	PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6Gb x8	34	AIX, Linux
EJ10	57B4	PCIe3 SAS Tape/DVD Adapter Quad-port 6Gb x8	34	AIX, Linux
EJ14	57B1	PCIe3 12GB Cache RAID PLUS SAS Adapter Quad-port 6Gb x8	24	AIX, Linux

2.9.5 Fibre Channel adapter

The Power E850 supports direct or SAN connection to devices that use Fibre Channel adapters. Table 2-23 summarizes the available Fibre Channel adapters, which all have LC connectors.

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

Table 2-23 Available Fibre Channel adapters

Feature Code	CCIN	Description	Max per system	OS support
5729		PCIe2 8 Gb 4-port Fibre Channel Adapter	50	AIX, Linux
5735	577D	8 Gigabit PCI Express Dual Port Fibre Channel Adapter	50	AIX, Linux
EN12		PCIe2 8Gb 4-port Fibre Channel Adapter		AIX, Linux
EN0A	577F	PCIe2 16Gb 2-port Fibre Channel Adapter	50	AIX, Linux
EN0G		PCIe2 8Gb 2-Port Fibre Channel Adapter	50	AIX, Lunux

2.9.6 Fibre Channel over Ethernet

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

Figure 2-17 compares existing Fibre Channel and network connections and FCoE connections.

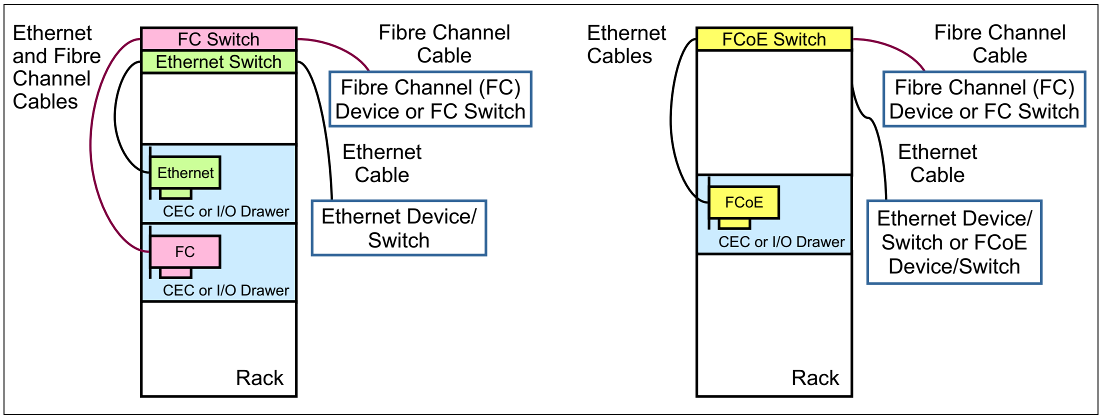


Figure 2-17 Comparison between existing FC and network connections and FCoE connections

Table 2-24 on page 66 lists the available FCoE adapters. They are high-performance Converged Network Adapters (CNAs) using SR optics. Each port can simultaneously provide network interface card (NIC) traffic and Fibre Channel functions.

Table 2-24 Available FCoE adapters

Feature Code	CCIN	Description	Max per system	OS support
EN0H	2B93	PCIe2 4-port (10Gb FCoE & 1 GbE) SR&RJ45	50	AIX, Linux
EN0K	2CC1	PCIe2 4-port (10Gb FCoE & 1 GbE) SFP+Copper&RJ45	50	AIX, Linux
EN0M	2CC0	PCIe2 4-port(10Gb FCoE & 1 GbE) LR&RJ45 Adapter	50	AIX, Linux

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

Note: Adapters #EN0H, #EN0K, and #EN0M support SR-IOV when minimum firmware and software levels are met. See 3.4, “Single root I/O virtualization” on page 114 for more information.

2.9.7 InfiniBand Host Channel adapter

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

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

http://www.infinibandta.org

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

The IB 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 - 120 Gbps. The architecture defines a layered hardware protocol and also a software layer to manage initialization and the communication between devices. Each link can support multiple transport services for reliability and multiple prioritized virtual communication channels.

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

A connection to supported IB switches is accomplished by using the QDR optical cables #3290 and #3293.

Table 2-25 lists the available IB adapter.

Table 2-25 Available IB adapter

Feature code	CCIN	Description	Max per system	OS support
5285	58E2	PCIe2 2-Port 4X IB QDR Adapter 40Gb	10	AIX, Linux

2.9.8 Asynchronous and USB adapters

Asynchronous PCIe adapters provide the connection of asynchronous EIA-232 or RS-422 devices. If you have a cluster configuration or high-availability configuration and plan to connect the IBM Power Systems using a serial connection, you can use the features that are listed in Table 2-26.

Table 2-26 Available asynchronous and USB adapters

Feature code	CCIN	Description	Max per system	OS support
5785	57D2	4 Port Async EIA-232 PCIe Adapter	9	AIX, Linux
EN27		PCIe 2-Port Async EIA-232 adapter	10	AIX, Linux
EC46		PCIe2 LP 4-Port USB 3.0 Adapter	30	AIX, Linux

2.9.9 Cryptographic coprocessor

The cryptographic coprocessor cards that are supported for the Power E850 are shown in Table 2-27.

Table 2-27 Available cryptographic coprocessor

Feature code	Description	Max per system	OS support
EJ27	PCIe Crypto Coprocessor No BSC 4765-001	10	AIX
EJ28	PCIe Crypto Coprocessor Gen3 BSC 4765-001	10	AIX

2.9.10 Flash storage adapters

The available flash storage adapters are shown in Table 2-28.

Table 2-28 Available flash storage adapters

Feature code	CCIN	Description	Max	OS support
EC55	58CB	PCIe3 1.6TB NVMe Flash Adapter	7	Linux
EC57	58CC	PCIe3 3.2TB NVMe Flash Adapter	7	Linux

2.10 Internal storage

There are three storage controller options for the Power E850 server. All of these options connect to the storage backplane that has front mounted storage bays, which include eight hot-plug Small Form Factor (SFF) disk bays and four 1.8-inch disk bays and one DVD bay. Figure 2-18 shows the positions of the different bays on the front of the Power E850 server.

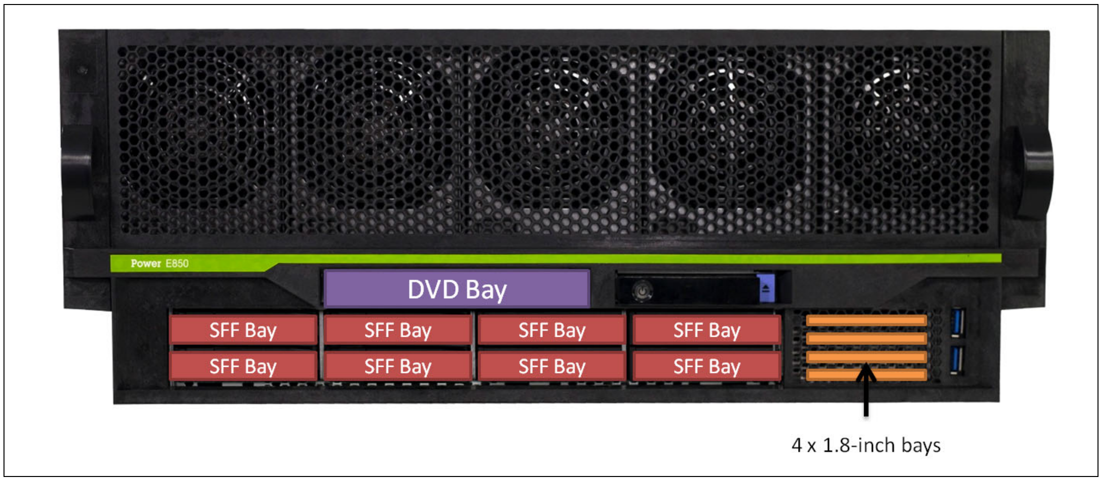


Figure 2-18 Disk bay positions on the front of the Power E850

2.10.1 Storage controller options

There are three storage controller options to choose from when configuring the Power E850 server. Following are the three choices:

- Dual controller disk backplane with write cache (#EPVN)

  The pair of controllers handles all 12 integrated disk bays and the DVD bay.
- Dual controller disk backplane without write cache (#EPVP)

  The pair of controllers handles all 12 integrated disk bays and the DVD bay.
- Split disk backplane (two single controllers) without write cache (#EPVQ)

  Each one of the two controllers handles four SFF disk bays and two 1.8-inch disk bays. One of the controllers handles the DVD bay.

None of the controller options provide any external SAS connections for further expansion. If you want to expand the storage capability of the server, you will need to add an EXP24S expansion drawer, using a PCIe SAS adapter. This expansion drawer handles further SFF hard disk drives (HDDs) and solid-state devices (SSDs). For more information about the EXP24S expansion drawer, see 2.12.1, “EXP24S SFF Gen2-bay Drawer” on page 82.

Dual controller disk backplane options

The dual controller disk backplane options provide both performance and protection advantages. Patented Active-Active capabilities enhance performance when there is more than one array configured. Each of the dual controllers has access to all the disk bays and can back up the other controller if there was a problem with the other controller. For the dual controller backplane with write cache, each controller mirrors the other’s write cache, providing redundancy protection.

The write-cache capability increases the speed of write operations by committing them to the write-cache flash memory first, and then writing to the disks attached to the controller. This allows write activities to be committed in a shorter time, reducing the latency of the write operation. When the write cache is full, write operation latency reverts to the speed of writes on the attached disks. The write cache has a raw capacity of 1.8 GB, but uses advanced compression techniques to store up to 7.2 GB of writes at a time. Integrated flash memory for the write-cache content provides protection against electrical power loss to the server and avoids the need for write cache battery protection and battery maintenance.

Clients with I/O performance-sensitive workloads with a large percentage of writes should consider using the dual controllers with write cache, or use PCIe SAS controllers with write cache to connect external storage. This is especially relevant for HDDs. Note also that RAID 5 and RAID 6 protection levels result in more drive write activity than mirroring or unprotected drives.

The dual controllers should be treated as a single resource, so both should be assigned to the same partition or VIOS. They both have access to all of the internal disks. If multiple arrays are configured on the internal disks, the controllers split primary responsibility for handling the arrays. If one of the dual controllers fails, the remaining controller takes over all work.

Split disk backplane option

The split disk backplane option has two independent controllers. One of these handles the top four SFF disk bays, the top two 1.8-inch disk bays, and the DVD bay. The other controller handles the bottom four SFF disk bays and the bottom two 1.8-inch disk bays. Figure 2-19 shows how the bays are split across the two controllers.

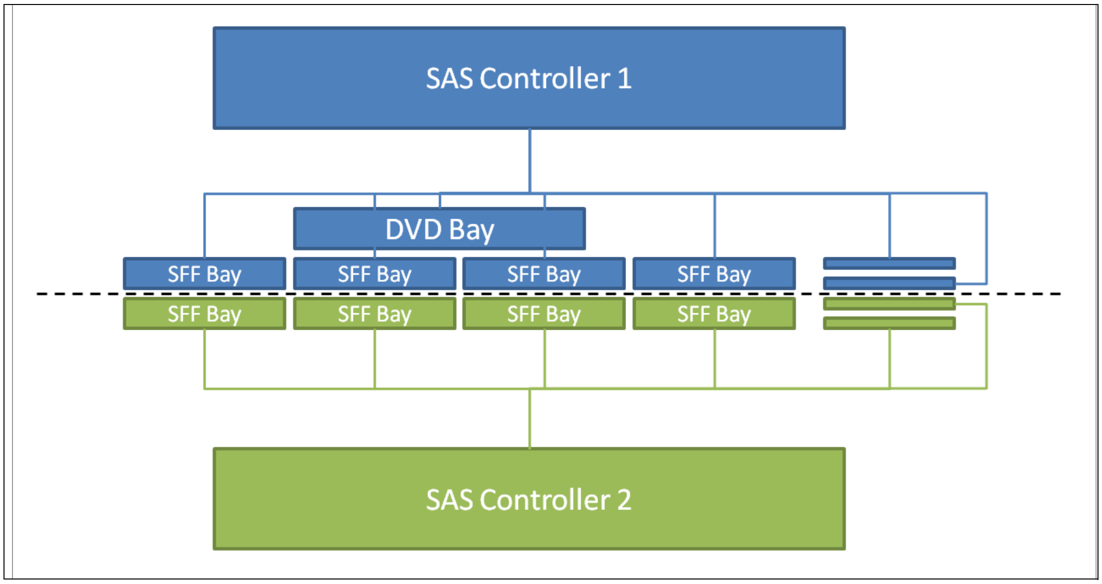


Figure 2-19 Controller to disk bay connections for split disk backplane option on Power E850

The independent split disk controllers should each be treated as a single resource. Each controller only has access to the six disk bays assigned to it, so it cannot take over control of any disks assigned to the other controller. If there is a failure of a disk controller, you might lose all access to a set of disks, and potentially to the DVD drive. This should be considered in the system planning phase. For instance, you might want to assign each controller to a VIOS partition, and then mirror the two VIOS for protection. Or assign each to the same partition and then mirror the two sets of drives.

2.10.2 RAID protection for internal disks

There are multiple protection options for HDD/SSD drives in the Power E850 server, whether they are contained in the internal SFF or 1.8-inch bays in the system unit or in disk-only I/O drawers like the EXP24S. Although protecting drives is always recommended, AIX and Linux users can choose to leave a few or all drives unprotected at their own risk, and IBM supports these configurations.

HDD/SSD drive protection can be provided by AIX and Linux, or by the available hardware controllers.

All three of these controller options can offer different drive protection options: RAID 0, RAID 5, RAID 6, or RAID 10. RAID 5 requires a minimum of three drives of the same capacity. RAID 6 requires a minimum of four drives of the same capacity. RAID 10 requires a minimum of two drives. Hot spare capability is supported by RAID 5 or RAID 6.

All three controller options offer Easy Tier functionality, which is also called *RAID 5T2* (2-tiered RAID 5), *RAID 6T2* (2-tiered RAID 6), and *RAID 10T2* (2-tiered RAID 10). The split disk backplane option supports RAID 10T2 but does not support RAID 5T2 or 6T2.

Table 2-29 details the drive protection options that are available with each storage controller option available with the Power E850 server.

Table 2-29 Drive protection capabilities of the Power E850 storage controller options

Protection type	Dual controller with write cache (#EPVN)	Dual controller without write cache (#EPVP)	Split disk backplane (#EPVQ)
JBOD	No	No	Yes
RAID 0/1	Yes	Yes	Yes
RAID 5/6/10	Yes	Yes	Yes
RAID 5T2 (Easy Tier)	Yes	Yes	No
RAID 6T2 (Easy Tier)	Yes	Yes	No
RAID 10T2 (Easy Tier)	Yes	Yes	Yes
Split backplane	No	No	Yes

2.10.3 Drive protection levels

The different available levels of drive protection are listed here:

- Just a bunch of disks (JBOD) provides no drive protection.

  JBOD presents the drives as just a bunch of disks to the system. The failure of a single drive results in the loss of all data on that disk. Any data protection must be provided through the operating system or software.

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

  The failure of a single drive results in the loss of all data on the array. This version of RAID increases I/O bandwidth by simultaneously accessing multiple data paths.

- RAID 1 provides mirroring for fault tolerance, but halves total drive capacity.

  The failure of a single drive results in no loss of data on the array. This version of RAID provides simple mirroring, requiring two copies of all data. However, this option has the highest cost because only half of the installed capacity is usable.

- RAID 5 uses block-level data striping with distributed parity.

  RAID 5 stripes both data and parity information across three or more drives. Fault tolerance is maintained by ensuring that the parity information for any given block of data is placed on a drive that is separate from the ones that are used to store the data itself. This version of RAID provides data resiliency if a single drive fails in a RAID 5 array.

- RAID 6 uses block-level data striping with dual distributed parity.

  RAID 6 is the same as RAID 5 except that it uses a second level of independently calculated and distributed parity information for additional fault tolerance. A RAID 6 configuration requires N+2 drives to accommodate the additional parity data, making it less cost-effective than RAID 5 for equivalent storage capacity. This version of RAID provides data resiliency if one or two drives fail in a RAID 6 array. When you work with large capacity disks, RAID 6 allows you to sustain data parity during the rebuild process.

- RAID 10 is a striped set of mirrored arrays.

  It is a combination of RAID 0 and RAID 1. A RAID 0 stripe set of the data is created across a two-disk array for performance benefits. A duplicate of the first stripe set is then mirrored on another two-disk array for fault tolerance. This version of RAID provides data resiliency if a single drive fails, and it can provide resiliency for multiple drive failures.

RAID 5T2, RAID 6T2, and RAID 10T2 are RAID levels with EasyTier enabled. It requires that both types of disks exist on the system under the same controller (HDDs and SSDs) and that both are configured under the same RAID type.

If Easy Tier functionality is not being used, an array can consist of only one drive type. So you can have HDD only arrays and SDD only arrays.

It is possible to mix drive capacities in an array. However, if an array has multiple capacity points, the utilized capacity might be lower. For example, with a mix of 300 GB HDDs and 600 GB HDDs, only 300 GB of the larger 600 GB HDDs will be used. Similarly, if an array has both 387 GB SSDs and 775 GB SSDs, only 387 GB of the 775 GB SSDs will be used.

Note: The block size of the drives in an array must match, covering both HDDs and SSDs. So either all drives must be formatted with a 4 k block size, or all must be formatted with a 5 xx block size.

2.10.4 Easy Tier

The Power E850 server supports both HDDs and SSDs in the SFF bays connected to the storage backplane and accessible from the front of the server. The 1.8-inch bays support only SSDs. It is possible to create multiple arrays using any of the storage controller options, however you cannot mix HDDs and SSDs in a standard (non-Easy Tier) array. Instead, you would need to create separate arrays for HDDs and SSDs.

All of the storage controller options also support Easy Tier functionality, which allows them to support mixed arrays of HDDs and SSDs. When the SSDs and HDDs are under the same array, the adapter can automatically move the most accessed data to faster storage (SSDs) and less accessed data to slower storage (HDDs).

There is no need for coding or software intervention after the RAID is configured correctly. Statistics on block accesses are gathered every minute, and after the adapter realizes that some portion of the data is being frequently requested, it moves this data to faster devices. The data is moved in chunks of 1 MB or 2 MB called *bands*.

By moving hot data onto faster SSDs, Easy Tier can dramatically improve the performance of applications that are limited by disk I/O operations. Cold data blocks are moved onto more cost effective HDDs, reducing the overall cost. Combining this with the dual controller storage backplane with write cache can provide even higher levels of overall storage performance.

From the operating system point-of-view, there is just a regular array disk. From the SAS controller point-of-view, there are two arrays with parts of the data being serviced by one tier of disks and parts by another tier of disks.

Figure 2-20 shows a representation of an Easy Tier array.

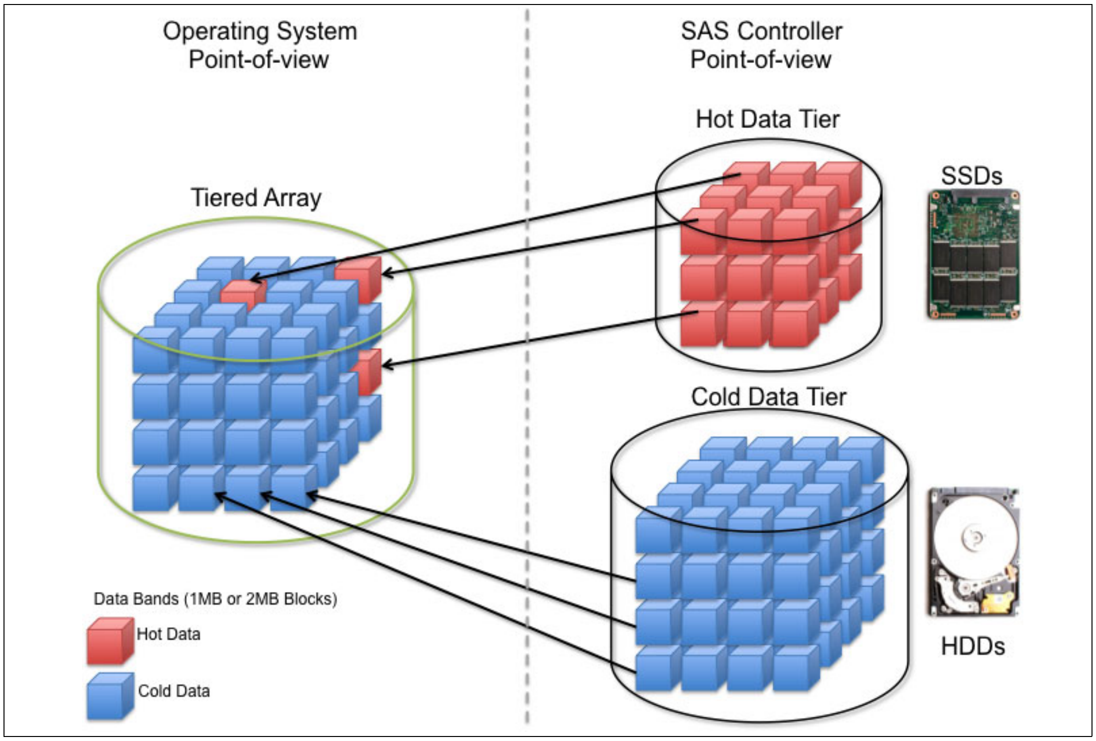
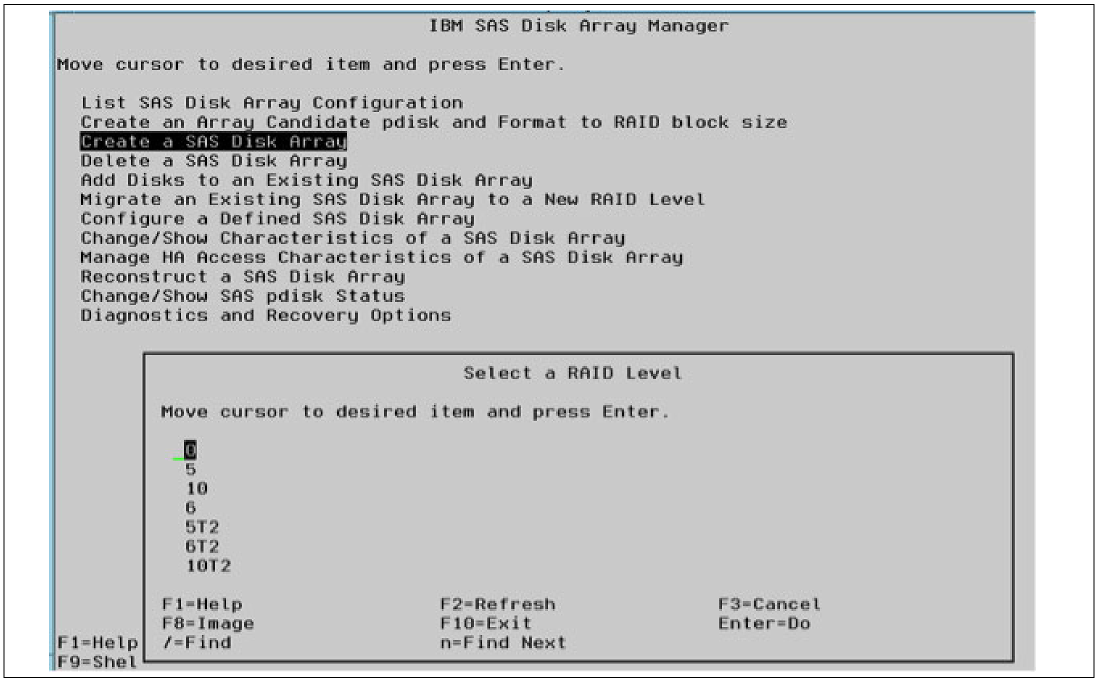

Operating System Point-of-view
SAS Controller Point-of-view
Hot Data Tier
SSDs
Tiered Array
Cold Data Tier
HDDs
Data Bands (1MB or 2MB Blocks)
Hot Data
Cold Data

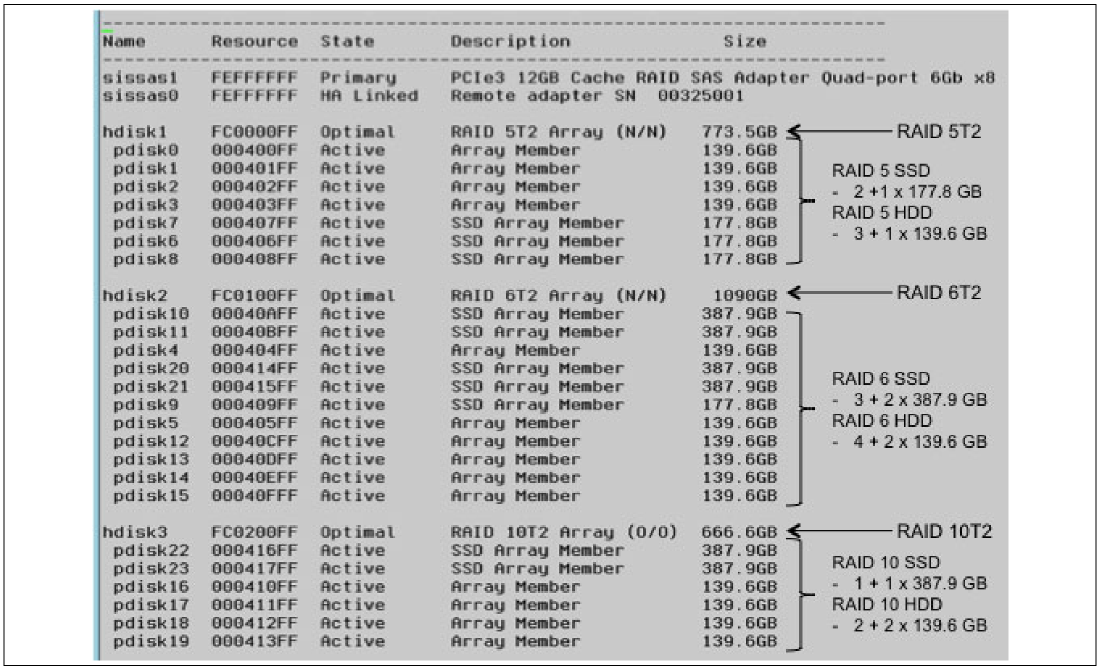
Figure 2-20 An Easy Tier array

The Easy Tier configuration is accomplished through a standard operating system SAS adapter configuration utility. Figure 2-21 and Figure 2-22 on page 74 show two examples of tiered array creation for AIX.



```
                               IBM SAS Disk Array Manager

Move cursor to desired item and press Enter.

  List SAS Disk Array Configuration
  Create an Array Candidate pdisk and Format to RAID block size
  Create a SAS Disk Array
  Delete a SAS Disk Array
  Add Disks to an Existing SAS Disk Array
  Migrate an Existing SAS Disk Array to a New RAID Level
  Configure a Defined SAS Disk Array
  Change/Show Characteristics of a SAS Disk Array
  Manage HA Access Characteristics of a SAS Disk Array
  Reconstruct a SAS Disk Array
  Change/Show SAS pdisk Status
  Diagnostics and Recovery Options

                         Select a RAID Level

         Move cursor to desired item and press Enter.

           0
           5
           10
           6
           5T2
           6T2
           10T2

         F1=Help                F2=Refresh             F3=Cancel
         F8=Image               F10=Exit               Enter=Do
F1=Help  /=Find                 n=Find Next
F9=Shel
```

Figure 2-21 Array type selection panel on AIX RAID Manager



```
-------------------------------------------------------------------------
Name        Resource  State       Description               Size
-------------------------------------------------------------------------
sissas1     FEFFFFFF  Primary     PCIe3 12GB Cache RAID SAS Adapter Quad-port 6Gb x8
sissas0     FEFFFFFF  HA Linked   Remote adapter SN  00325001

hdisk1      FC0000FF  Optimal     RAID 5T2 Array (N/N)     773.5GB
 pdisk0     000400FF  Active      Array Member             139.6GB
 pdisk1     000401FF  Active      Array Member             139.6GB
 pdisk2     000402FF  Active      Array Member             139.6GB
 pdisk3     000403FF  Active      Array Member             139.6GB
 pdisk7     000407FF  Active      SSD Array Member         177.8GB
 pdisk6     000406FF  Active      SSD Array Member         177.8GB
 pdisk8     000408FF  Active      SSD Array Member         177.8GB

hdisk2      FC0100FF  Optimal     RAID 6T2 Array (N/N)      1090GB
 pdisk10    00040AFF  Active      SSD Array Member         387.9GB
 pdisk11    00040BFF  Active      SSD Array Member         387.9GB
 pdisk4     000404FF  Active      Array Member             139.6GB
 pdisk20    000414FF  Active      SSD Array Member         387.9GB
 pdisk21    000415FF  Active      SSD Array Member         387.9GB
 pdisk9     000409FF  Active      SSD Array Member         177.8GB
 pdisk5     000405FF  Active      Array Member             139.6GB
 pdisk12    00040CFF  Active      Array Member             139.6GB
 pdisk13    00040DFF  Active      Array Member             139.6GB
 pdisk14    00040EFF  Active      Array Member             139.6GB
 pdisk15    00040FFF  Active      Array Member             139.6GB

hdisk3      FC0200FF  Optimal     RAID 10T2 Array (O/O)    666.6GB
 pdisk22    000416FF  Active      SSD Array Member         387.9GB
 pdisk23    000417FF  Active      SSD Array Member         387.9GB
 pdisk16    000410FF  Active      Array Member             139.6GB
 pdisk17    000411FF  Active      Array Member             139.6GB
 pdisk18    000412FF  Active      Array Member             139.6GB
 pdisk19    000413FF  Active      Array Member             139.6GB
```

Figure 2-22 Tiered arrays (RAID 5T2, RAID 6T2, and RAID 10T2) example on AIX RAID Manager

Each Easy Tier array is made up of two individual arrays of the same type, one populated with HDDs and one with SSDs. So a RAID 5T2 Array is formed of a RAID 5 array of HDDs and a RAID 5 array of SSDs. As such, following are the minimum device quantities required:

- RAID 5T2 requires at least three HDDs and three SSDs
- RAID 6T2 requires at least four HDDs and four SSDs
- RAID 10T2 requires at least two HDDs and two SSDs

The HDD and SSD can be different capacities in an Easy Tier array. However, if either half of the array has multiple capacity points, the utilized capacity might be lower. For example, with a mix of 300 GB HDDs and 600 GB HDDs, only 300 GB of the larger 600 GB HDDs are used. Similarly, if an array has both 387 GB SSDs and 775 GB SSDs, only 387 GB of the 775 GB SSDs are used. A combination of 600 GB HDDs and 387 GB SSDs will allow you to use the full capacity of all devices.

Note: The block size of the drives in an array must match, covering both HDDs and SSDs. So either all drives must be formatted with a 4 k block size, or all must be formatted with a 5 xx block size.

2.10.5 Internal disk options

Each of the controller options that are available provides a backplane with eight SFF-3 disk bays in the server. These 2.5-inch (SFF) SAS bays can support both HDDs and SSDs mounted in a Gen3 carrier. Previous generation SFF disk types (SFF-1 and SFF-2) do not fit in these bays. All SFF-3 bays support concurrent maintenance or hot-plug capability. All three of the controller options support HDDs or SSDs or a mixture of HDDs and SSDs in the SFF-3 bays. If mixing HDDs and SSDs, they must be in separate arrays (unless using the Easy Tier function).

The storage backplane also has four 1.8-inch storage bays. These can hold 1.8-inch SSDs to provide high performance storage for Easy Tier arrays.

Table 2-30 lists the drives that are available for the internal bays of the Power E850 server.

Table 2-30 Drives available for internal storage bays on the Power E850

Feature Code	Capacity	Type	Placement
ES0L	387 GB	SSD	SSF-3 bay
ES0N	775 GB	SSD	SSF-3 bay
ES0U	387 GB	4k block SSD	SFF-3 bay
ES0W	775 GB	4k block SSD	SFF-3 bay
ES0Y	177 GB	4k block Read intensive SSD	1.8-inch bay
ES0Z	177 GB	Read intensive SSD	1.8-inch bay
ES16	387 GB	SSD	1.8-inch bay
ES1C	387 GB	5xx SSD eMLC4	1.8-inch bay
ES2V	387 GB	4k SSD eMLC4	1.8-inch bay
ES2X	775 GB	5xx SSD eMLC4	1.8-inch bay
ES4K	775 GB	4k SSD eMLC4	1.8-inch bay
ES7K	387 GB	SSD 5xx eMLC4	SFF-3 bay
ES7P	775 GB	SSD 5xx eMLC4	SFF-3 bay
ES8J	1.9 TB	4k blovk Read Intensive SSD	SFF-3 bay
ES8N	387 GB	SSD 4k eMLC4	SFF-3 bay
ES8Q	775 GB	SSD 4k eMLC4	SFF-3 bay
ES8V	1.55 TB	SSD 4k eMLC4	SFF-3 bay
ESD5	600 GB	10k RPM HDD	SFF-3 bay
ESD9	1.2 TB	10k RPM HDD	SFF-3 bay
ESDB	300 GB	15k RPM HDD	SFF-3 bay
ESDF	600 GB	15k RPM HDD	SFF-3 bay
ESDR	300 GB	10k RPM HDD	SFF-3 bay
ESDT	146 GB	15k RPM HDD	SFF-3 bay
ESF5	600 GB	4k block 10k RPM HDD	SFF-3 bay
ESF9	1.2 TB	4k block 10k RPM HDD	SFF-3 bay
ESFB	300 GB	4k block 15k RPM HDD	SFF-3 bay
ESFF	600 GB	4k block 15k RPM HDD	SFF-3 bay
ESFV	1.8 TB	4k block 10k RPM HDD	SFF-3 bay

If you want to expand the storage capability of the server, you need to add an EXP24S expansion drawer, using a PCIe SAS adapter. This expansion drawer handles further SFF HDDs and SSDs. For more information about the EXP24S expansion drawer, see 2.12.1, “EXP24S SFF Gen2-bay Drawer” on page 82.

2.10.6 DVD drive

A slimline media bay is included with all of the controller options on the Power E850 server. This bay can hold the optional DVD-RAM drive (#5771). If using one of the two dual controller options, both integrated controllers can access the DVD device if fitted. If using the split disk backplane option, only one controller can access the DVD. This is the controller that has access to the top disks in the system.

A DVD drive can be included in the system, and is then available to perform operating system installation, maintenance, problem determination, and service actions, such as maintaining system firmware and I/O microcode at their latest levels. Alternatively, the system must be attached to a network with software such as AIX Network Installation Manager (NIM) or Linux Install Manager to perform these functions.

2.11 External I/O subsystems

This section describes the PCIe Gen3 I/O expansion drawer that can be attached to the Power E850 server.

2.11.1 PCIe Gen3 I/O expansion drawer

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

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

A PCIe x16 to Optical CXP converter adapter (#EJ08) and two 3.0 m (#ECC7), or two 10.0 m (#ECC8) CXP 16X Active Optical cables (AOC) connect the system node to a PCIe Fan Out module in the I/O expansion drawer. One feature #ECC7, or one #ECC8 ships two AOC cables.

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

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

A maximum of four PCIe Gen3 I/O drawers can be attached to the Power E850 server, if equipped with four processor modules. The maximum number of PCIe Gen3 I/O drawers depends on the number of installed processor modules. Table 2-31 lists the maximum PCIe Gen3 I/O drawer configurations for the Power E850 server.

Table 2-31 Maximum PCIe Gen3 I/O drawer configurations

Power E850 configuration	Maximum number of attached PCIe Gen3 I/O drawers
Power E850 with two processor modules	2
Power E850 with three processor modules	3
Power E850 with four processor modules	4

Figure 2-23 shows the back view of the PCIe Gen3 I/O expansion drawer.

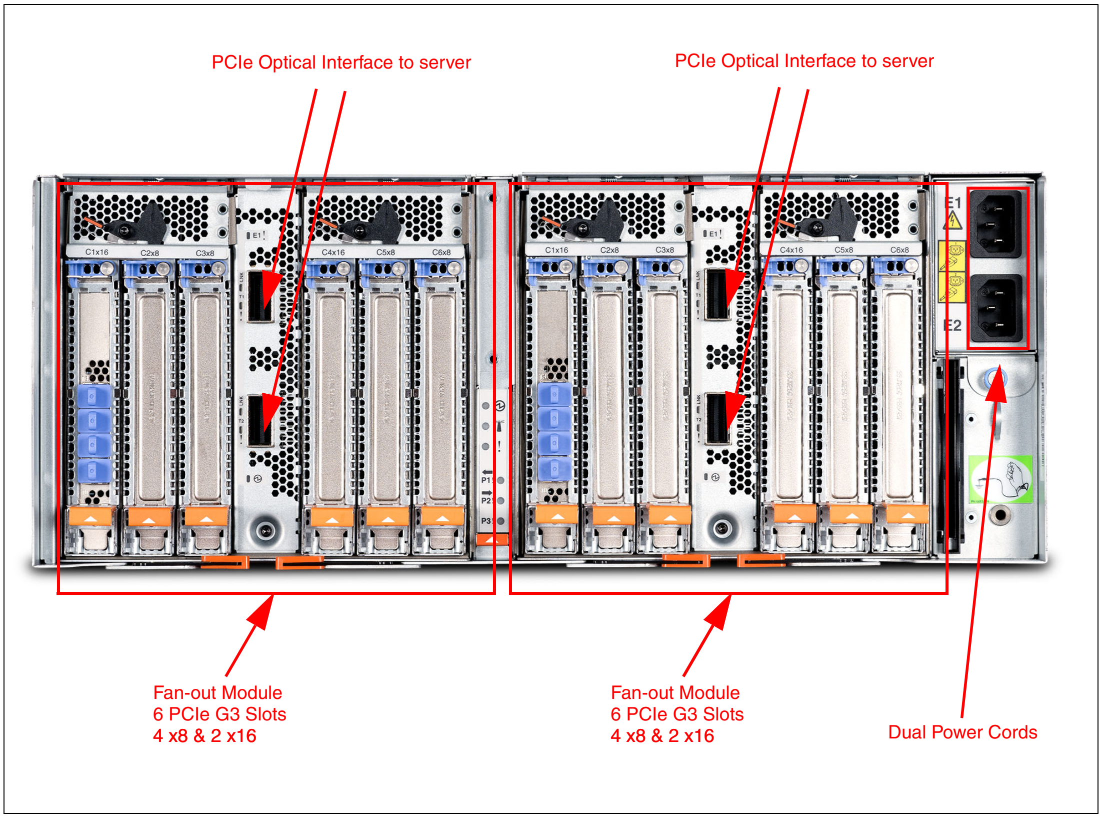


Figure 2-23 Rear view of the PCIe Gen3 I/O expansion drawer

2.11.2 PCIe Gen3 I/O expansion drawer optical cabling

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

- 3.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC7)
- 10.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC8)

Cable lengths: Use the 3.0 m cables for intra-rack installations. Use the 10.0 m cables for inter-rack installations.

A minimum of one PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ08) is required to connect to the PCIe3 6-slot Fan Out module in the I/O expansion drawer. The top port of the fan-out module must be cabled to the top port of the #EJ08 port. Likewise, the bottom two ports must be cabled together:

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

Figure 2-24 shows connector locations for the PCIe Gen3 I/O expansion drawer.

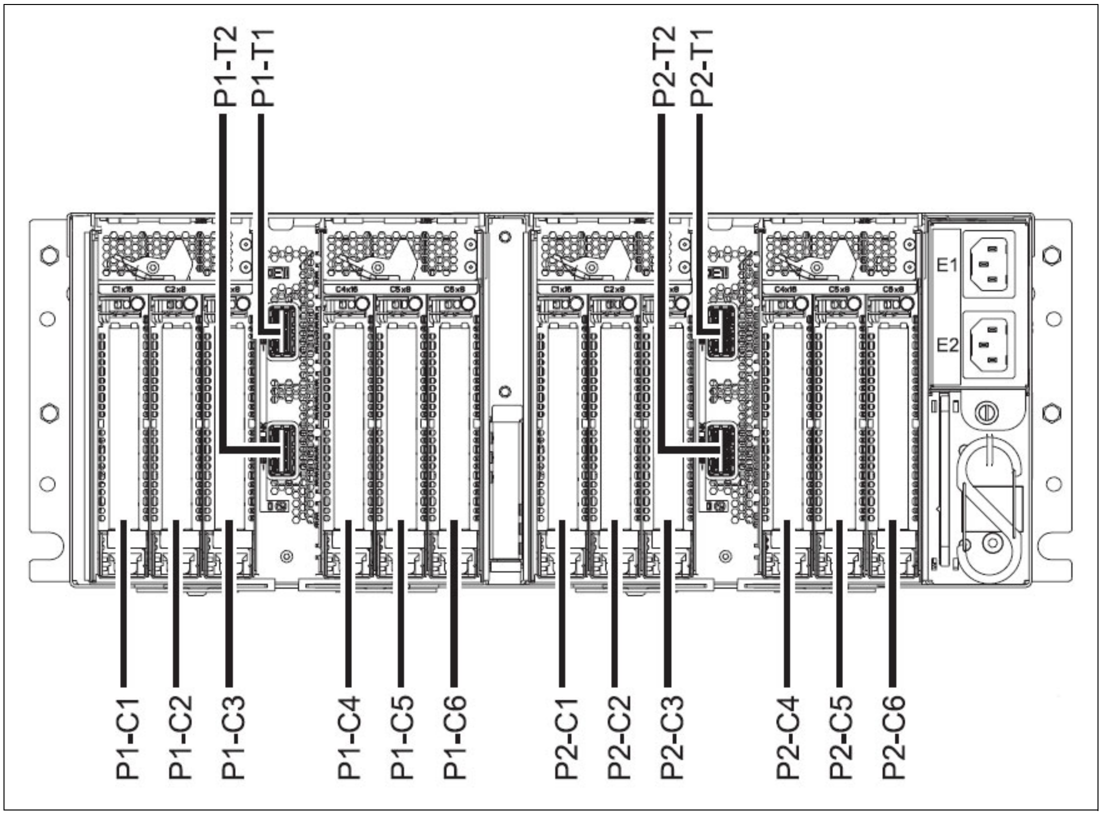


Figure 2-24 Connector locations for the PCIe Gen3 I/O expansion drawer

Figure 2-25 shows typical optical cable connections.

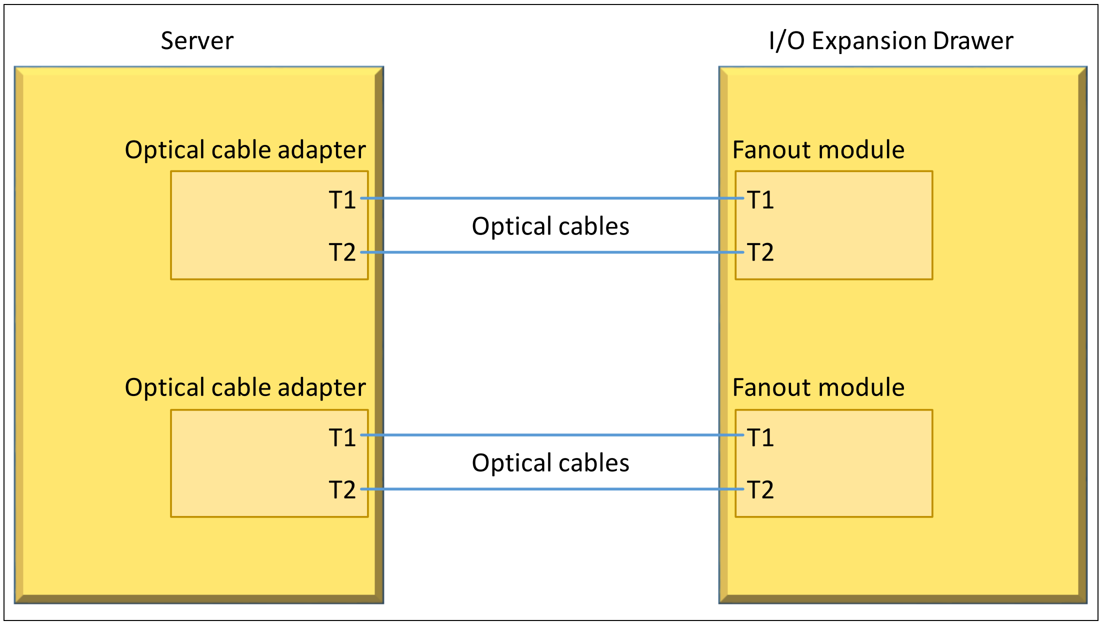


Figure 2-25 Typical optical cable connection

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

The PCIe3 optical cable adapter can be in any of the x16 PCIe Gen3 adapter slots in the Power E850 system node. However, we advise that you use the PCIe adapter slot priority information while selecting slots for installing PCIe3 Optical Cable Adapter (#EJ08).

Table 2-32 shows PCIe adapter slot priorities in the Power E850 server.

Table 2-32 PCIe adapter slot priorities

Feature code	Description	Slot priorities		
		Two processor modules	**Three processor modules**	**Four processor modules**
EJ08	PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer	9, 7, 11, 8[a]	9, 7, 3, 11, 8, 4[a]	9, 7, 3, 1, 11, 8, 4, 2[a]

[a] For information about how the slot numbers listed relate to physical location codes, see Table 2-18 on page 61.

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

Figure 2-26 shows an example of a Power E850 with two processor modules and a maximum of two PCIe Gen3 I/O expansion drawers.

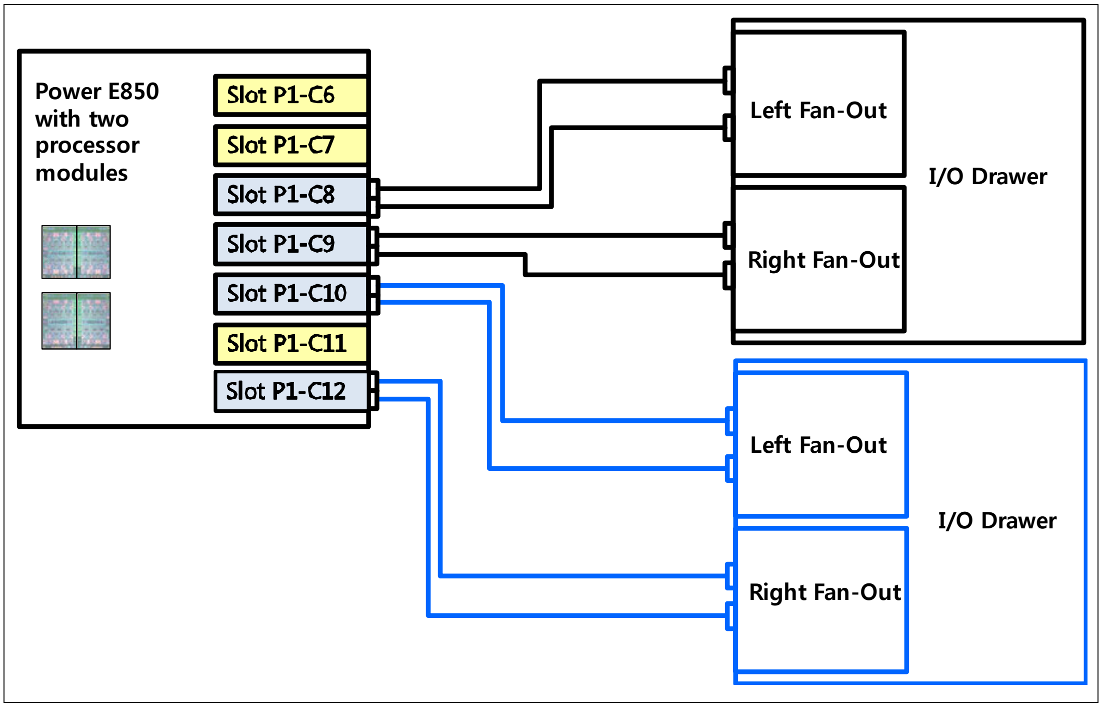


Figure 2-26 Example of a Power E850 and a maximum of two I/O drawers

Figure 2-27 shows an example of a Power E850 with three processor modules and a maximum of three PCIe Gen3 I/O expansion drawers.

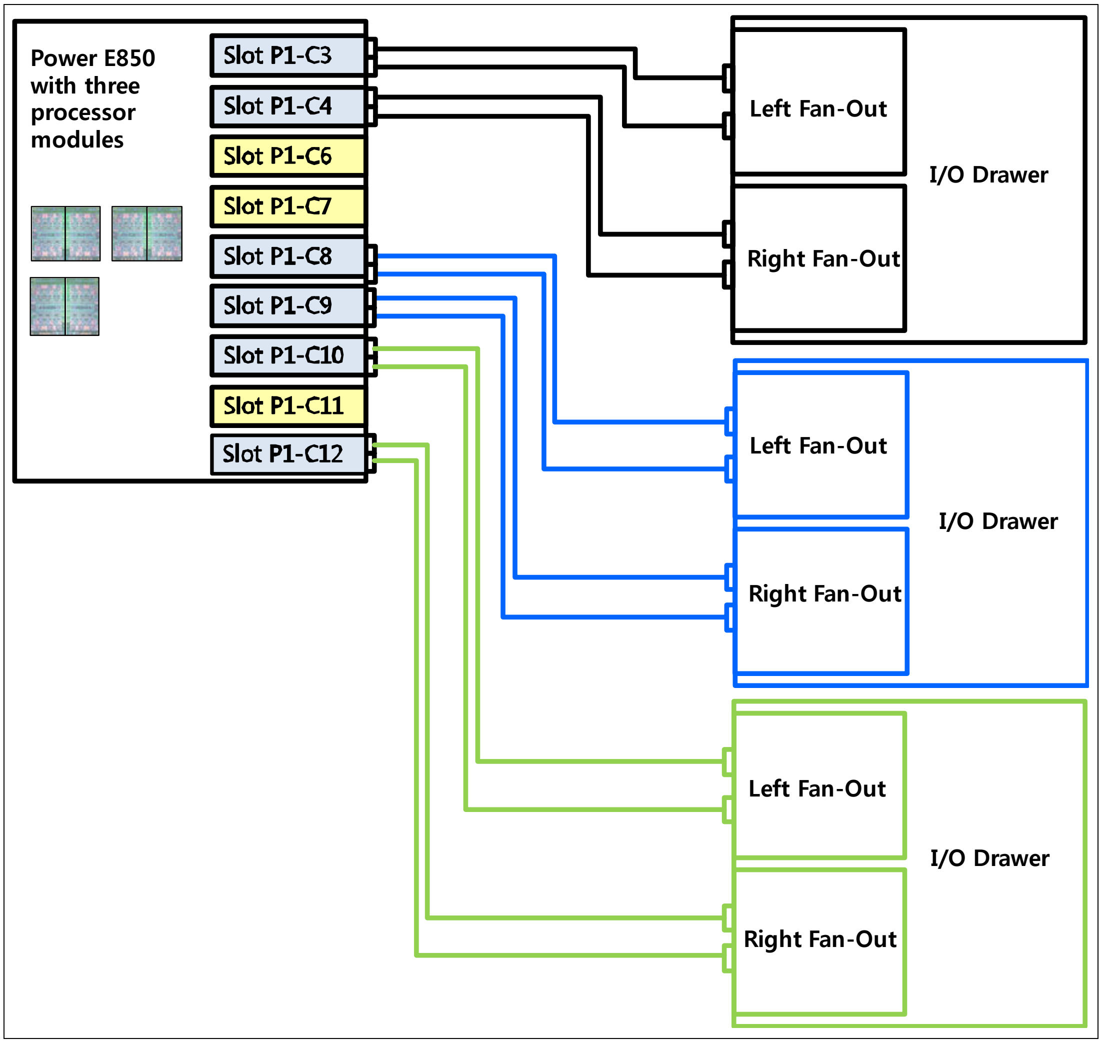


Figure 2-27 Example of a Power E850 and a maximum of three PCI Gen3 I/O expansion drawers

Figure 2-28 shows an example of Power E850 with four processor modules and a maximum of four PCIe Gen3 I/O expansion drawers.

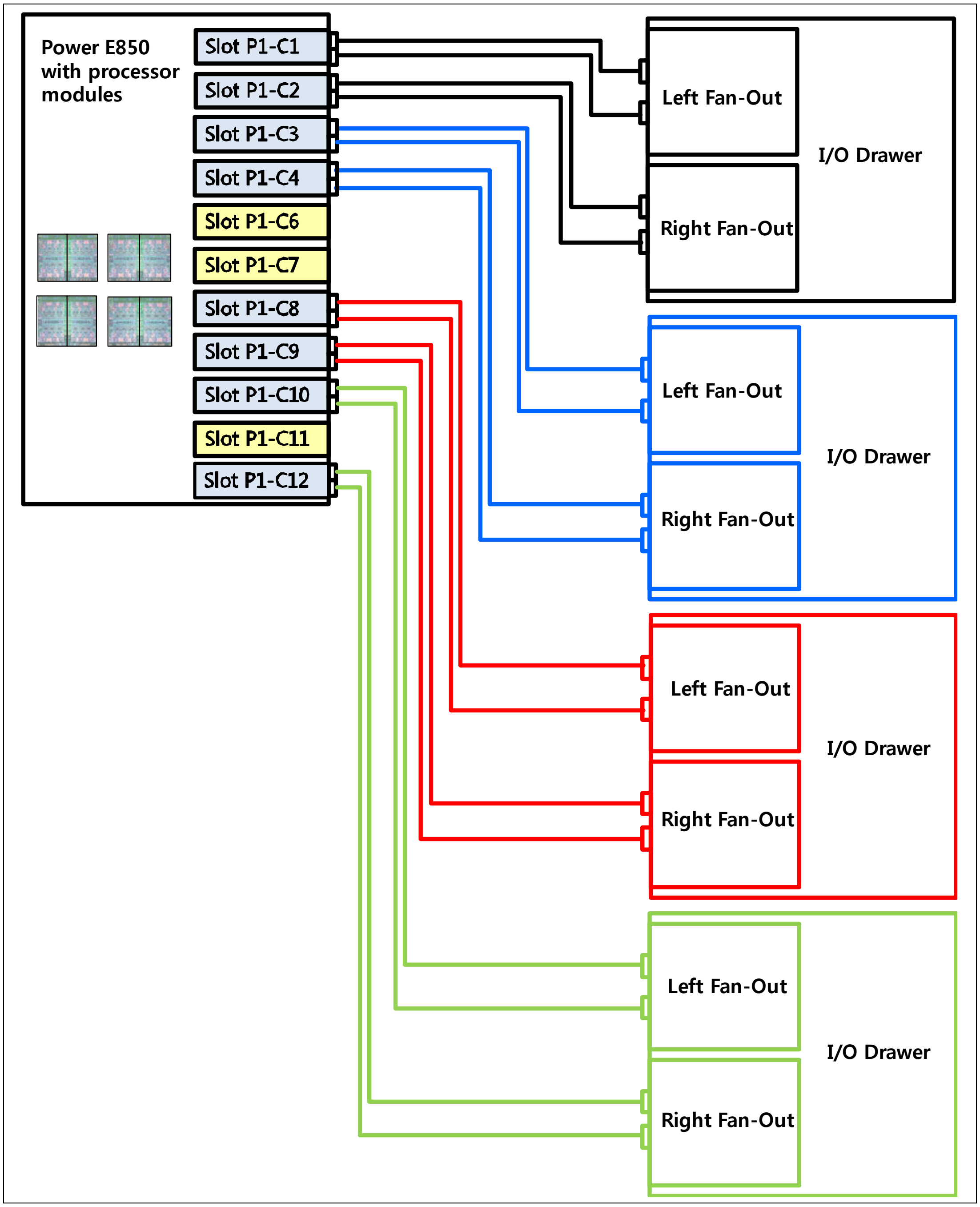


Figure 2-28 Example of a Power E850 and a maximum of four I/O drawers

2.11.3 PCIe Gen3 I/O expansion drawer SPCN cabling

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

2.12 External disk subsystems

This section describes the following external disk subsystems that can be attached to the Power E850 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 E850 server.
- IBM offers a 1U multimedia drawer that can hold one or more DVD drives, tape drives, or RDX docking stations. The 7226-1U3 is the most current offering. The earlier 7216-1U2 and 7214-1U2 are also supported. Up to six of these multimedia drawers can be attached, via a PCIe SAS adapter.

2.12.1 EXP24S SFF Gen2-bay Drawer

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

Note: A maximum of 64 EXP24S drawers can be attached to the Power E850 server providing an extra quantity of 1536 disks.

To further reduce possible single points of failure, EXP24S configuration rules consistent with previous Power Systems are used. 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). Figure 2-29 shows the front of the unit and the groups of disks on each mode.

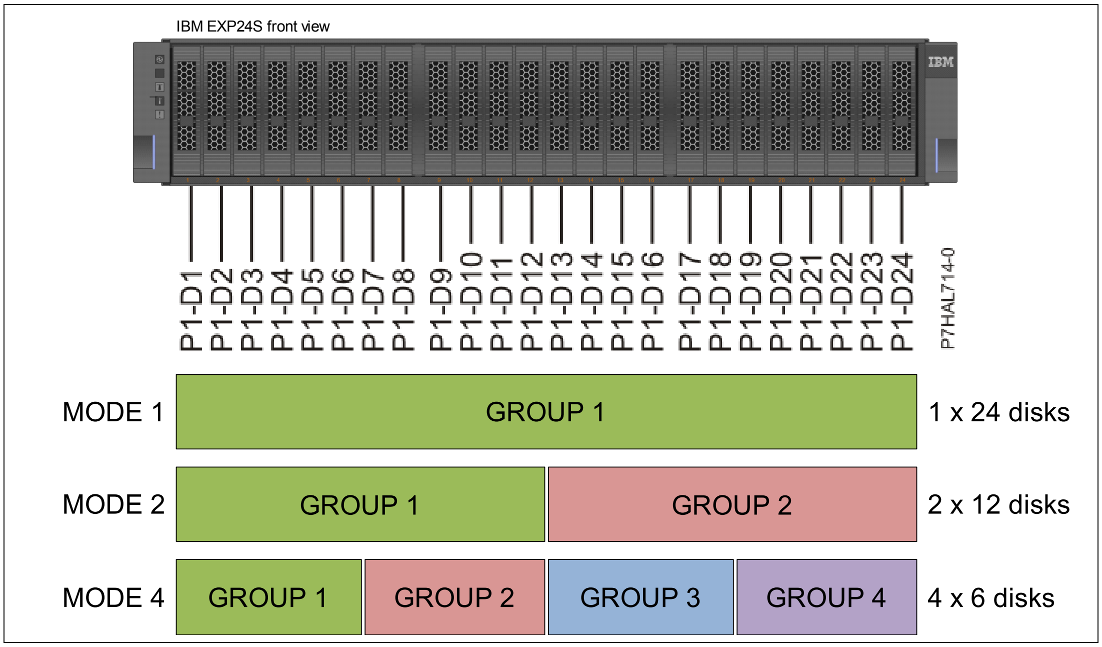


Figure 2-29 EXP24S front view with location codes and disk groups depending on its mode of operation

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

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

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

Figure 2-30 shows the mode stickers.

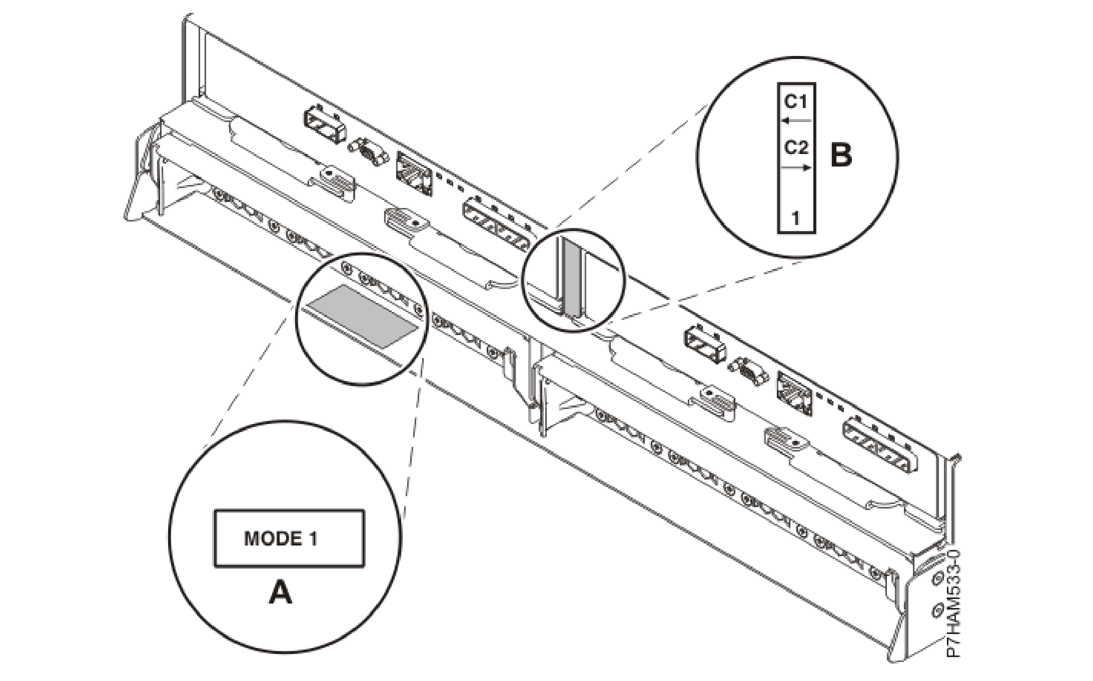


Figure 2-30 Mode sticker locations at the rear of the 5887 disk drive enclosure

The EXP24S SAS ports are attached to a SAS PCIe adapter or pair of adapters using SAS YO or X cables. Cable length varies depending on the feature code, and proper length should be calculated considering routing for proper airflow and ease of handling. A diagram of both types of SAS cables can be seen in Figure 2-31.

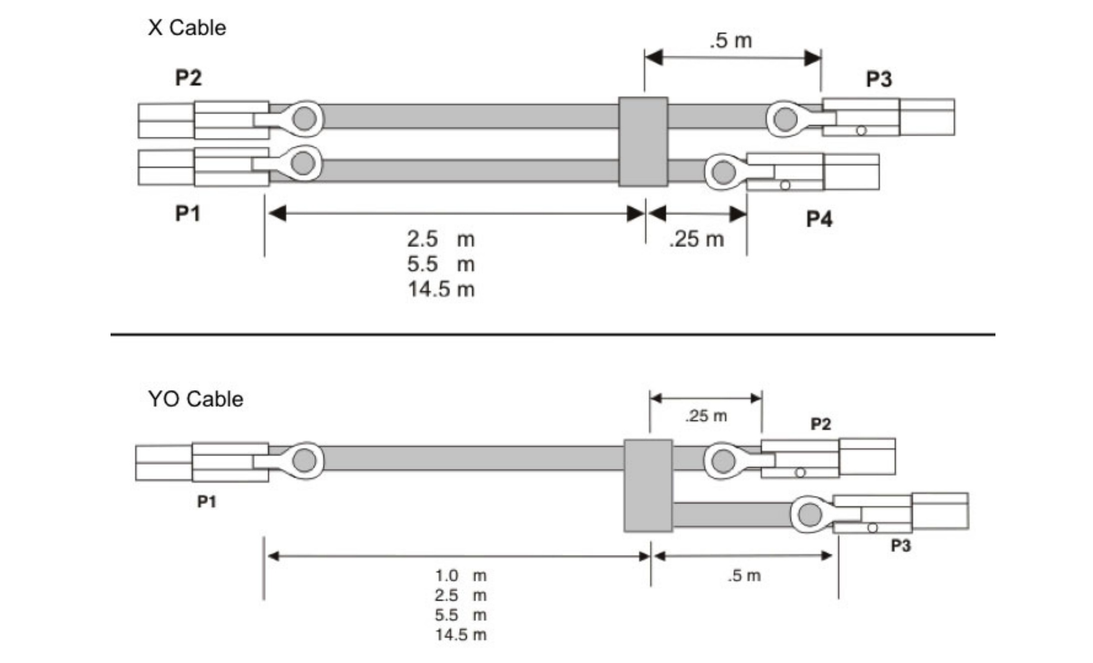


Figure 2-31 Diagram of SAS cable types X and YO

The following SAS adapters support the EXP24S:

- PCIe Dual-x4 SAS Adapter (#5901)
- PCIe3 RAID SAS Adapter Quad-port 6Gb x8 (#EJ0J)
- PCIe3 12GB Cache RAID SAS Adapter Quad-port 6Gb x8 (#EJ0L)
- PCIe3 SAS Tape/DVD Adapter Quad-port 6Gb x8 (#EJ10)

The EXP24S drawer can support up to 24 SAS SFF Gen-2 disks. Table 2-33 lists the available disk options.

Table 2-33 Available disks for the EXP24S

Feature code	Description	Max per server	OS support
ES0G	775 GB SFF-2 SSD for AIX/Linux	768	AIX, Linux
ES0Q	387 GB SFF-2 4 K SSD for AIX/Linux	768	AIX, Linux
ES0S	775 GB SFF-2 4 K SSD for AIX/Linux	768	AIX, Linux
ES19	387 GB SFF-2 SSD for AIX/Linux	768	AIX, Linux
1752	900 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX, Linux
1917	146 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX, Linux
1925	300 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX, Linux
1953	300 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX, Linux
1964	600 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX, Linux
ES0Q	387 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	768	AIX, Linux
ES78	775 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	768	AIX, Linux
ES7E	1.9 TB Read Intensive SAS 4k SFF-2 SSD for AIX/Linux	768	AIX, Linux
ES80	387 GB SFF-2 SSD 4k eMLC4 for AIX/Linux	768	AIX, Linux
ES85	775 GB SFF-2 SSD 4k eMLC4 for AIX/Linux	768	AIX, Linux
ES8C	1.55 TB SFF-2 SSD 4k eMLC4 for AIX/Linux	768	AIX, Linux
ESD3	1.2 TB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX, Linux
ESDP	600 GB 15 K RPM SAS SFF-2 Disk Drive - 5xx Block (AIX/Linux)	1536	AIX, Linux
ESEV	600 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	1536	AIX, Linux
ESEZ	300 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	1536	AIX, Linux
ESF3	1.2 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	1536	AIX, Linux
ESFP	600 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	1536	AIX, Linux
ESFT	1.8 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	1536	AIX, Linux

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

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

Figure 2-32 shows the rear connectors of the EXP24S drawer, how they relate with the modes of operation, and disk grouping.

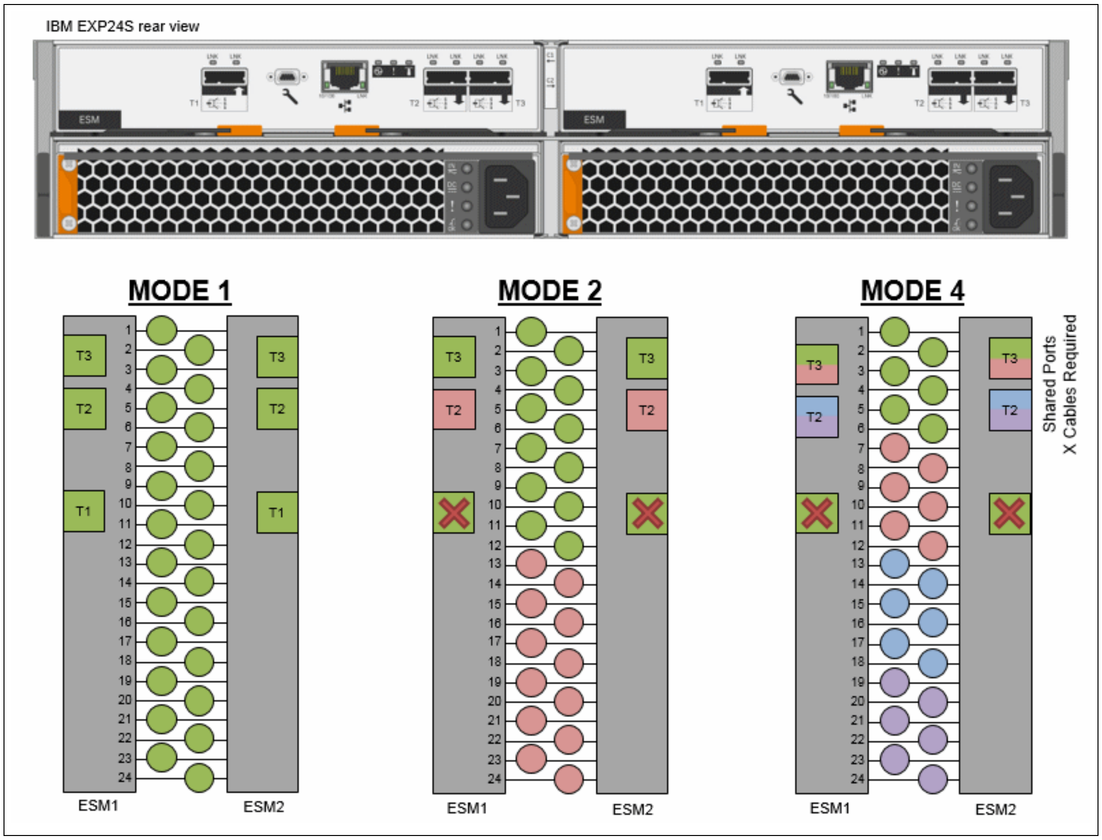


Figure 2-32 Rear view of EXP24S with the 3 modes of operation and the disks assigned to each port

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.

The most common configurations for EXP24S with Power Systems are detailed in 2.12.2, “EXP24S common usage scenarios” on page 87. Not all possible scenarios are included. For more information about SAS cabling and cabling configurations, search “Planning for serial-attached SCSI cables” in the IBM Knowledge Center, which can be accessed at:

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

2.12.2 EXP24S common usage scenarios

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

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

Scenario 1: Basic non-redundant connection

This scenario assumes a single Virtual I/O Server with a single PCIe SAS adapter #5901 and an EXP24S set on mode 1, allowing for up to 24 disks to be attached to the server. Figure 2-33 shows the connection diagram and components of the solution.

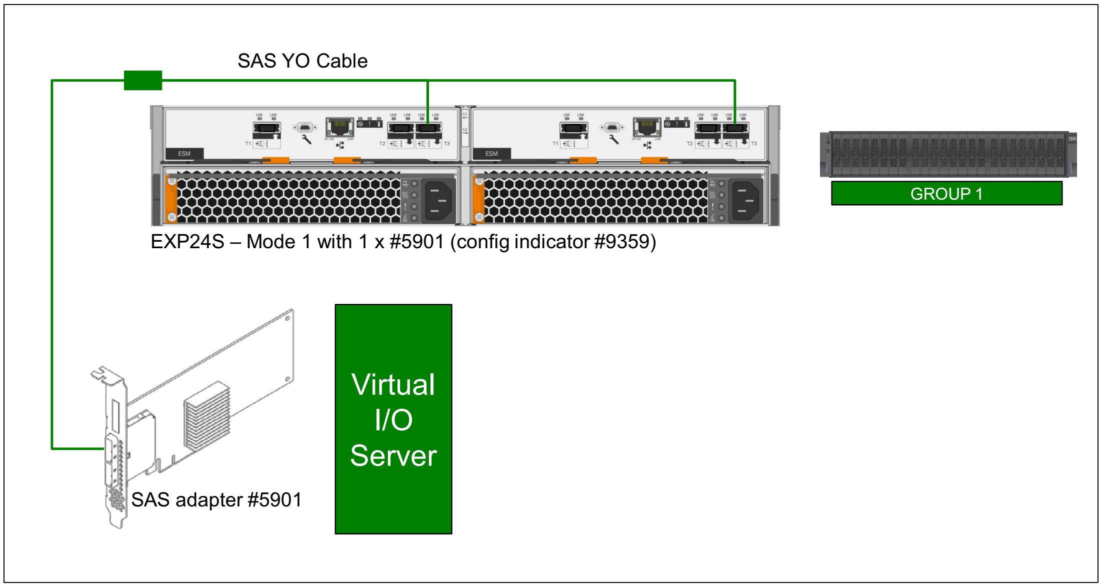


Figure 2-33 Scenario 1: Basic non-redundant connection

For this scenario, these are the required feature codes:

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

Scenario 2: Basic redundant connection

This scenario assumes a single Virtual I/O Server with two PCIe SAS adapters #5901 and an EXP24S set on mode 1, allowing for up to 24 disks to be attached to the server. Figure 2-34 shows the connection diagram and components of the solution.

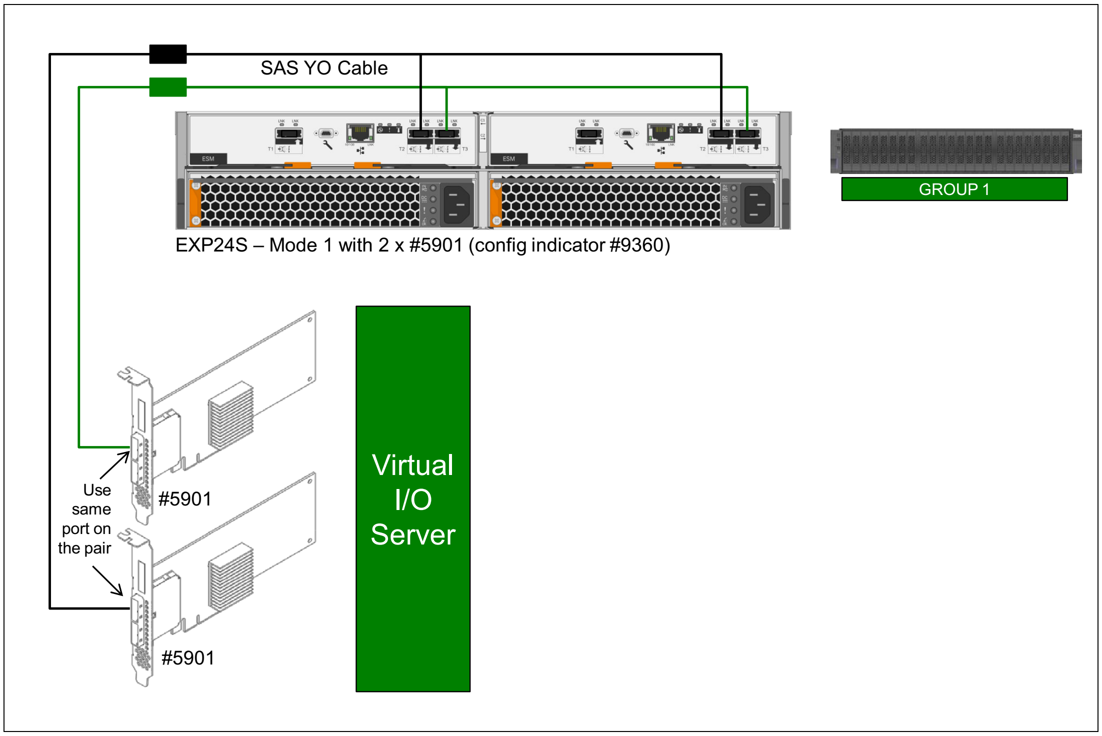

SAS YO Cable
GROUP 1
EXP24S – Mode 1 with 2 x #5901 (config indicator #9360)
Use same port on the pair
#5901
#5901
Virtual I/O Server

Figure 2-34 Scenario 2: Basic redundant connection

For this scenario, these are the required feature codes:

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

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

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

This scenario assumes a dual Virtual I/O Server with two PCIe SAS adapters #5901 each and an EXP24S set on mode 2, allowing for up to 12 disks to be attached to each Virtual I/O Server. Figure 2-35 shows the connection diagram and components of the solution.

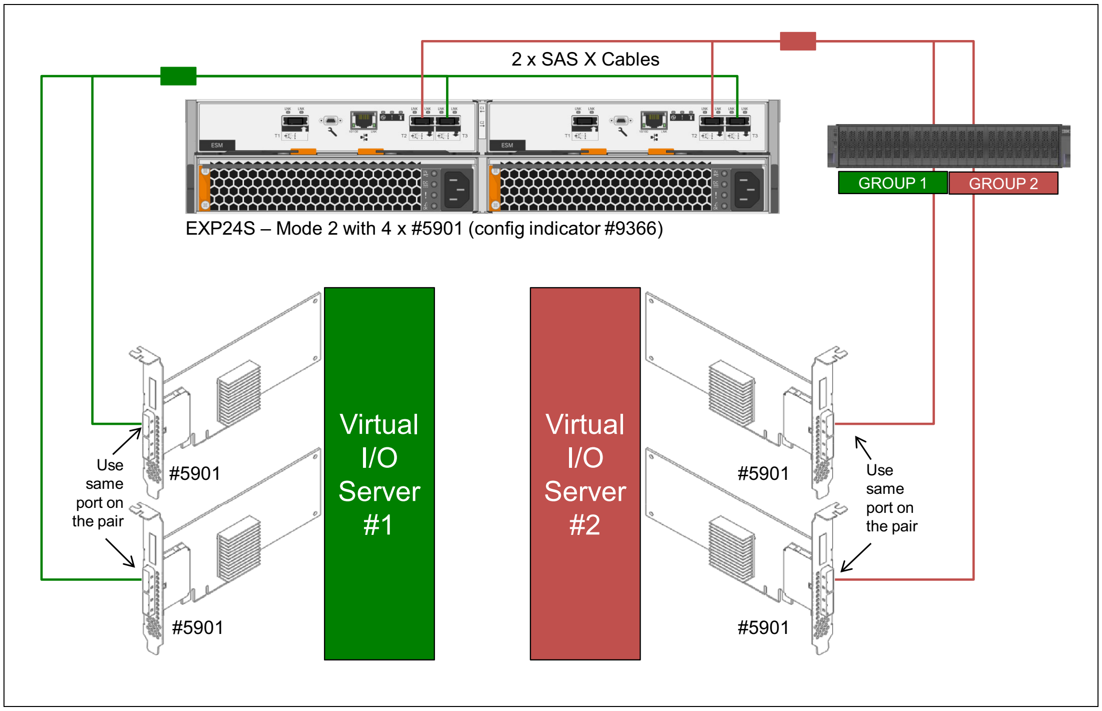

2 x SAS X Cables
GROUP 1
GROUP 2
EXP24S – Mode 2 with 4 x #5901 (config indicator #9366)
Use same port on the pair
#5901
#5901
Virtual I/O Server #1
Virtual I/O Server #2
#5901
#5901
Use same port on the pair

Figure 2-35 Dual Virtual I/O Servers sharing a single EXP24S

For this scenario, these are the required feature codes:

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

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

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

This scenario assumes a dual Virtual I/O Server with two PCIe SAS adapters #5901 each and two EXP24S set on mode 2, allowing for up to 24 disks to be attached to each Virtual I/O Server (2 per drawer). If compared to scenario 3, this scenario has the benefit of allowing disks from different EXP24S drawers to be mirrored, allowing for hot maintenance of the whole EXP24S drawers if all data is properly mirrored. Figure 2-36 shows the connection diagram and components of the solution.

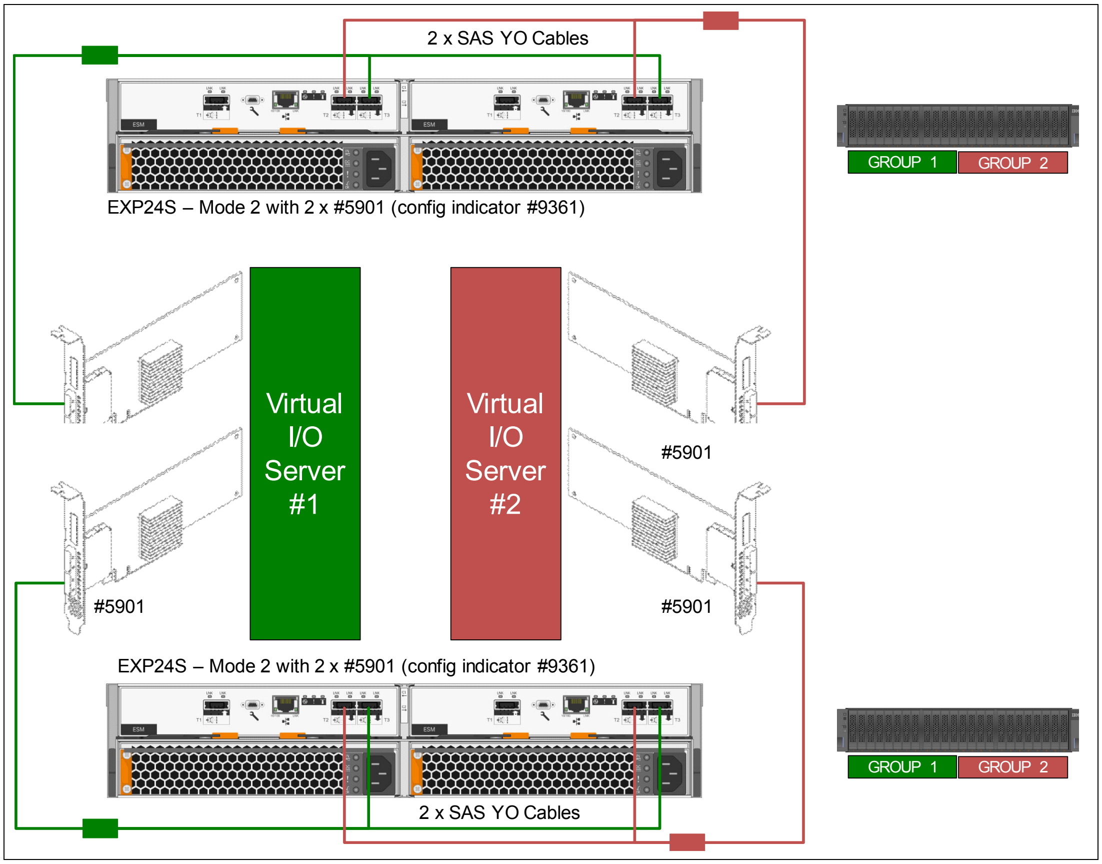


Figure 2-36 Dual Virtual I/O Servers sharing two EXP24S

For this scenario, these are the required feature codes:

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

There is no SSD support on this scenario.

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

This scenario assumes four Virtual I/O Servers with two PCIe SAS adapters #5901 each and two EXP24S set on mode 4, allowing for up to 12 disks to be attached to each Virtual I/O Server (6 per drawer). This scenario has the benefit to allow disks from different EXP24S drawers to be mirrored, allowing for hot maintenance of the whole EXP24S drawers if all data is properly mirrored. Figure 2-37 shows the connection diagram and components of the solution.

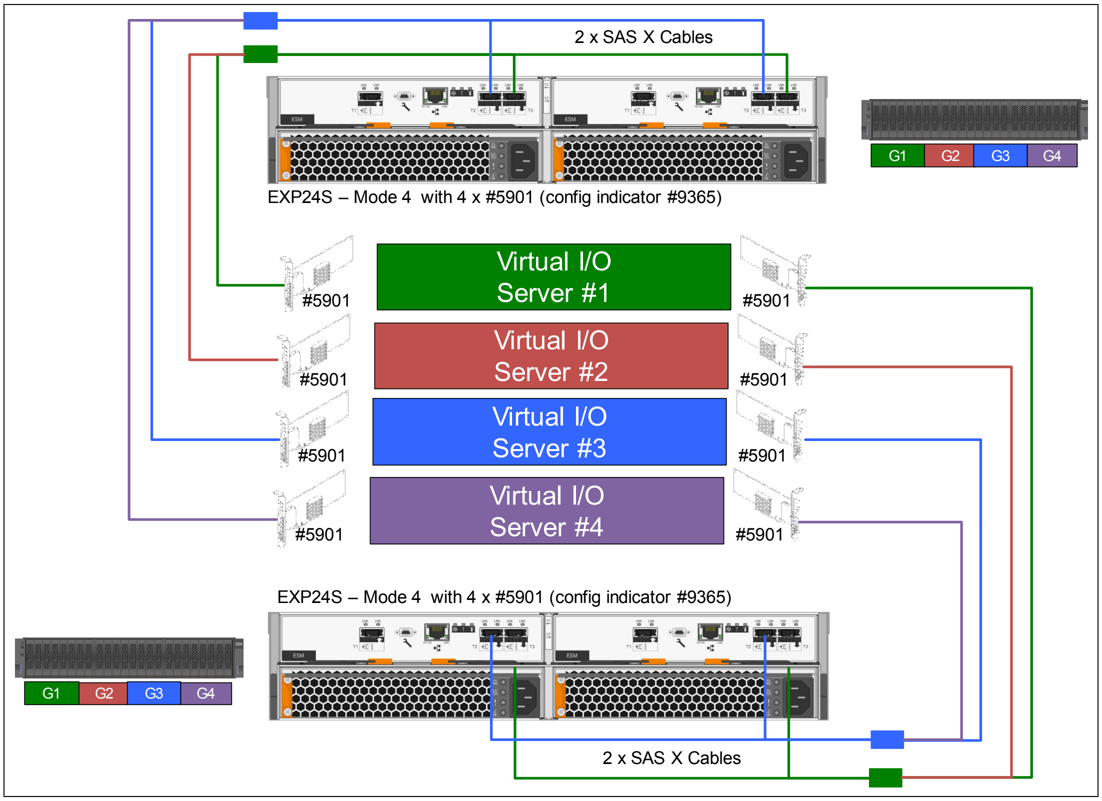


Figure 2-37 Four Virtual I/O Servers sharing two EXP24S

For this scenario, these are the required feature codes:

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

There is no SSD support on this scenario.

Other scenarios

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

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

2.12.3 IBM System Storage

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

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

The following section highlights a few of the offerings.

IBM network-attached storage

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

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

IBM Storwize family

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

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

IBM flash storage

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

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

IBM XIV Storage System

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

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

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

IBM System Storage DS8000

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

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

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

2.13 Hardware Management Console

The Power E850 platforms support two main service environments:

- Attachment to one or more HMCs. This environment is the common configuration for servers supporting logical partitions with dedicated or virtual I/O. In this case, all servers have at least one logical partition.
- No HMC. There are two service strategies for non-HMC systems.
  - Full-system partition with PowerVM: A single partition owns all the server resources and only one operating system may be installed.
  - Partitioned system with PowerVM: In this configuration, the system can have more than one partition and can be running more than one operating system. In this environment, partitions are managed by the Integrated Virtualization Manager (IVM), which provides some of the functions provided by the HMC.

The HMC is a dedicated appliance that allows administrators to configure and manage system resources on IBM Power Systems servers. The latest HMC can manage 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 Power Systems servers. The HMC also supports advanced service functions, including guided repair and verification, concurrent firmware updates for managed systems, and around-the-clock error reporting through IBM Electronic Service Agent™ for faster support.

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

Requirements:

- When using the HMC with the Power E850 servers, the HMC code must be running at V8R8.3.0 level, or later.
- When PowerVC is enabled, 4 GB of RAM is recommended. To use the enhanced login mode, 8 GB is recommended. HMC 7042-CR5 ships with a default of 2 GB RAM.
- A single HMC can manage multiple Power Systems servers.
- HMCs supported on POWER8 hardware are 7042-CR5 through 7042-CR9.
- HMC is required to enable Active Memory Expansion.
- HMC is required to use SR-IOV.
- HMC is required for elastic, utility, or trial CoD.

Hardware support for customer-replaceable units. This support comes standard along with the HMC. In addition, users can upgrade this support level to IBM onsite support to be consistent with other Power Systems servers.

2.13.1 HMC code level

HMC V8R8.3.0 contains the following new features:

- Support for Power E850 servers
- Console Management:
  - Full release of the new enhanced user interface
  - Browser currency
  - Improved log retention (through file system resizing, rotation changes, and content reduction)
  - Call Home support for modem (dial-in via AT&T Global Network) and VPN is removed

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

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

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

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

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

2.13.2 HMC RAID 1 support

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

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

HMC models

To use an existing HMC to manage any POWER8 processor-based server, the HMC must be a model CR5, or later, rack-mounted HMC, or model C08, or later, deskside HMC. The latest HMC model is the 7042-CR9. For your reference, Table 2-34 on page 95 lists a comparison between the 7042-CR8 and the 7042-CR9 HMC models.

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

Table 2-34 Comparison between 7042-CR8 and 7042-CR9 models

Feature	CR8	CR9
IBM System x model	x3550 M4 7914 PCH	x3550 M5 5463 AC1
HMC model	7042-CR8	7042-CR9
Processor	Intel 8-Core Xeon v2 2.00 GHz	Intel 18-core Xeon v3 2.4 GHz
Memory max:	16 GB (when featured)	16 GB DDR4 expandable to 192 GB
DASD	500 GB	500 GB
RAID 1	Default	Default
USB ports	Two front, four back	Two front, four rear
Integrated network	Four 1 Gb Ethernet	Four 1 Gb Ethernet
I/O slots	One PCI Express 3.0 slot	One PCI Express 3.0 slot

2.13.3 HMC connectivity to the POWER8 processor-based systems

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

For the HMC to communicate properly with the managed server, eth0 of the HMC must be connected to either the HMC1 or HMC2 ports of the managed server, although other network configurations are possible. You 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.

Figure 2-38 shows a simple network configuration to enable the connection from the HMC to the server and to allow for dynamic LPAR operations. For more information about HMC and the possible network connections, see *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491.

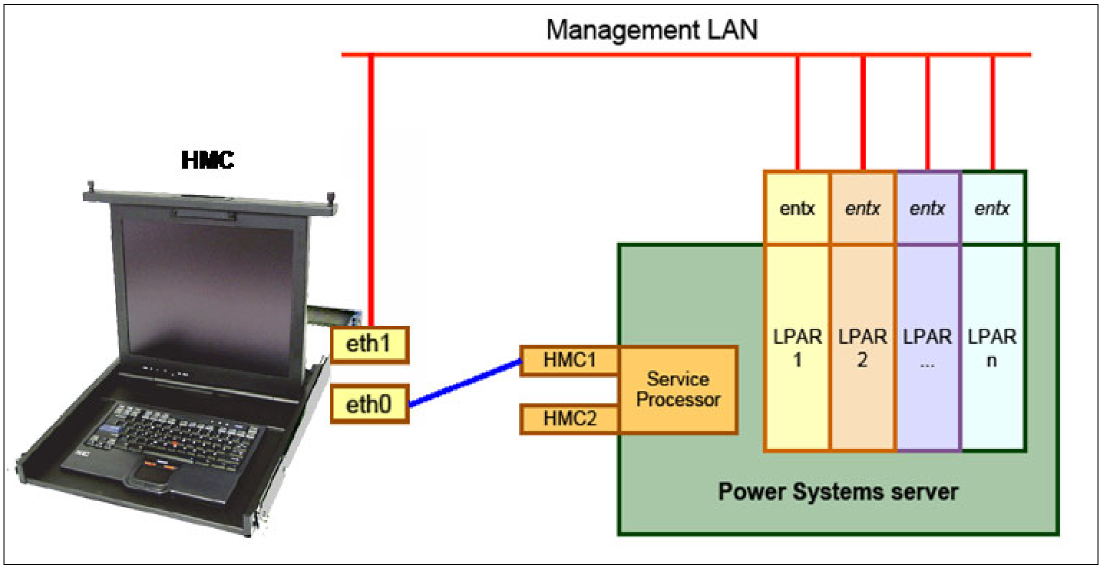


Figure 2-38 Network connections from the HMC to service processor and LPARs

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

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

> **Note:** The two service processor HMC ports have the following features:
>
> - Run at a speed of 1 Gbps
> - Are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI options from a client directly from a client web browser
> - Use the following network configuration if no IP addresses are set:
>   - Service processor eth0 (HMC1 port): 169.254.2.147 with netmask 255.255.255.0
>   - Service processor eth1 (HMC2 port): 169.254.3.147 with netmask 255.255.255.0
>
> For more information about the service processor, see 2.8.1, “System ports” on page 62.

2.13.4 High availability HMC configuration

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

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

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

Figure 2-39 shows one possible highly available HMC configuration that is managing two servers. Each HMC is connected to one FSP port of each managed server.

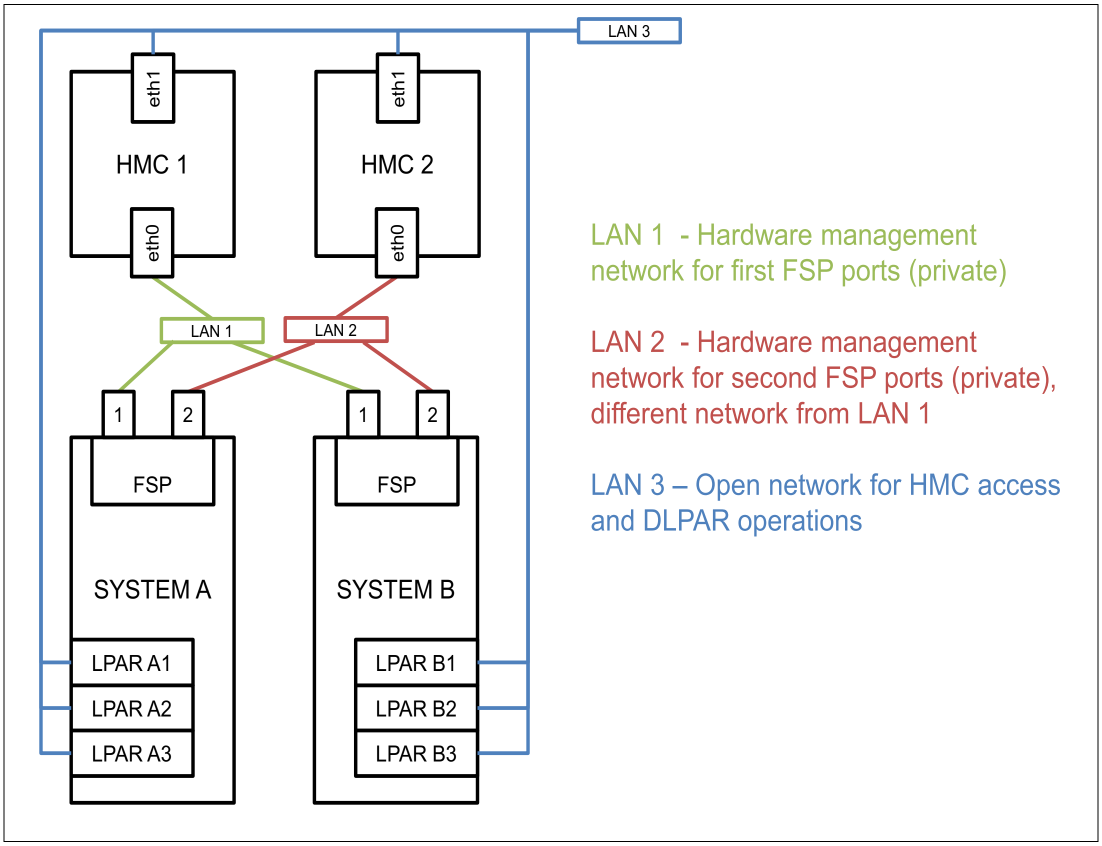

LAN 3
eth1
HMC 1
eth0
eth1
HMC 2
eth0
LAN 1
LAN 2
1
2
1
2
FSP
FSP
SYSTEM A
SYSTEM B
LPAR A1
LPAR A2
LPAR A3
LPAR B1
LPAR B2
LPAR B3
LAN 1 - Hardware management network for first FSP ports (private)
LAN 2 - Hardware management network for second FSP ports (private), different network from LAN 1
LAN 3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

Figure 2-39 Highly available HMC networking example

For simplicity, only the hardware management networks (LAN1 and LAN2) are highly available (Figure 2-39). However, the open network (LAN3) can be made highly available by using a similar concept and adding a second network between the partitions and HMCs.

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

2.14 Operating system support

The IBM Power E850 systems support the following operating systems:

- AIX
- Linux

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

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

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

2.14.1 Virtual I/O Server

The minimum required level of Virtual I/O Server for Power E850 is VIOS 2.2.3.51 or later.

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

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

2.14.2 IBM AIX operating system

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

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

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

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

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

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

IBM AIX Version 6.1

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

The minimum level of AIX Version 6.1 supported on the Power E850 depends on the partition having 100% virtualized resources or not.

For partitions that have all of their resources virtualized via Virtual I/O Server, the minimum levels of AIX Version 6.1 supported on the Power E850 is as follows:

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

For all other partitions, the minimum levels of AIX Version 6.1 supported on the Power E850 is as follows:

- AIX Version 6.1 with the 6100-08 Technology Level Service Pack 7, or later (planned availability: September 30, 2015)
- AIX Version 6.1 with the 6100-09 Technology Level Service Pack 5, and APAR IV68443, or later

IBM AIX Version 7.1

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

A partition running AIX 7.1 in POWER6, POWER6+, or POWER7 mode can be migrated to a POWER8 system and upgraded to POWER8 mode. In POWER8 mode, it can make full use of the latest POWER8 capabilities.

The minimum level of AIX Version 7.1 supported on the Power E850 depends on the partition having 100% virtualized resources or not.

For partitions that have all of their resources virtualized via Virtual I/O Server, the minimum levels of AIX Version 7.1 supported on the Power E850 is as follows:

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

For all other partitions, the minimum levels of AIX Version 7.1 supported on the Power E850 is as follows:

- AIX Version 7.1 with the 7100-02 Technology Level Service Pack 7, or later (planned availability: September 30, 2015)
- AIX Version 7.1 with the 7100-03 Technology Level Service Pack 5, and APAR IV68444, or later

2.14.3 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 E850 are as follows:

- Big endian
  - Red Hat Enterprise Linux 7.1, or later
  - Red Hat Enterprise Linux 6.6, or later
  - SUSE Linux Enterprise Server 11 Service Pack 3 (IBM Power Systems Solution Editions for SAP HANA clients only)
- Little endian
  - Red Hat Enterprise Linux 7.1, or later
  - SUSE Linux Enterprise Server 12 and later service packs
  - Ubuntu 15.04

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

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

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

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

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

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

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

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

For information about Ubuntu Server, see this site:

`http://www.ubuntu.com/server`

2.14.4 Java versions that are supported

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

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

2.14.5 Boosting performance and productivity with IBM compilers

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

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

XL C, XL C/C++, and XL Fortran features introduced to exploit the latest POWER8 processor include vector unit and vector scalar extension (VSX) instruction set to efficiently manipulate vector operations in your application, vector functions within the Mathematical Acceleration Subsystem (MASS) libraries for improved application performance, built-in functions or intrinsics and directives for direct control of POWER instructions at the application level, and architecture and tune compiler options to optimize and tune your applications.

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

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

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

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

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

2.15 Energy management

The Power E850 systems have features to help clients become more energy efficient. EnergyScale technology enables advanced energy management features to conserve power dramatically and dynamically and further improve energy efficiency. Intelligent Energy optimization capabilities enable the POWER8 processor to operate at a higher frequency for increased performance and performance per watt, or dramatically reduce frequency to save energy.

2.15.1 IBM EnergyScale technology

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

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

- Power trending

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

- Power saver mode

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

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

- Dynamic power saver mode

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

  When a system is idle, the system firmware lowers the frequency and voltage to power energy saver mode values. When fully used, the maximum frequency varies, depending on whether the user favors power savings or system performance. If an administrator prefers energy savings and a system is fully used, the system is designed to reduce the maximum frequency to about 95% of nominal values.

  Dynamic power saver mode is mutually exclusive with power saver mode. Only one of these modes can be enabled at a given time.

- Power capping

  Power capping enforces a user-specified limit on power usage. Power capping is not a power-saving mechanism. It enforces power caps by throttling the processors in the system, degrading performance significantly. The idea of a power cap is to set a limit that must never be reached but that frees extra power that was never used in the data center. The *margined* power is this amount of extra power that is allocated to a server during its installation in a data center. It is based on the server environmental specifications that usually are never reached because server specifications are always based on maximum configurations and worst-case scenarios.

- Soft power capping

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

- Processor core nap mode

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

- Processor core sleep mode

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

- Processor chip winkle mode

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

- Fan control and altitude input

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

- Processor folding

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

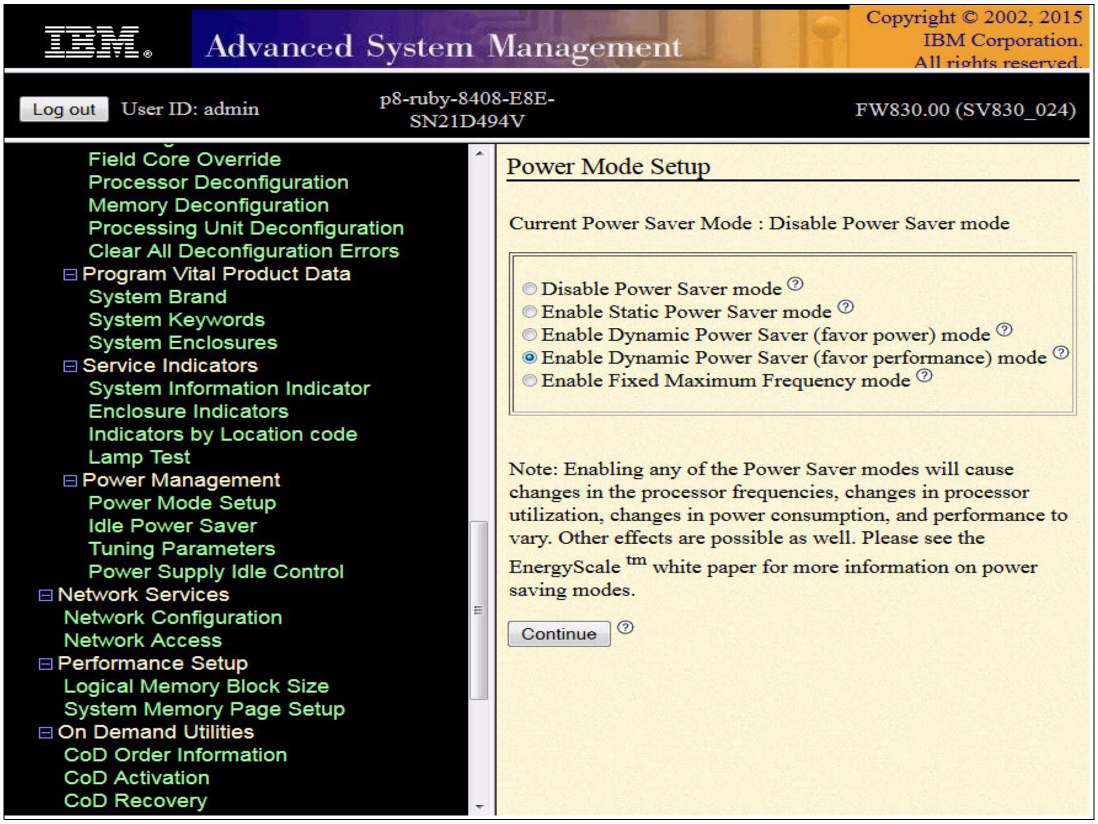
- EnergyScale for I/O

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

- Server power down

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

  On POWER8 processor-based systems, several EnergyScale technologies are embedded in the hardware and do not require an operating system or external management component. Fan control, environmental monitoring, and system energy management are controlled by the On Chip Controller (OCC) and associated components. The power mode can also be set up without external tools, by using the ASMI interface, as shown in Figure 2-40.



User ID: admin

p8-ruby-8408-E8E-SN21D494V

FW830.00 (SV830_024)

Power Mode Setup

Current Power Saver Mode : Disable Power Saver mode

○ Disable Power Saver mode
○ Enable Static Power Saver mode
○ Enable Dynamic Power Saver (favor power) mode
● Enable Dynamic Power Saver (favor performance) mode
○ Enable Fixed Maximum Frequency mode

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

Figure 2-40 Setting the power mode in ASMI

2.15.2 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 using an embedded IBM PowerPC core with 512 KB of SRAM runs real-time control firmware to respond to workload variations by adjusting the per-core frequency and voltage based on activity, thermal, voltage, and current sensors.

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

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

2.15.3 Energy consumption estimation

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

- Maximum power consumption and power source loading values

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

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

  Search for type and model number and “server specifications”. For example, for the Power E850 system, search for “8408-E8E server specifications”.

- An estimation of the energy consumption for a certain configuration

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

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

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

3

Virtualization

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

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

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

Single Root I/O Virtualization (SR-IOV) is supported on the Power E850 server. For more information about SR-IOV, see chapter 3.4, “Single root I/O virtualization” on page 114.

PowerVM license: The Standard and Enterprise editions are available on Power E850.

PowerKVM is not supported on the Power E850. PowerVM is the only virtualization technology available for these systems.

3.1 IBM POWER Hypervisor

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

- Provides an abstraction between the physical hardware resources and the logical partitions that use them.
- Enforces partition integrity by providing a security layer between logical partitions.
- Controls the dispatch of virtual processors to physical processors (see "Processing mode" on page 118).
- Saves and restores all processor state information during a logical processor context switch.
- Controls hardware I/O interrupt management facilities for logical partitions.
- Provides virtual LAN channels between logical partitions that help reduce the need for physical Ethernet adapters for inter-partition communication.
- Monitors the service processor and performs a reset or reload if it detects the loss of the service processor, notifying the operating system if the problem is not corrected.

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

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

The minimum amount of physical memory that is required to create a partition is the size of the system's logical memory block (LMB). The default LMB size varies according to the amount of memory that is configured in the system (see Table 3-1).

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

Configurable system memory	Default logical memory block
Up to 32 GB	128 MB
Greater than 32 GB	256 MB

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

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

- Virtual Small Computer System Interface (SCSI)
- Virtual Ethernet
- Virtual Fibre Channel
- Virtual (TTY) console

3.1.1 Virtual SCSI

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

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

A Virtual I/O Server (VIOS) partition can define virtual SCSI server adapters. Other partitions are *client* partitions. The VIOS partition is a special logical partition, which is described in 3.5.4, “Virtual I/O Server” on page 121. The VIOS software is included on all PowerVM editions. When using the PowerVM Standard Edition and PowerVM Enterprise Edition, dual VIOS can be deployed to provide maximum availability for client partitions when performing VIOS maintenance.

3.1.2 Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions on the same server to use fast and secure communication without any need for physical interconnection. The virtual Ethernet allows a transmission speed up to 20 Gbps, depending on the maximum transmission unit (MTU) size, type of communication, and CPU entitlement. Virtual Ethernet support began with IBM AIX Version 5.3, Red Hat Enterprise Linux 4, and SUSE Linux Enterprise Server, 9, and is supported on all later versions.
(For more information, see 3.5.8, “Operating system support for PowerVM” on page 130). The virtual Ethernet is part of the base system configuration.

Virtual Ethernet has the following major features:

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

Any virtual Ethernet can also have connectivity outside of the server if a Layer 2 bridge to a physical Ethernet adapter is set in one VIOS partition, also known as *Shared Ethernet Adapter*. For more information about shared Ethernet, see 3.5.4, “Virtual I/O Server” on page 121.

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

3.1.3 Virtual Fibre Channel

A virtual Fibre Channel adapter is a virtual adapter that provides client logical partitions with a Fibre Channel connection to a storage area network through the VIOS logical partition. The VIOS logical partition provides the connection between the virtual Fibre Channel adapters on the VIOS logical partition and the physical Fibre Channel adapters on the managed system. Figure 3-1 shows the connections between the client partition virtual Fibre Channel adapters and the external storage. For more information, see 3.5.8, “Operating system support for PowerVM” on page 130.

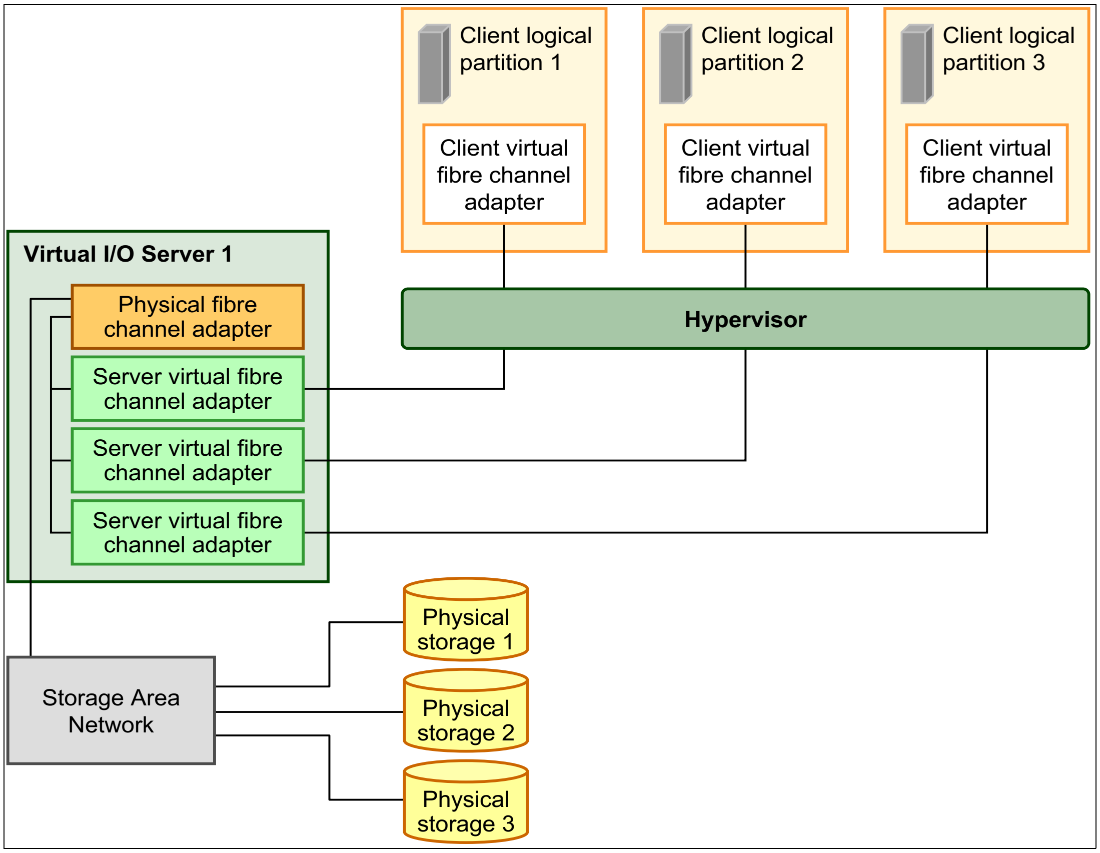

Client logical partition 1
Client logical partition 2
Client logical partition 3
Client virtual fibre channel adapter
Client virtual fibre channel adapter
Client virtual fibre channel adapter
Virtual I/O Server 1
Physical fibre channel adapter
Server virtual fibre channel adapter
Server virtual fibre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

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

3.1.4 Virtual (TTY) console

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

Depending on the system configuration, the operating system console can be provided by the Hardware Management Console (HMC) virtual TTY, Integrated Virtualization Manager (IVM) virtual TTY, or from a terminal emulator that is connected to a system port.

3.2 POWER processor modes

Although they are not virtualization features, the POWER processor modes are described here because they affect various virtualization features.

On POWER8 based Power System servers, partitions can be configured to run in several modes, including the following modes:

- POWER6 compatibility mode

  This execution mode is compatible with Version 2.05 of the Power Instruction Set Architecture (ISA). For more information, visit the following website:

  `http://power.org/wp-content/uploads/2012/07/PowerISA_V2.05.pdf`

- POWER6+ compatibility mode

  This mode is similar to POWER6, with eight more storage protection keys.

- POWER7 compatibility mode

  This is the mode for POWER7+ and POWER7 processors, implementing Version 2.06 of the Power Instruction Set Architecture. For more information, visit the following website:

  `http://power.org/wp-content/uploads/2012/07/PowerISA_V2.06B_V2_PUBLIC.pdf`

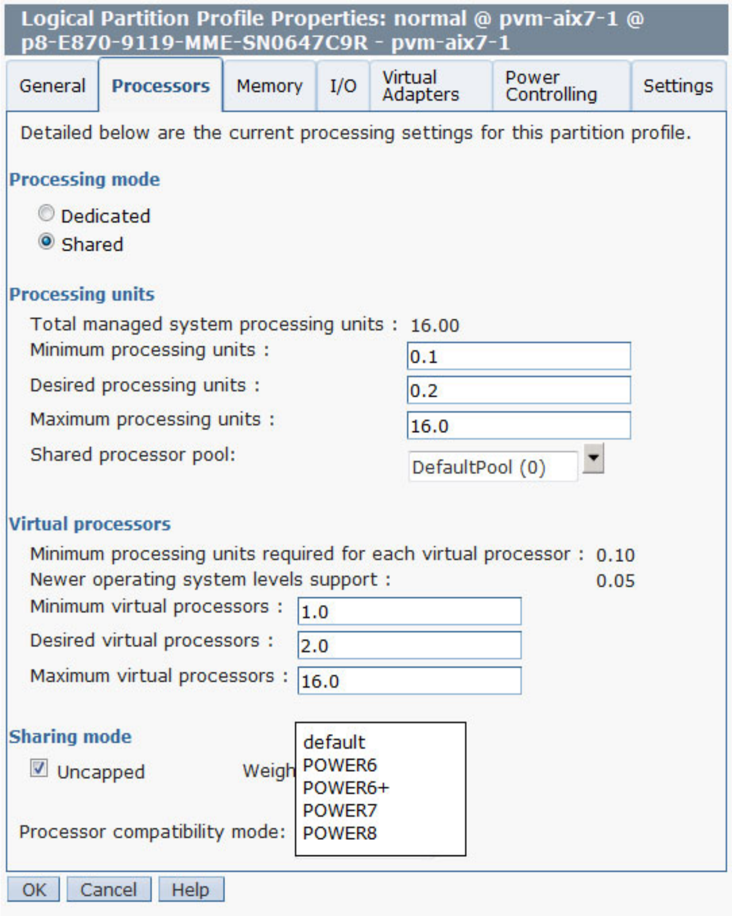
- POWER8 compatibility mode

  This is the native mode for POWER8 processors implementing Version 2.07 of the Power Instruction Set Architecture. For more information, visit the following website:

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

The selection of the mode is made on a per-partition basis, from the management console, by editing the partition profile.

Figure 3-2 shows the compatibility modes within the logical partition (LPAR) profile.



Logical Partition Profile Properties: normal @ pvm-aix7-1 @ p8-E870-9119-MME-SN0647C9R - pvm-aix7-1

Detailed below are the current processing settings for this partition profile.

Processing mode

○ Dedicated
● Shared

Processing units

Total managed system processing units :	16.00
Minimum processing units :	0.1
Desired processing units :	0.2
Maximum processing units :	16.0
Shared processor pool:	DefaultPool (0)

Virtual processors

Minimum processing units required for each virtual processor :	0.10
Newer operating system levels support :	0.05
Minimum virtual processors :	1.0
Desired virtual processors :	2.0
Maximum virtual processors :	16.0

Sharing mode

☑ Uncapped Weigh

Processor compatibility mode:

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

Table 3-2 lists the differences between the processor modes.

Table 3-2 Differences between POWER6, POWER7, and POWER8 compatibility modes

POWER6 and POWER6+ mode	POWER7 mode	POWER8 mode	Customer value
2-thread simultaneous multithreading (SMT)	4-thread SMT	8-thread SMT	Throughput performance, and processor core usage
Vector Multimedia Extension/AltiVec (VMX)	Vector scalar extension (VSX)	VSX2 In-Core Encryption Acceleration	High-performance computing
Affinity off by default	3-tier memory, micropartition affinity, and dynamic platform optimizer	► HW memory affinity tracking assists ► Micropartition prefetch ► Concurrent LPARs per core	Improved system performance for system images spanning sockets and nodes

POWER6 and POWER6+ mode	POWER7 mode	POWER8 mode	Customer value
64-core and 128-thread scaling	► 32-core and 128-thread scaling ► 64-core and 256-thread scaling ► 128-core and 512-thread scaling ► 256-core and 1024-thread scaling	► 48-core and 384-thread scaling ► Hybrid threads ► Transactional memory ► Active system optimization hardware assists	Performance and scalability for large scale-up single system image workloads (such as OLTP, ERP scale-up, and WPAR consolidation)
EnergyScale CPU Idle	EnergyScale CPU Idle and Folding with NAP and SLEEP	WINKLE, NAP, SLEEP, and Idle power saver	Improved energy efficiency

3.3 Active Memory Expansion

Active Memory Expansion is an optional feature for the Power E850, which can be ordered with feature code #4798.

This feature enables memory expansion on the system. By using compression and decompression of memory, content can effectively expand the maximum memory capacity, providing additional server workload capacity and performance.

Active Memory Expansion is a technology that allows the effective maximum memory capacity to be much larger than the true physical memory maximum. Compression and decompression of memory content can allow memory expansion up to 100% and more for AIX partitions, which in turn enables a partition to perform more work or support more users with the same physical amount of memory. Similarly, it can allow a server to run more partitions and do more work for the same physical amount of memory.

Note: The Active Memory Expansion feature is only supported by the AIX operating system.

Active Memory Expansion uses the CPU resource to compress and decompress the memory contents. The trade-off of memory capacity for processor cycles can be an excellent choice, but the degree of expansion varies based on how compressible the memory content is, and it also depends on having adequate spare CPU capacity available for this compression and decompression.

The POWER8 processor includes an Active Memory Expansion co-processor on the processor chip to provide dramatic improvement in performance and greater processor efficiency. To take advantage of the hardware compression offload, AIX 6.1 Technology Level 8 and 9, or AIX 7.1 Technology Level 2 and 3 is required for Power E850.

Tests in IBM laboratories, using sample work loads, showed excellent results for many workloads in terms of memory expansion per additional CPU used. Other test workloads had more modest results. The ideal scenario is when there are many cold pages, that is, infrequently referenced pages. However, if many memory pages are referenced frequently, Active Memory Expansion might not be a preferred choice.

Tip: If the workload is based on Java, the garbage collector must be tuned so that it does not access the memory pages so often, that is, turning cold pages to hot.

A planning tool is included within AIX, which allows you to sample actual workloads and estimate the level of expansion and processor usage expected. This can be run on any Power Systems server running PowerVM as a hypervisor. A one-time, 60-day trial of Active Memory Expansion is available on each server to confirm the estimated results. You can request the trial and find more information about the IBM Capacity on Demand website:

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

For more information about Active Memory Expansion, download the document *Active Memory Expansion: Overview and Usage Guide*, found at:

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

3.4 Single root I/O virtualization

Single root I/O virtualization (SR-IOV) is an extension to the PCI Express (PCIe) specification that allows multiple operating systems to simultaneously share a PCIe adapter with little or no runtime involvement from a hypervisor or other virtualization intermediary.

SR-IOV is PCI standard architecture that enables PCIe adapters to become self-virtualizing. It enables adapter consolidation, through sharing, much like logical partitioning enables server consolidation. With an adapter capable of SR-IOV, you can assign virtual *slices* of a single physical adapter to multiple partitions through logical ports. All of this is done without the need for a Virtual I/O Server (VIOS).

Initial SR-IOV deployment supports up to 48 logical ports per adapter, depending on the adapter. You can provide additional fan-out for more partitions by assigning a logical port to a VIOS, and then using that logical port as the physical device for a Shared Ethernet Adapter (SEA). VIOS clients can then use that SEA through a traditional virtual Ethernet configuration.

Overall, SR-IOV provides integrated virtualization without VIOS and with greater server efficiency as more of the virtualization work is done in the hardware and less in the software.

The following are the hardware requirements to enable SR-IOV:

- One of the following pluggable PCIe adapters:
  - PCIe2 4-port (10Gb FCoE and 1GbE) SR&RJ45 Adapter (#EN0H)
  - PCIe2 4-port (10Gb FCoE and 1GbE) SFP+Copper&RJ45 Adapter (#EN0K)
  - PCIe2 4-port (10Gb FCoE and 1GbE) LR&RJ45 Adapter (#EN0M)
  - PCIe3 4-port (10Gb) SR Adapter (#EN15)
  - PCIe3 4-port (10Gb) SFP&Copper Adapter (#EN17)

The minimum operating system requirements, related to SR-IOV functions, are as follows:

- VIOS

  Virtual I/O Server Version 2.2.3.51
- AIX
  - AIX 6.1 Technology Level 9 with Service Pack 5 and APAR IV68443 or later
  - AIX 7.1 Technology Level 3 with Service Pack 5 and APAR IV68444 or later

Firmware level: SR-IOV is supported from firmware level 8.3 for POWER8 processor-based Power E850 servers. Check the Fix Central portal to verify the specific firmware level for your type of the machine at:

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

The entire adapter (all four ports) is configured for SR-IOV or none of the ports are. (FCoE not supported when using SR-IOV).

SR-IOV provides significant performance and usability benefits, as described in the following sections.

All internal PCIe slots in the Power E850 are SR-IOV capable.

3.4.1 Direct access I/O and performance

The primary benefit of allocating adapter functions directly to a partition, as opposed to using a virtual intermediary (VI) like VIOS, is performance. The processing overhead involved in passing client instructions through a VI, to the adapter and back, are substantial.

With direct access I/O, SR-IOV-capable adapters running in shared mode allow the operating system to directly access the slice of the adapter that has been assigned to its partition, so there is no control or data flow through the hypervisor. From the partition perspective, the adapter appears to be physical I/O. Regarding CPU and latency, it exhibits the characteristics of physical I/O. And because the operating system is directly accessing the adapter, if the adapter has special features, like multiple queue support or receive side scaling (RSS), the partition can leverage those also, if the operating system has the capability in its device drivers.

3.4.2 Adapter sharing

The current trend of consolidating servers to reduce cost and improve efficiency is increasing the number of partitions per system, driving a requirement for more I/O adapters per system to accommodate them. SR-IOV addresses and simplifies that requirement by enabling the sharing of SR-IOV-capable adapters. Because each adapter can be shared and directly accessed by up to 48 partitions, depending on the adapter, the partition to PCI slot ratio can be significantly improved without adding the overhead of a virtual intermediary.

3.4.3 Adapter resource provisioning (QoS)

Power Systems SR-IOV provides quality of service (QoS) controls to specify a capacity value for each logical port, improving the ability to share adapter ports effectively and efficiently. The capacity value determines the wanted minimum percentage of the physical port’s resources that should be applied to the logical port.

The exact resource represented by the capacity value can vary based on the physical port type and protocol. In the case of Ethernet physical ports, capacity determines the minimum percentage of the physical port’s transmit bandwidth that the user wants for the logical port.

For example, consider Partitions A, B, and C, with logical ports on the same physical port. If Partition A is assigned an Ethernet logical port with a capacity value of 20%, Partitions B and C cannot use more than 80% of the physical port’s transmission bandwidth unless Partition A is using less than 20%. Partition A can use more than 20% if bandwidth is available. This ensures that, although the adapter is being shared, the partitions maintain their portion of the physical port resources when needed.

3.4.4 Flexible deployment

Power Systems SR-IOV enables flexible deployment configurations, ranging from a simple, single-partition deployment, to a complex, multi-partition deployment involving VIOS partitions and VIOS clients running different operating systems.

In a single-partition deployment, the SR-IOV capable adapter in shared mode is wholly owned by a single partition, and no adapter sharing takes place. This scenario offers no practical benefit over traditional I/O adapter configuration, but the option is available.

In a more complex deployment scenario, an SR-IOV-capable adapter could be shared by both VIOS and non-VIOS partitions, and the VIOS partitions could further virtualize the logical ports as shared Ethernet adapters for VIOS client partitions. This scenario leverages the benefits of direct access I/O, adapter sharing, and QoS that SR-IOV provides, and also the benefits of higher-level virtualization functions, such as Live Partition Mobility (for the VIOS clients), that VIOS can offer.

3.4.5 Reduced costs

SR-IOV facilitates server consolidation by reducing the number of physical adapters, cables, switch ports, and I/O slots required per system. This translates to reduced cost in terms of physical hardware required, and also reduced associated energy costs for power consumption, cooling, and floor space. You may save additional cost on CPU and memory resources, relative to a VIOS adapter sharing solution because SR-IOV does not have the resource overhead inherent in using a virtualization intermediary to interface with the adapters.

3.5 PowerVM

The PowerVM platform is the family of technologies, capabilities, and offerings that delivers industry-leading virtualization on the IBM Power Systems. It is the umbrella branding term for Power Systems virtualization (logical partitioning, IBM Micro-Partitioning®, POWER Hypervisor, VIOS, Live Partition Mobility, and more). As with Advanced POWER Virtualization in the past, PowerVM is a combination of hardware enablement and software. The licensed features of each of the two separate editions of PowerVM are described here.

3.5.1 PowerVM edition

PowerVM editions are available on Power E850 server in Standard (#EPVU) and Enterprise (#EPVV) Edition.

PowerVM supports up to 20 partitions per core, VIOS, and multiple shared processor pools.

> **PowerVM Enterprise Edition:** Adds support for Live Partition Mobility, Active Memory Sharing, and PowerVP performance monitoring. To verify the details about what is specifically supported by the operating system, see Table 3-3 on page 131.

3.5.2 Logical partitions

LPARs and virtualization increase the usage of system resources and add a level of configuration possibilities.

Logical partitioning

Logical partitioning was introduced with the POWER4 processor-based product line and AIX Version 5.1, Red Hat Enterprise Linux 3.0, and SUSE Linux Enterprise Server 9.0 operating systems. This technology was able to divide an IBM eServer™ pSeries (now IBM Power Systems) server into separate logical systems, allowing each LPAR to run an operating environment on dedicated attached devices, such as processors, memory, and I/O components.

Later, dynamic logical partitioning increased the flexibility, allowing selected system resources, such as processors, memory, and I/O components, to be added and deleted from logical partitions while they are running. AIX Version 5.2, with all the necessary enhancements to enable dynamic LPAR, was introduced in 2002. At the same time, Red Hat Enterprise Linux 5 and SUSE Linux Enterprise 9.0 were also able to support dynamic logical partitioning. The ability to reconfigure dynamic LPARs encourages system administrators to dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

Micro-Partitioning

When you use Micro-Partitioning technology, you can allocate fractions of processors to a logical partition. This technology was introduced with POWER5 processor-based systems. A logical partition using fractions of processors is also known as a *shared processor partition* or *micropartition*. Micropartitions run over a set of processors that are called a *shared processor pool*, and virtual processors are used to let the operating system manage the fractions of processing power that are assigned to the logical partition. From an operating system perspective, a virtual processor cannot be distinguished from a physical processor, unless the operating system is enhanced to determine the difference. Physical processors are abstracted into virtual processors that are available to partitions. The meaning of the term *physical processor* in this section is a *processor core*.

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

- The minimum, wanted, and maximum processing units

  Processing units are defined as processing power, or the fraction of time that the partition is dispatched on physical processors. Processing units define the capacity entitlement of the partition.

- The shared processor pool

  Select a pool from the list with the names of each configured shared processor pool. This list also shows, in parentheses, the pool ID of each configured shared processor pool. If the name of the wanted shared processor pool is not available here, you must first configure the shared processor pool by using the shared processor pool Management window. Shared processor partitions use the default shared processor pool, called *DefaultPool by default*. For more information about multiple shared processor pools, see 3.5.3, “Multiple shared processor pools” on page 120.

- Whether the partition can access extra processing power to “fill up” its virtual processors above its capacity entitlement (you select either to cap or uncap your partition)

  If spare processing power is available in the shared processor pool or other partitions are not using their entitlement, an uncapped partition can use additional processing units if its entitlement is not enough to satisfy its application processing demand.

- The weight (preference) if there is an uncapped partition
- The minimum, wanted, and maximum number of virtual processors

The POWER Hypervisor calculates partition processing power based on minimum, wanted, and maximum values, processing mode, and the requirements of other active partitions. The actual entitlement is never smaller than the processing unit's wanted value, but can exceed that value if it is an uncapped partition and up to the number of virtual processors that are allocated.

On the POWER8 processors, a partition can be defined with a processor capacity as small as 0.05 processing units. This number represents 0.05 of a physical core. Each physical core can be shared by up to 20 shared processor partitions, and the partition's entitlement can be incremented fractionally by as little as 0.01 of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC, or IVM.

The Power E850 supports up to 48 cores in a single system. Here are the maximum numbers:

- 48 dedicated partitions
- 960 micropartitions

An important point is that the maximum amounts are supported by the hardware, but the practical limits depend on application workload demands.

Consider the following additional information about virtual processors:

- A virtual processor can be running (dispatched) either on a physical core or as standby waiting for a physical core to became available.
- Virtual processors do not introduce any additional abstraction level. They are only a dispatch entity. When running on a physical processor, virtual processors run at the same speed as the physical processor.
- Each partition's profile defines a CPU entitlement that determines how much processing power any given partition should receive. The total sum of CPU entitlement of all partitions cannot exceed the number of available physical processors in a shared processor pool.
- The number of virtual processors can be changed dynamically through a dynamic LPAR operation.

Processing mode

When you create a logical partition, you can assign entire processors for dedicated use, or you can assign partial processing units from a shared processor pool. This setting defines the processing mode of the logical partition. Figure 3-3 shows a diagram of the concepts that are described in this section.

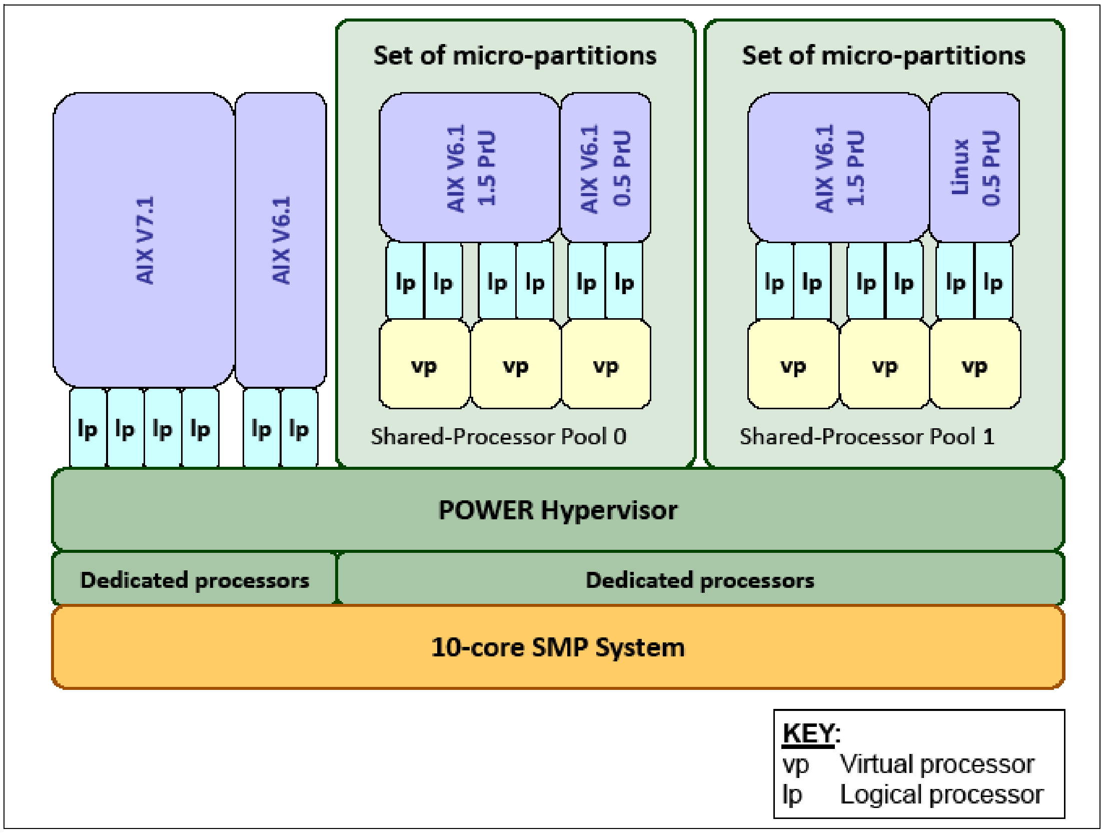


Figure 3-3 Logical partitioning concepts

Dedicated mode

In dedicated mode, physical processors are assigned as a whole to partitions. The simultaneous multithreading feature in the POWER8 processor core allows the core to run instructions from two, four, or eight independent software threads simultaneously.

To support this feature, consider the concept of *logical processors*. The operating system (AIX or Linux) sees one physical core as two, four, or eight logical processors if the simultaneous multithreading feature is on. It can be turned off and on dynamically while the operating system is running (for AIX, run **smtctl**, for Linux, run **ppc64_cpu --smt**). If simultaneous multithreading is off, each physical core is presented as one logical processor in AIX or Linux, and thus only one thread.

Shared dedicated mode

On POWER8 processor technology-based servers, you can configure dedicated partitions to become processor donors for idle processors that they own, allowing for the donation of spare CPU cycles from dedicated processor partitions to a shared processor pool. The dedicated partition maintains absolute priority for dedicated CPU cycles. Enabling this feature can help increase system usage without compromising the computing power for critical workloads in a dedicated processor.

Shared mode

In shared mode, logical partitions use virtual processors to access fractions of physical processors. Shared partitions can define any number of virtual processors (the maximum number is 20 times the number of processing units that are assigned to the partition). From the POWER Hypervisor perspective, virtual processors represent dispatching objects. The POWER Hypervisor dispatches virtual processors to physical processors according to the partition's processing units entitlement. One processing unit represents one physical processor's processing capacity.

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

3.5.3 Multiple shared processor pools

Multiple shared processor pools (MSPPs) are supported on POWER8 processor-based servers. This capability allows a system administrator to create a set of micropartitions with the purpose of controlling the processor capacity that can be consumed from the physical shared processor pool.

Implementing MSPPs depends on a set of underlying techniques and technologies. Figure 3-4 shows an overview of the architecture of multiple shared processor pools.

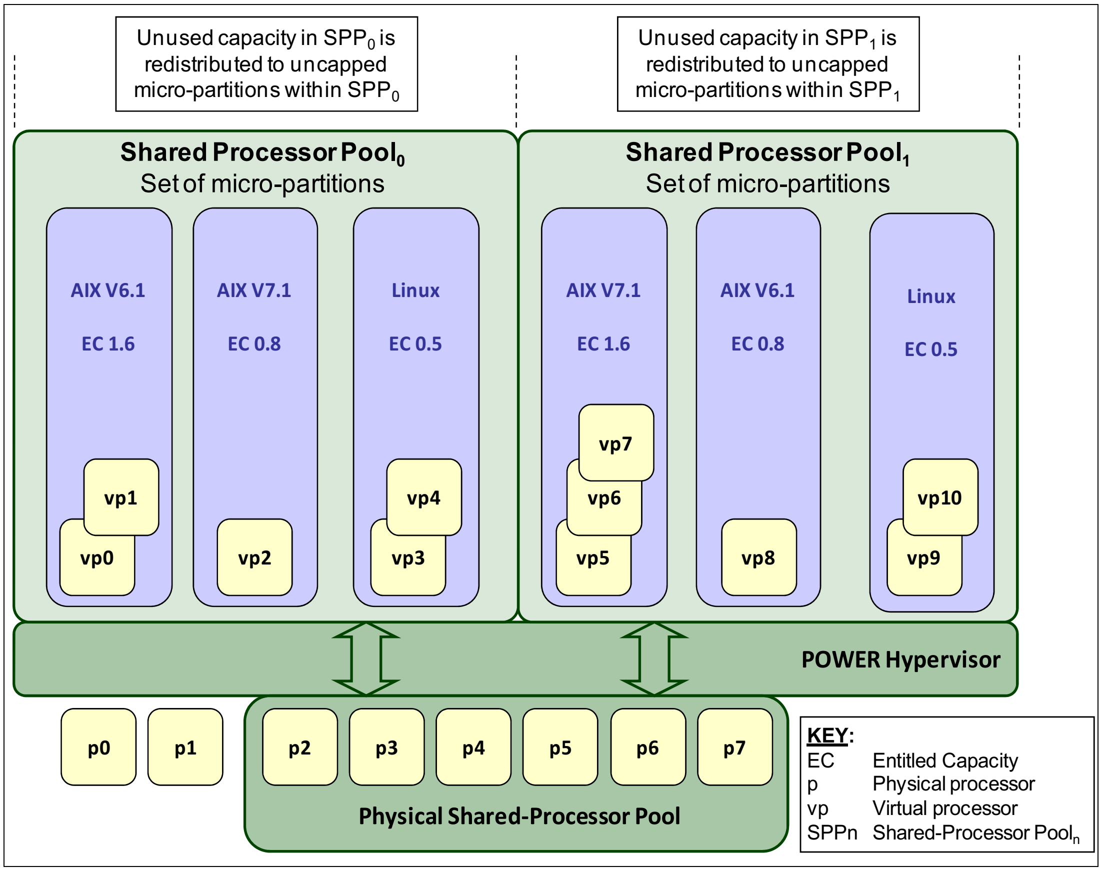

Unused capacity in SPP_0 is redistributed to uncapped micro-partitions within SPP_0
Unused capacity in SPP_1 is redistributed to uncapped micro-partitions within SPP_1
Shared Processor $Pool_0$
Set of micro-partitions
Shared Processor $Pool_1$
Set of micro-partitions
AIX V6.1
EC 1.6
AIX V7.1
EC 0.8
Linux
EC 0.5
AIX V7.1
EC 1.6
AIX V6.1
EC 0.8
Linux
EC 0.5
vp1
vp0
vp2
vp4
vp3
vp7
vp6
vp5
vp8
vp10
vp9
POWER Hypervisor
p0
p1
p2
p3
p4
p5
p6
p7
Physical Shared-Processor Pool
KEY:
EC Entitled Capacity
p Physical processor
vp Virtual processor
SPPn Shared-Processor $Pool_n$

Figure 3-4 Overview of the architecture of multiple shared processor pools

Micropartitions are created and then identified as members of either the default shared processor pool$_0$ or a user-defined shared processor pool$_n$. The virtual processors that exist within the set of micropartitions are monitored by the POWER Hypervisor, and processor capacity is managed according to user-defined attributes.

If the Power Systems server is under heavy load, each micropartition within a shared processor pool is ensured its processor entitlement plus any capacity that it might be allocated from the reserved pool capacity if the micropartition is uncapped.

If certain micropartitions in a shared processor pool do not use their capacity entitlement, the unused capacity is ceded and other uncapped micropartitions within the same shared processor pool are allocated the additional capacity according to their uncapped weighting. In this way, the entitled pool capacity of a shared processor pool is distributed to the set of micropartitions within that shared processor pool.

All Power Systems servers that support the multiple shared processor pools capability have a minimum of one (the default) shared processor pool and up to a maximum of 64 shared processor pools.

3.5.4 Virtual I/O Server

The VIOS is part of all PowerVM editions. It is a special-purpose partition that allows the sharing of physical resources between logical partitions to allow more efficient usage (for example, consolidation). In this case, the VIOS owns the physical resources (SCSI, Fibre Channel, network adapters, and optical devices) and allows client partitions to share access to them, thus minimizing the number of physical adapters in the system.

The VIOS eliminates the requirement that every partition owns a dedicated network adapter, disk adapter, and disk drive. The VIOS supports OpenSSH for secure remote logins. It also provides a firewall for limiting access by ports, network services, and IP addresses. Figure 3-5 shows an overview of a VIOS configuration.

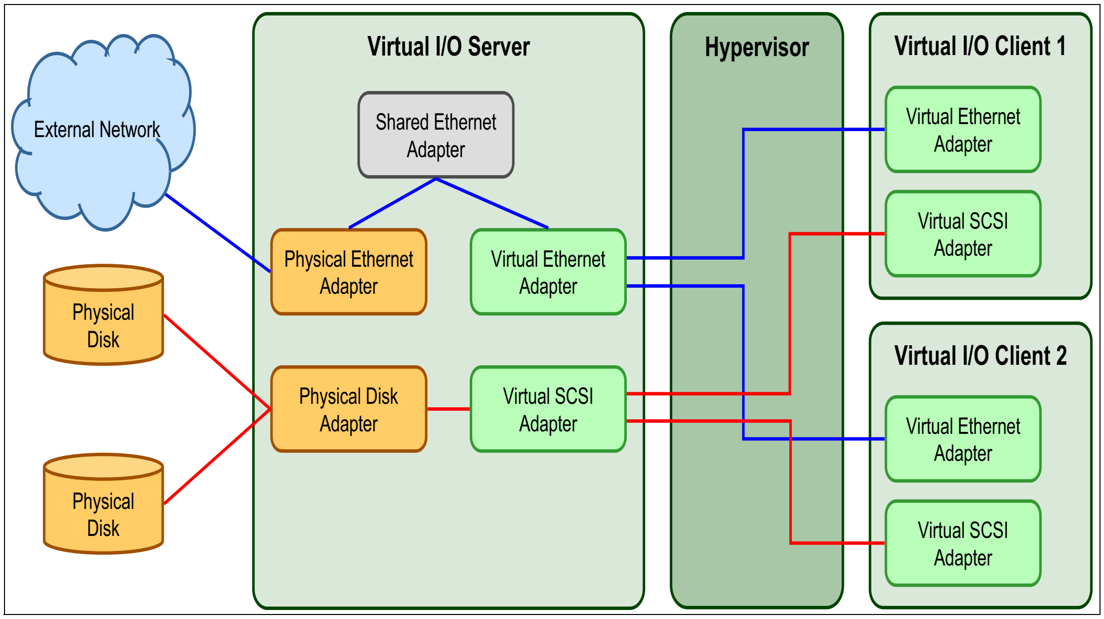


Figure 3-5 Architectural view of the VIOS

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

Installation of the VIOS partition is performed from a special system backup DVD or downloaded image that is provided to clients who order any PowerVM edition. This dedicated software is only for the VIOS, and is supported only in special VIOS partitions. Partially automated VIOS installation can also be performed through HMC, if available. Three major virtual devices are supported by the VIOS:

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

The Virtual Fibre Channel adapter is used with the NPIV feature, as described in 3.5.8, “Operating system support for PowerVM” on page 130.

Shared Ethernet Adapter

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

The SEA also provides the ability for several client partitions to share one physical adapter. With an SEA, you can connect internal and external VLANs by using a physical adapter. The Shared Ethernet Adapter service can be hosted only in the VIOS, not in a general-purpose AIX or Linux partition, and acts as a Layer 2 network bridge to securely transport network traffic between virtual Ethernet networks (internal) and one or more (Etherchannel) physical network adapters (external). These virtual Ethernet network adapters are defined by the POWER Hypervisor on the VIOS.

Figure 3-6 on page 123 shows a configuration example of an SEA with one physical and two virtual Ethernet adapters. An SEA can include up to 16 virtual Ethernet adapters on the VIOS that shares the physical access.

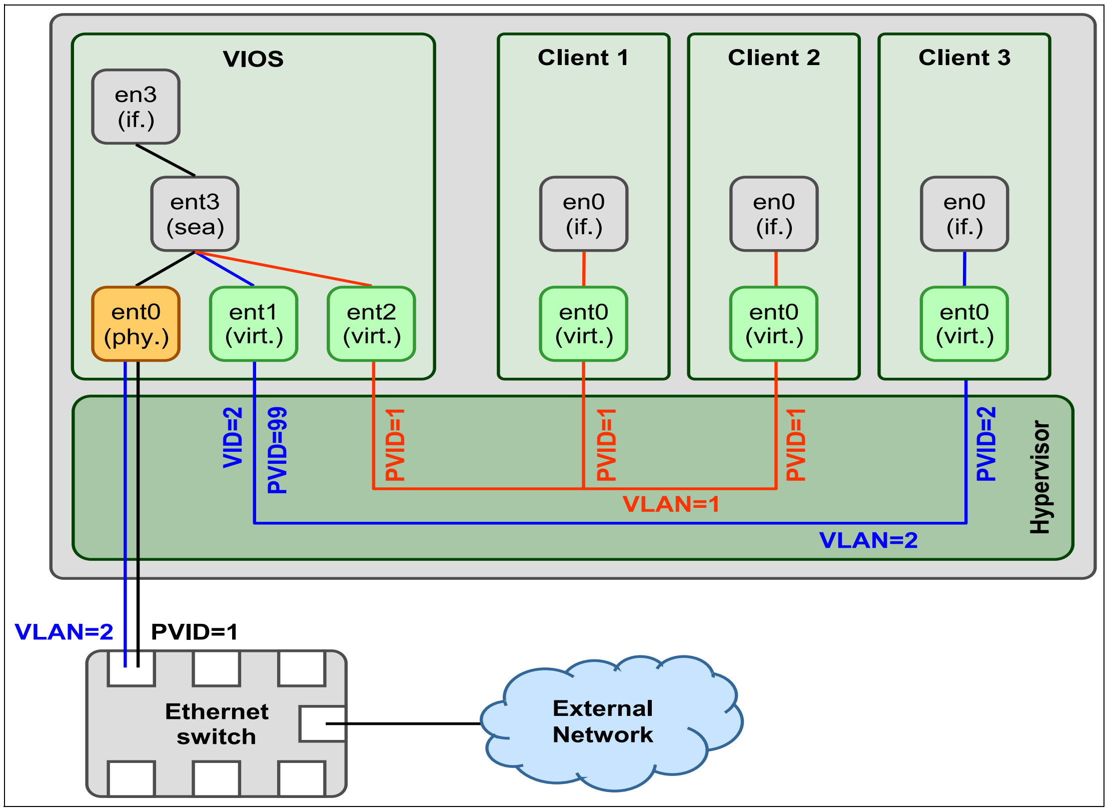


Figure 3-6 Architectural view of a Shared Ethernet Adapter

A single SEA setup can have up to 16 virtual Ethernet trunk adapters and each virtual Ethernet trunk adapter can support up to 20 VLAN networks. Therefore, a possibility is for a single physical Ethernet adapter to be shared between 320 internal VLAN networks. The number of shared Ethernet adapters that can be set up in a VIOS partition is limited only by the resource availability because there are no configuration limits.

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

Virtual SCSI

Virtual SCSI is used to see a virtualized implementation of the SCSI protocol. Virtual SCSI is based on a client/server relationship. The VIOS logical partition owns the physical resources and acts as a server or, in SCSI terms, a target device. The client logical partitions access the virtual SCSI backing storage devices that are provided by the VIOS as clients.

The virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured by using an HMC or through the Integrated Virtualization Manager. The virtual SCSI server (target) adapter is responsible for running any SCSI commands that it receives. It is owned by the VIOS partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN-attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the VIOS.

Physical disks that are presented to the Virtual I/O Server can be exported and assigned to a client partition in various ways:

- The entire disk is presented to the client partition.
- The disk is divided into several logical volumes, which can be presented to a single client or multiple clients.
- As of VIOS 1.5, files can be created on these disks, and file-backed storage devices can be created.

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

For more information about specific storage devices that are supported for VIOS, see the following website:

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

N_Port ID Virtualization

N_Port ID Virtualization (NPIV) is a technology that allows multiple logical partitions to access independent physical storage through the same physical Fibre Channel adapter. This adapter is attached to a VIOS partition that acts only as a pass-through, managing the data transfer through the POWER Hypervisor.

Each partition that uses NPIV is identified by a pair of unique worldwide port names, enabling you to connect each partition to independent physical storage on a SAN. Unlike virtual SCSI, only the client partitions see the disk.

For more information and requirements for NPIV, see the following resources:

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

Virtual I/O Server functions

The VIOS has many features, including monitoring solutions and the following capabilities:

- Support for Live Partition Mobility starting on POWER6 processor-based systems with the PowerVM Enterprise Edition. For more information about Live Partition Mobility, see 3.5.6, “Active Memory Sharing” on page 129.
- Support for virtual SCSI devices that are backed by a file, which are then accessed as standard SCSI-compliant LUNs.
- Support for virtual Fibre Channel devices that are used with the NPIV feature.
- Virtual I/O Server Expansion Pack with additional security functions, such as Kerberos (Network Authentication Service for users and client and server applications), Simple Network Management Protocol (SNMP) v3, and Lightweight Directory Access Protocol (LDAP) client function.
- System Planning Tool (SPT) and Workload Estimator, which are designed to ease the deployment of a virtualized infrastructure. For more information about the System Planning Tool, see 3.6, “System Planning Tool” on page 139.
- IBM Systems Director agent and several preinstalled IBM Tivoli® agents, such as the following examples:
  - IBM Tivoli Identity Manager, which allows easy integration into an existing Tivoli Systems Management infrastructure

  - IBM Tivoli Application Dependency Discovery Manager (ADDM), which creates and automatically maintains application infrastructure maps, including dependencies, change histories, and deep configuration values

- vSCSI enterprise reliability, availability, and serviceability (eRAS).
- Additional CLI statistics in **`svmon`**, **`vmstat`**, **`fcstat`**, and **`topas`**.
- The VIOS Performance Advisor tool provides advisory reports based on key performance metrics for various partition resources that are collected from the VIOS environment.
- Monitoring solutions to help manage and monitor the VIOS and shared resources. Commands and views provide additional metrics for memory, paging, processes, Fibre Channel HBA statistics, and virtualization.

For more information about the VIOS and its implementation, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

3.5.5 Live Partition Mobility

Live Partition Mobility (LPM) is a technique that allows a partition running on one server to be migrated dynamically to another server.

This feature can be extremely useful in a situation where a system needs to be evacuated for maintenance but its partitions do not allow for downtime. LPM allows for all the partitions to be moved while running to another server so the system can be properly shut down without impacts for the applications.

In simplified terms, LPM typically works in an environment where all of the I/O from one partition is virtualized through PowerVM and VIOS and all partition data is stored in a storage area network (SAN) accessed by both servers.

To migrate a partition from one server to another, a partition is identified on the new server and configured to have the same virtual resources as the primary server including access to the same logical volumes as the primary using the SAN.

When an LPM migration is initiated on a server for a partition, PowerVM in the first system starts copying the state of memory in the first partition over to a destination partition in another server through PowerVM on each system. This is done across a LAN while the initial partition continues to run. PowerVM has control of I/O operations through I/O virtualization and keeps track of memory changes that occur throughout the process.

At some point, when all of the memory state is copied from the primary partition, the primary partition is paused. PowerVM in the second server takes over control of the shared storage resources and allows the partition now running in that server to resume processing at the point where the first server left off.

Thinking in terms of using LPM for hardware repairs, if all of the workloads on a server are migrated by LPM to other servers, after all have been migrated, the first server could be turned off to repair components.

LPM can also be used for performing firmware upgrades or adding additional hardware to a server when the hardware cannot be added concurrently in addition to software maintenance within individual partitions.

In successful LPM situations, while there may be a short time when applications are not processing new workload, the applications do not fail or crash and do not need to be restarted.

Minimum configuration

For LPM to work, it is necessary that the system containing a partition to be migrated, and the system being migrated to, both have a local LAN connection using a virtualized LAN adapter. The LAN adapter should be high speed for better migration performance. The LAN used should be a local network and should be private and have only two uses. The first is for communication between servers; the second is for communication between partitions on each server and the HMC for resource monitoring and control functions (RMC.)

LPM also needs all systems in the LPM cluster to be attached to the same SAN (when using SAN for required common storage), which typically requires use of Fibre Channel adapters.

If a single HMC is used to manage both systems in the cluster, connectivity to the HMC also needs to be provided by an Ethernet connection to each service processor.

The LAN and SAN adapters used by the partition must be assigned to a Virtual I/O Server and the partitions access to these would be by virtual LAN (VLAN) and virtual SCSI (vSCSI) connections within each partition to the VIOS.

Each partition to be migrated must only use virtualized I/O through a VIOS. There can be no non-virtualized adapters assigned to such partitions.

A diagram with the minimum requirements can be seen in Figure 3-7.

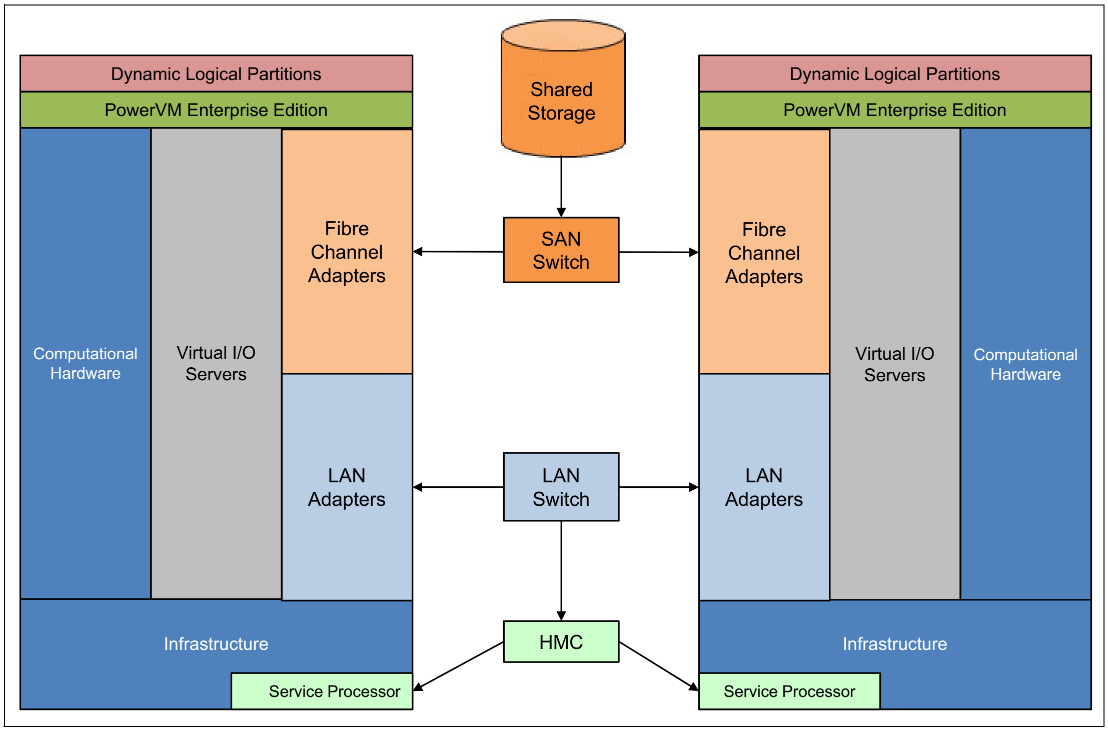


Figure 3-7 Minimum Live Partition Mobility requirements

Suggested configuration

LPM connectivity in the minimum configuration discussion is vulnerable to a number of different hardware and firmware faults that would lead to the inability to migrate partitions. Multiple paths to networks and SANs are therefore advised. To accomplish this, a VIOS server can be configured to use dual Fibre Channel and LAN adapters.

Externally to each system, redundant Hardware Management Consoles (HMCs) can be used for greater availability. There can also be options to maintain redundancy in SANs and local network hardware. A diagram with the suggested scenario can be seen in Figure 3-8.

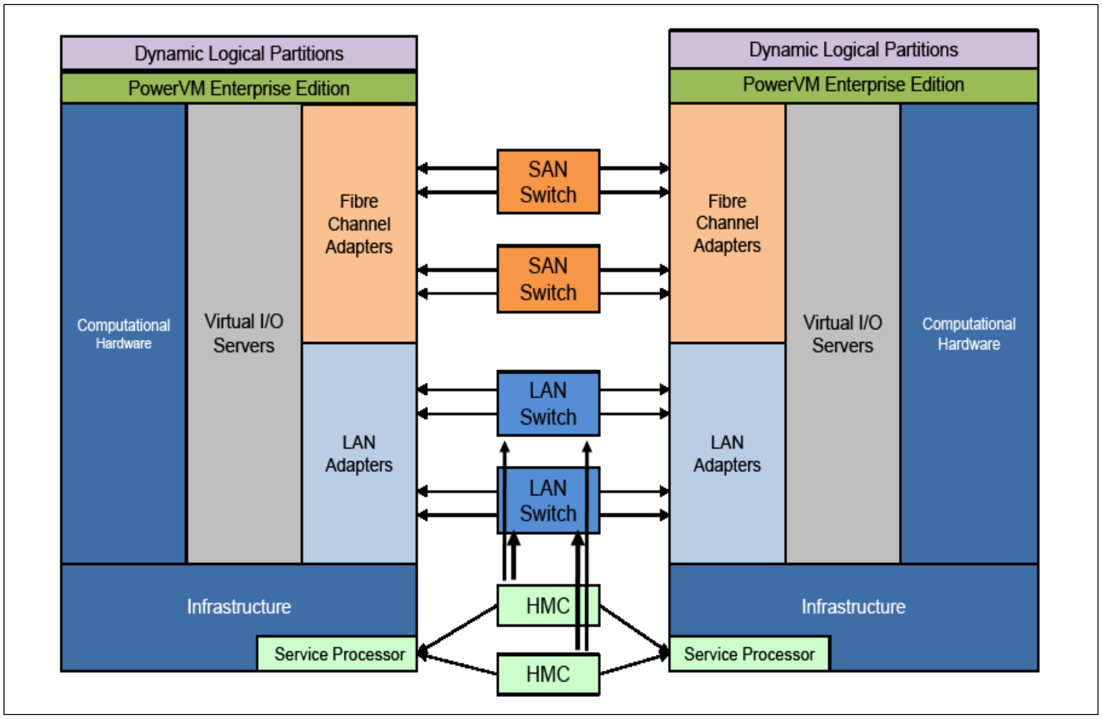


Figure 3-8 Redundant infrastructure for LPM

PCIe slot selection

The POWER8 processor has PCIe controllers integrated on the chip allowing for a single processor to have two or more PCIe Gen3 slots directly attached to it. In a complete processor failure, these slots might become unusable.

When this affects availability of PCIe slots, it must be considered while selecting the slot placement for the adapters on the Virtual I/O Servers.

The Power E850 server has up to 11 PCIe hot-plug Gen3 slots, which depends on the number of processor modules, providing excellent configuration flexibility and expandability. Eight adapter slots are x16 Gen3, and three adapter slots are x8 Gen3.

Figure 3-9 on page 128 illustrates how such a system could be configured to maximize redundancy in a VIOS environment, presuming that the boot disks for each VIOS are accessed from storage area networks.

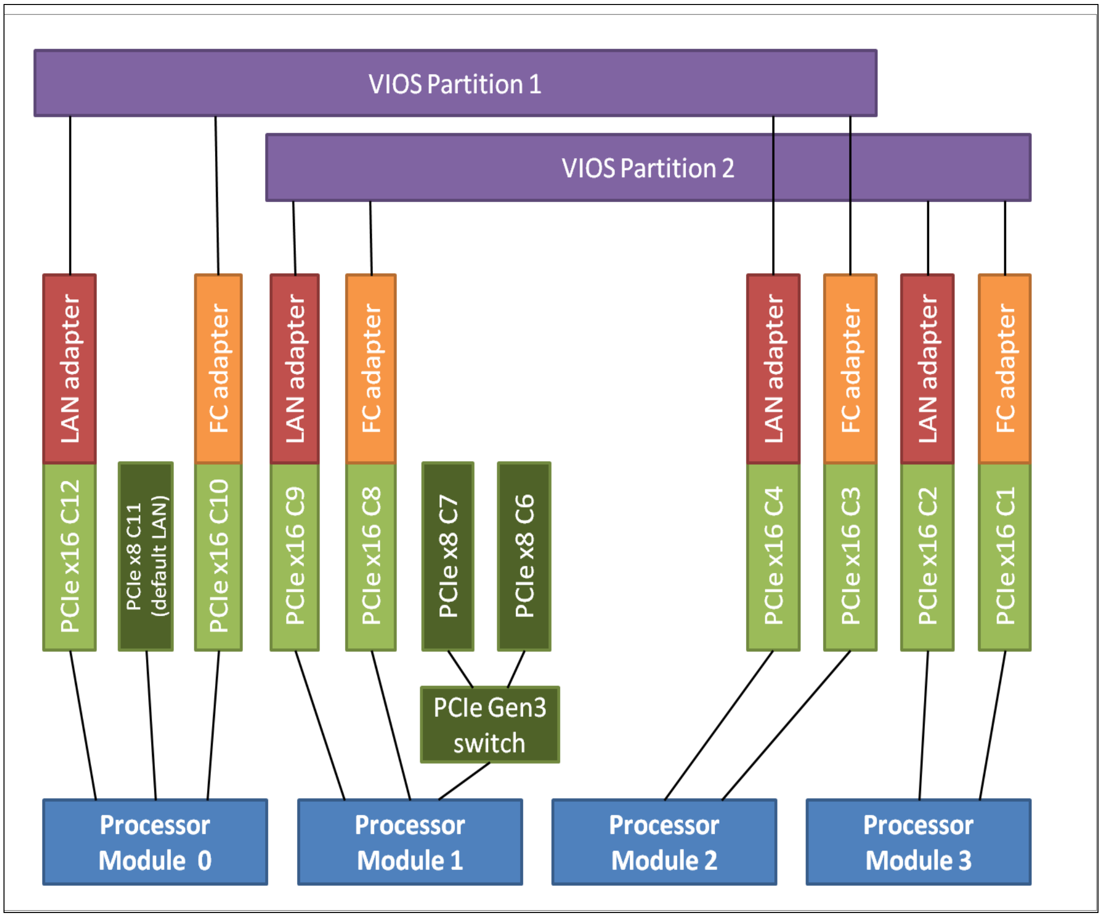


Figure 3-9 I/O subsystem of a 4-socket E850 system

On the Power E850, I/O drawers for expanding I/O are supported. When these drawers are used, all the slots connected to a PCIe Adapter for Expansion Drawer are bound to a given processor. A similar concept for I/O redundancy can be used to maximize availability of I/O access using two Fan-Out Modules, one connected to each of four processor sockets in the system. A logical diagram with the PCIe slots and its processors can be seen in Figure 3-10.

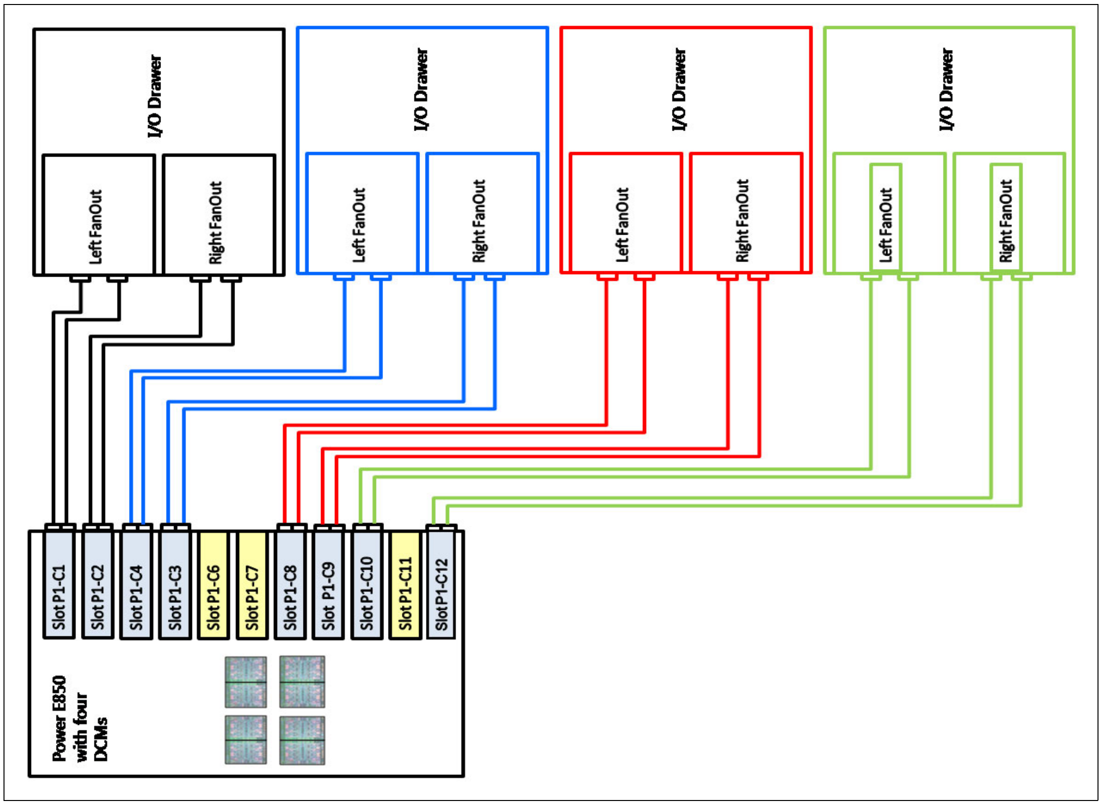


Figure 3-10 Logical diagram of processors and its associated PCIe slots

More information

For more information about PowerVM and Live Partition Mobility, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940:

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

3.5.6 Active Memory Sharing

Active Memory Sharing is an IBM PowerVM advanced memory virtualization technology that provides system memory virtualization capabilities to IBM Power Systems, allowing multiple partitions to share a common pool of physical memory.

Active Memory Sharing is available only with the Enterprise version of PowerVM.

The physical memory of an IBM Power System can be assigned to multiple partitions in either dedicated or shared mode. The system administrator can assign some physical memory to a partition and some physical memory to a pool that is shared by other partitions. A single partition can have either dedicated or shared memory:

- With a pure dedicated memory model, the system administrator's task is to optimize available memory distribution among partitions. When a partition suffers degradation because of memory constraints and other partitions have unused memory, the administrator can manually issue a dynamic memory reconfiguration.
- With a shared memory model, the system automatically decides the optimal distribution of the physical memory to partitions and adjusts the memory assignment based on partition

load. The administrator reserves physical memory for the shared memory pool, assigns partitions to the pool, and provides access limits to the pool.

Active Memory Sharing can be used to increase memory usage on the system either by decreasing the global memory requirement or by allowing the creation of additional partitions on an existing system. Active Memory Sharing can be used in parallel with Active Memory Expansion on a system running a mixed workload of several operating systems. For example, AIX partitions can take advantage of Active Memory Expansion. Other operating systems take advantage of Active Memory Sharing also.

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

3.5.7 Active Memory Deduplication

In a virtualized environment, the systems might have a considerable amount of duplicated information that is stored on RAM after each partition has its own operating system, and some of them might even share the same kinds of applications. On heavily loaded systems, this behavior might lead to a shortage of the available memory resources, forcing paging by the Active Memory Sharing partition operating systems, the Active Memory Deduplication pool, or both, which might decrease overall system performance.

Active Memory Deduplication allows the POWER Hypervisor to map dynamically identical partition memory pages to a single physical memory page within a shared memory pool. This enables a better usage of the Active Memory Sharing shared memory pool, increasing the system's overall performance by avoiding paging. Deduplication can cause the hardware to incur fewer cache misses, which also leads to improved performance.

Active Memory Deduplication depends on the Active Memory Sharing feature being available, and it consumes CPU cycles that are donated by the Active Memory Sharing pool's VIOS partitions to identify deduplicated pages. The operating systems that are running on the Active Memory Sharing partitions can "hint" to the POWER Hypervisor that some pages (such as frequently referenced read-only code pages) are good for deduplication.

To perform deduplication, the hypervisor cannot compare every memory page in the Active Memory Sharing pool with every other page. Instead, it computes a small signature for each page that it visits and stores the signatures in an internal table. Each time that a page is inspected, a look-up of its signature is done in the known signatures in the table. If a match is found, the memory pages are compared to be sure that the pages are really duplicates. When a duplicate is found, the hypervisor remaps the partition memory to the existing memory page and returns the duplicate page to the Active Memory Sharing pool.

From the LPAR perspective, the Active Memory Deduplication feature is not apparent. If an LPAR attempts to modify a deduplicated page, the Power Hypervisor grabs a free page from the Active Memory Sharing pool, copies the duplicate page contents into the new page, and maps the LPAR's reference to the new page so the LPAR can modify its own unique page.

For more information regarding Active Memory Deduplication, see *Power Systems Memory Deduplication*, REDP-4827.

3.5.8 Operating system support for PowerVM

Table 3-3 on page 131 shows operating system support for virtualization features.

Table 3-3 Virtualization features supported by AIX, and Linux

Feature	AIX 6.1 TL8& TL9 SP1	AIX 7.1 TL2& TL3 SP1	RHEL 6.6	RHEL 7.1	SUSE 11 SP3	SUSE 12	Ubuntu 15.04
Virtual SCSI	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Ethernet	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared Ethernet Adapter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Fibre Channel	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Tape	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Logical partitioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR I/O adapter add/remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR processor add/remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR memory add	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DLPAR memory remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Micro-Partitioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared dedicated capacity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Shared Processor Pools	Yes	Yes	Yes	Yes	Yes	Yes	Yes
VIOS	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Integrated Virtualization Manager	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Suspend/resume and hibernation[a]	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared Storage Pools	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Thin provisioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Active Memory Sharing[b]	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Active Memory Deduplication	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Live Partition Mobility	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Simultaneous multithreading (SMT)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Active Memory Expansion	Yes	Yes	No	No	No	No	No

[a] At the time of writing, Suspend/Resume is not available. Check with your IBM System Services Representative (SSR) for availability on POWER8 platforms.

[b] At the time of writing, Active Memory Sharing when used with Live Partition Mobility is not supported. Check with your IBM SSR for availability on POWER8 platforms.

For more information about specific features for Linux, see the following website:

http://www.ibm.com/support/knowledgecenter/linuxonibm/liaam/supportedfeaturesforlinuxonpowersystemsservers.htm

3.5.9 Linux support

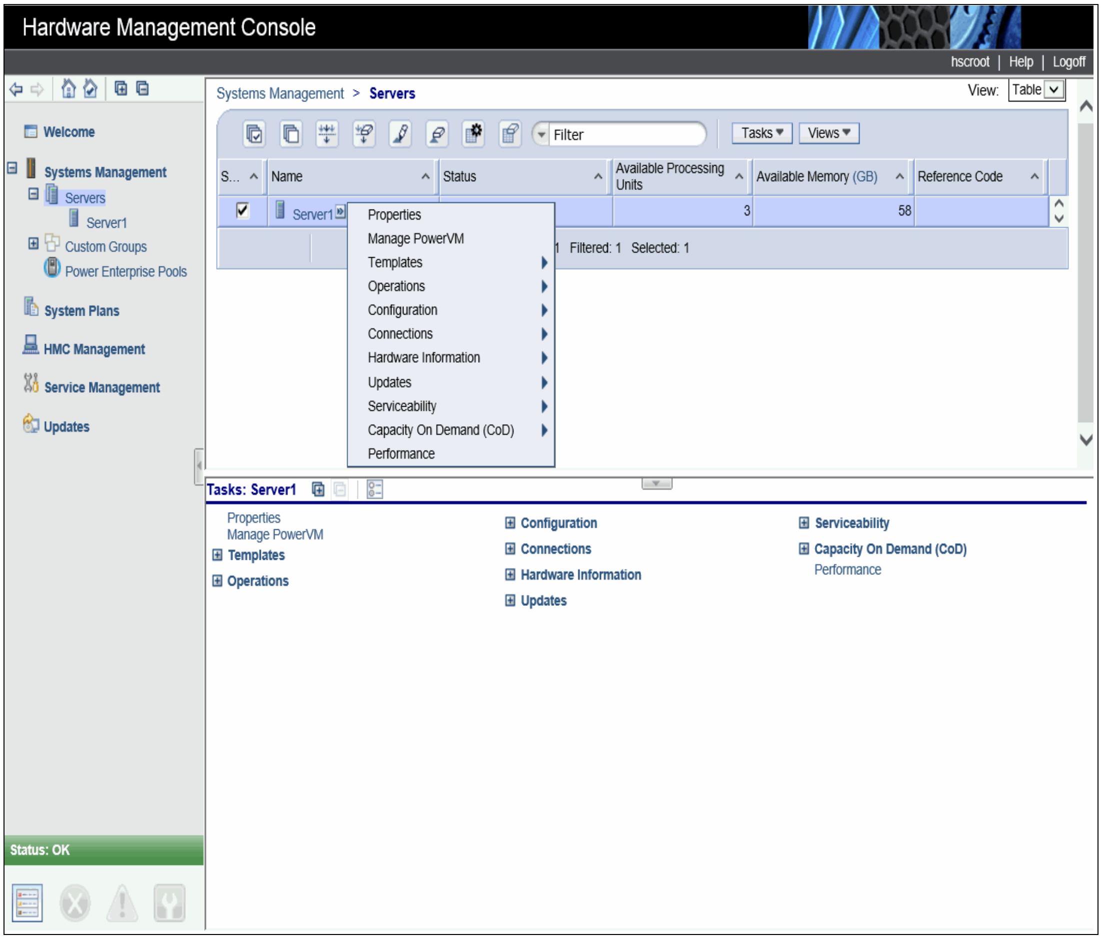
The IBM Linux Technology Center (LTC) contributes to the development of Linux by providing support for IBM hardware in Linux distributions. In particular, the LTC has available tools and code for the Linux communities to take advantage of the POWER8 technology and develop POWER8 optimized software.

For more information about specific Linux distributions, see the following website:

http://www.ibm.com/support/knowledgecenter/linuxonibm/liaam/liaamdistros.htm

3.5.10 PowerVM simplification

With the availability of HMC V8R8.2.0, PowerVM simplification was introduced. Figure 3-11 shows new options in HMC V8R8.2.0 that are available for PowerVM simplification. Selecting **Manage PowerVM** opens a new window, where a user can view and manage all the aspects of a PowerVM configuration, such as SEA, virtual networks, and virtual storage using the graphical interface only.



S...	Name	Status	Available Processing Units	Available Memory (GB)	Reference Code
☑	Server1		3	58	

Figure 3-11 PowerVM new level task

Templates

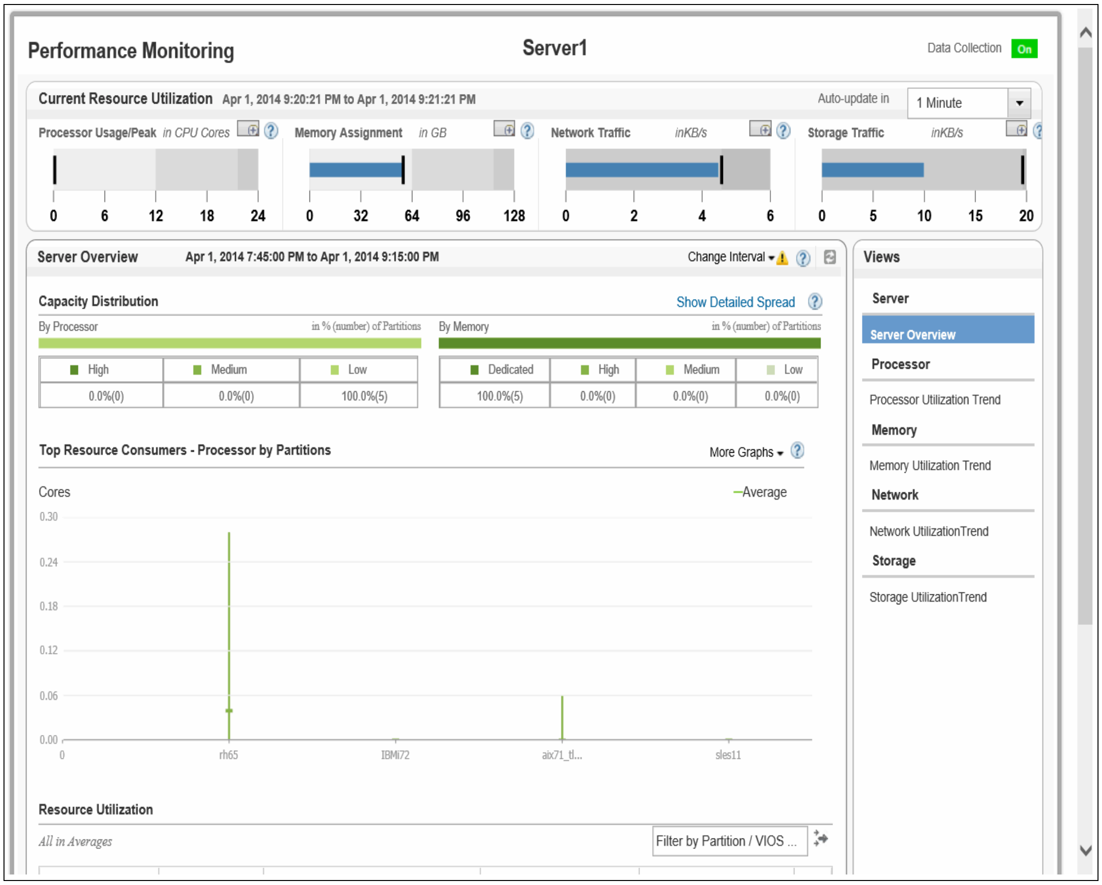
Templates allow you to specify the configuration details for the system I/O, memory, storage, network, processor, and other partition resources. A user can use the predefined or captured templates that are available in the template library to deploy a system. Two types of templates are available in the template library:

- The *System Template* contains configuration details for system resources, such as system I/O, memory, storage, processors, network, and physical I/O adapters. You can use these system templates to deploy the system.
- The *Partition Template* contains configuration details for the partition resources. A user can configure logical partitions with the predefined templates or by creating custom templates.

By using these options, a user can deploy a system, select a System Template, and click **Deploy**. After deploying a system, a user can choose a Partition Template from the library.

Performance

The Performance function opens the Performance and Capacity Monitoring window, as shown in Figure 3-12 and Figure 3-13 on page 134.



Performance Monitoring Server1 Data Collection On

Current Resource Utilization Apr 1, 2014 9:20:21 PM to Apr 1, 2014 9:21:21 PM Auto-update in 1 Minute

Server Overview Apr 1, 2014 7:45:00 PM to Apr 1, 2014 9:15:00 PM

Capacity Distribution

By Processor in % (number) of Partitions

High	Medium	Low
0.0%(0)	0.0%(0)	100.0%(5)

By Memory in % (number) of Partitions

Dedicated	High	Medium	Low
100.0%(5)	0.0%(0)	0.0%(0)	0.0%(0)

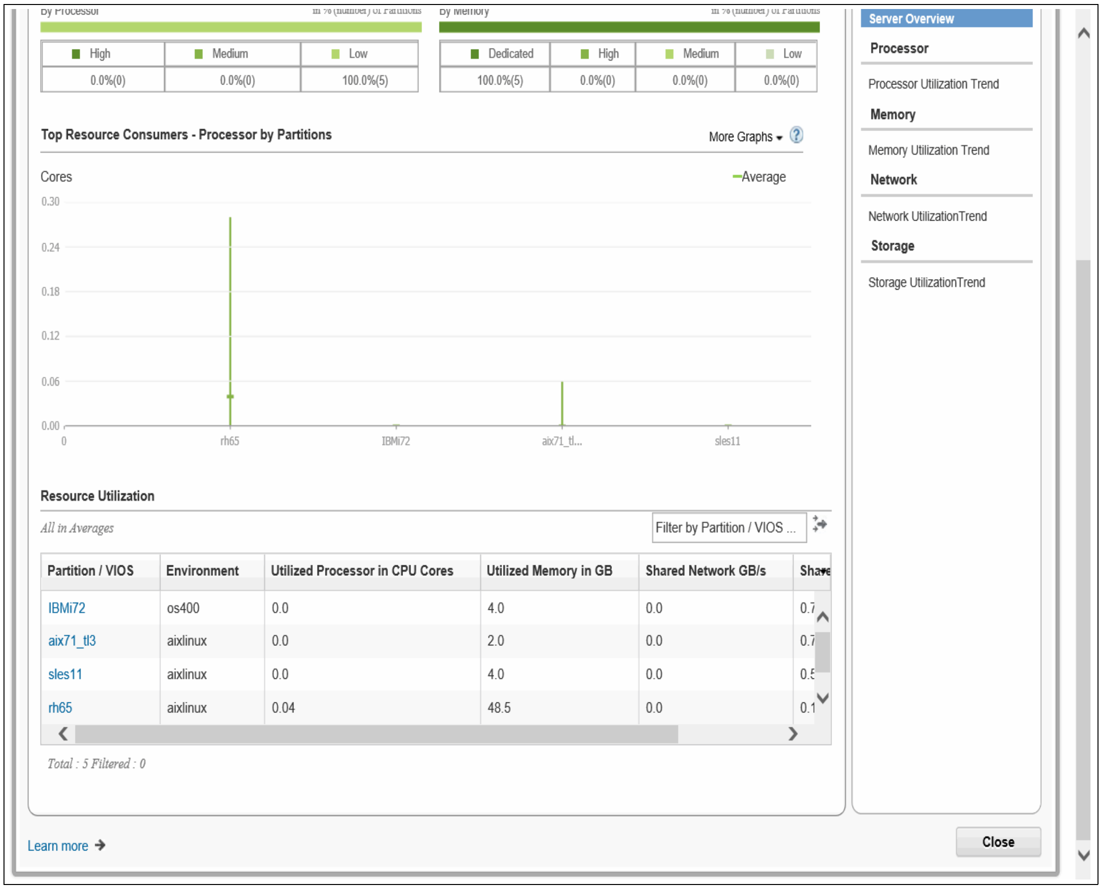
Top Resource Consumers - Processor by Partitions

Resource Utilization

All in Averages

Figure 3-12 HMC performance monitoring CPU Memory assignment

Figure 3-13 shows the bottom of the window where performance and capacity data are presented in a graphical format.



By Processor — in % (number) of Partitions

■ High	■ Medium	■ Low
0.0%(0)	0.0%(0)	100.0%(5)

By Memory — in % (number) of Partitions

■ Dedicated	■ High	■ Medium	■ Low
100.0%(5)	0.0%(0)	0.0%(0)	0.0%(0)

Resource Utilization

All in Averages

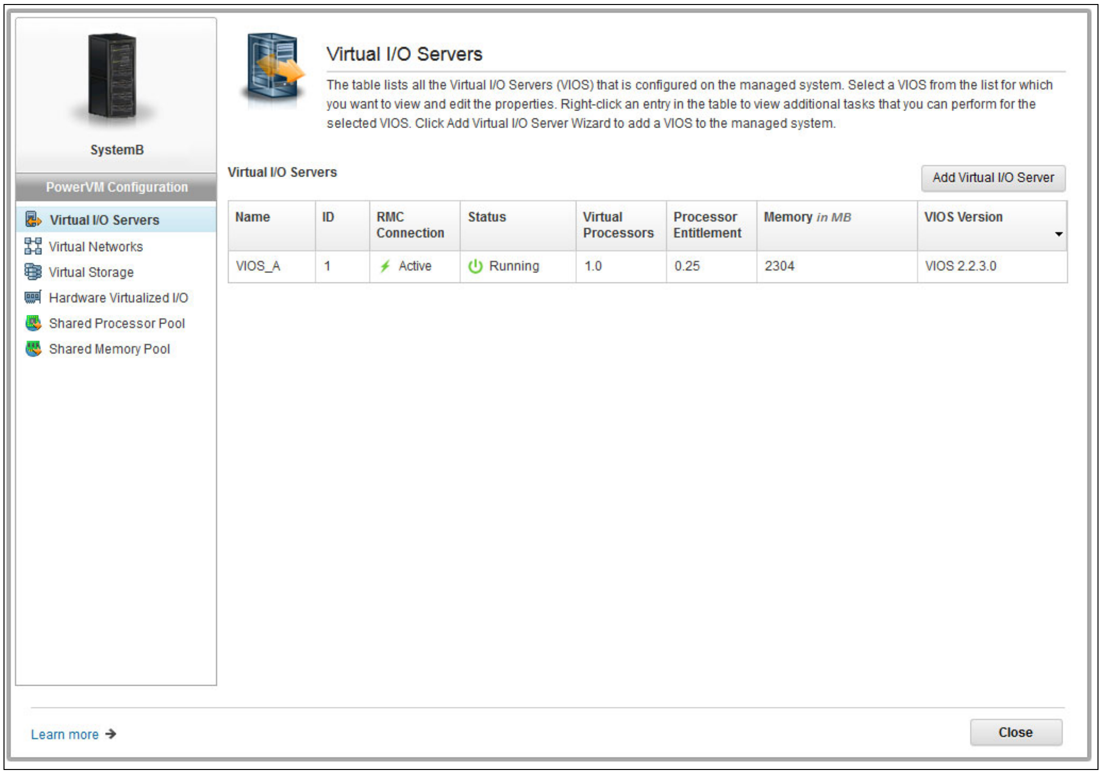
Partition / VIOS	Environment	Utilized Processor in CPU Cores	Utilized Memory in GB	Shared Network GB/s	Share
IBMi72	os400	0.0	4.0	0.0	0.7
aix71_tl3	aixlinux	0.0	2.0	0.0	0.7
sles11	aixlinux	0.0	4.0	0.0	0.5
rh65	aixlinux	0.04	48.5	0.0	0.1

Total : 5 Filtered : 0

Figure 3-13 HMC performance monitoring CPU Memory assignment

New GUI functions

As part of the enhancements introduced in this version, this new capability aids in the management of PowerVM from a single user interface (UI). Figure 3-14 shows how all virtualization components are now displayed in a single UI.



Virtual I/O Servers

The table lists all the Virtual I/O Servers (VIOS) that is configured on the managed system. Select a VIOS from the list for which you want to view and edit the properties. Right-click an entry in the table to view additional tasks that you can perform for the selected VIOS. Click Add Virtual I/O Server Wizard to add a VIOS to the managed system.

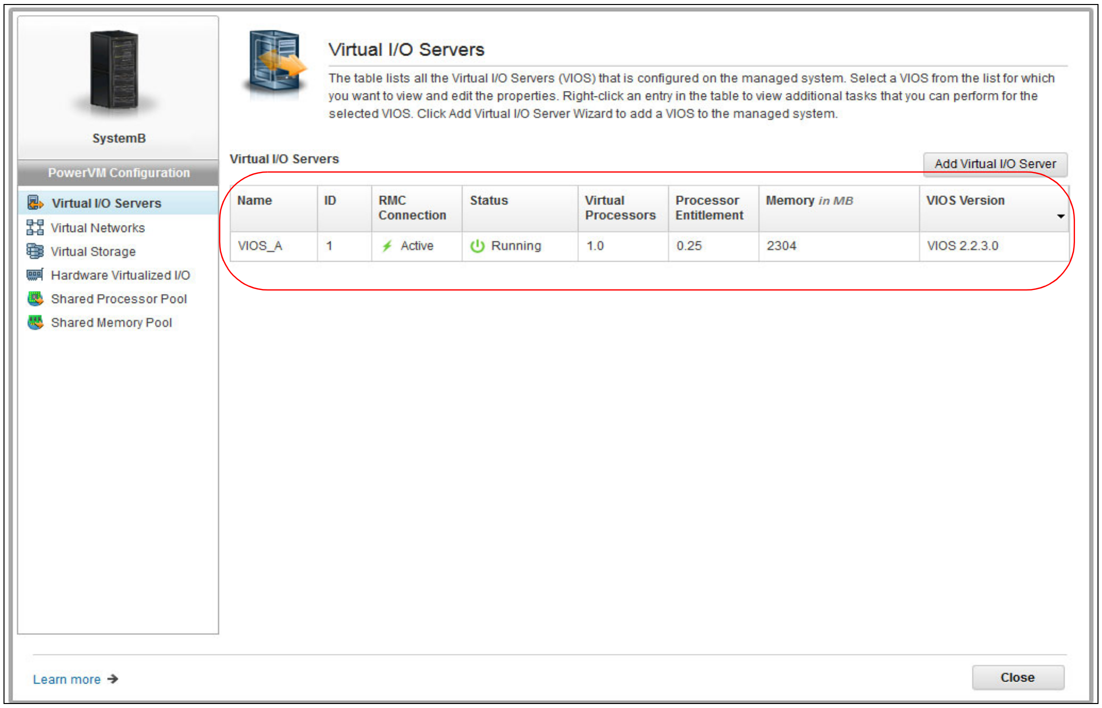
Virtual I/O Servers

Name	ID	RMC Connection	Status	Virtual Processors	Processor Entitlement	Memory *in MB*	VIOS Version
VIOS_A	1	⚡ Active	⏻ Running	1.0	0.25	2304	VIOS 2.2.3.0

Figure 3-14 VIO servers, Virtual Network, and Virtual Storage are accessible from the UI

3.5.11 Single view of all Virtual I/O servers

The *Manage Virtual I/O Servers* function displays a list of VIOS that is configured in the managed system. It also displays information about each VIOS configuration, such as allocated memory, allocated processing units, allocated virtual processors, and status. This can be seen in Figure 3-15.



Virtual I/O Servers

The table lists all the Virtual I/O Servers (VIOS) that is configured on the managed system. Select a VIOS from the list for which you want to view and edit the properties. Right-click an entry in the table to view additional tasks that you can perform for the selected VIOS. Click Add Virtual I/O Server Wizard to add a VIOS to the managed system.

Virtual I/O Servers

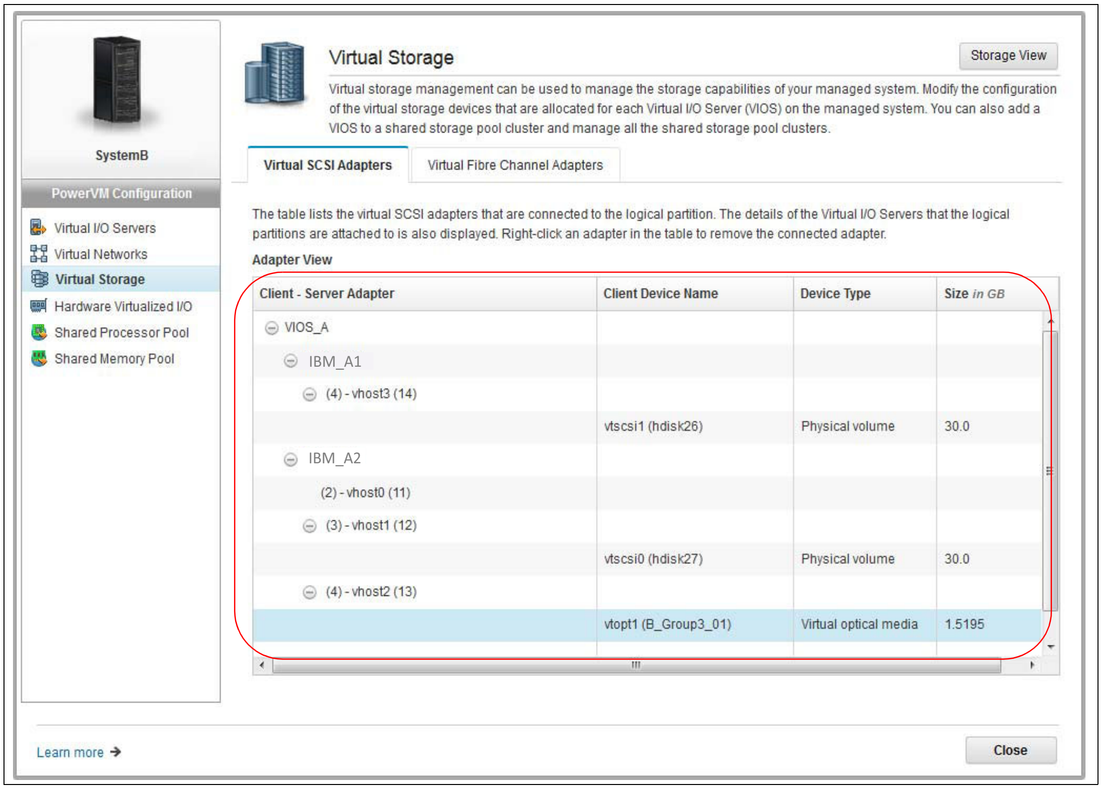
Name	ID	RMC Connection	Status	Virtual Processors	Processor Entitlement	Memory *in MB*	VIOS Version
VIOS_A	1	Active	Running	1.0	0.25	2304	VIOS 2.2.3.0

Figure 3-15 Virtual I/O Server properties displayed in a single UI

3.5.12 Virtual Storage Management

HMC V8R8.3.0 allows you to manage and monitor storage devices in a PowerVM virtual storage environment. Is possible to change the configuration of the virtual storage devices that are allocated to each VIOS on the managed system. Also it lets you add a VIOS to a shared storage pool cluster and manage all the shared storage pool clusters.

The GUI lets the user view the adapter configuration of the virtual storage devices that are allocated to the VIOS. The adapter view provides a mapping of the adapters to the physical storage device. By selecting a VIOS, you can manage the virtual storage devices that are configured to a particular partition and select and view all the partitions with storage provisioned by the VIOS. Figure 3-16 shows a single VIOS scenario managing multiple logical partitions and its virtual SCSI devices.



Virtual Storage

Virtual storage management can be used to manage the storage capabilities of your managed system. Modify the configuration of the virtual storage devices that are allocated for each Virtual I/O Server (VIOS) on the managed system. You can also add a VIOS to a shared storage pool cluster and manage all the shared storage pool clusters.

The table lists the virtual SCSI adapters that are connected to the logical partition. The details of the Virtual I/O Servers that the logical partitions are attached to is also displayed. Right-click an adapter in the table to remove the connected adapter.

Adapter View

Client - Server Adapter	Client Device Name	Device Type	Size *in GB*
⊖ VIOS_A			
⊖ IBM_A1			
⊖ (4) - vhost3 (14)			
	vtscsi1 (hdisk26)	Physical volume	30.0
⊖ IBM_A2			
(2) - vhost0 (11)			
⊖ (3) - vhost1 (12)			
	vtscsi0 (hdisk27)	Physical volume	30.0
⊖ (4) - vhost2 (13)			
	vtopt1 (B_Group3_01)	Virtual optical media	1.5195

Figure 3-16 Adapter assignment to each partition

3.5.13 Virtual Network Management

HMC V8R8.8.3.0 helps to manage PowerVM virtual networks through a UI. This UI uses a defined set of concepts about networking technologies with specific terminology introduced by IBM Power Architecture®.

As part of Virtual Network Management, new functions are added to perform actions on the virtual network devices, such as these:

- Add Virtual Network wizard
- Network Bridge Management
- Link Aggregation Management
- PowerVM networking concepts review

Add Virtual Network wizard

Use the **Add Virtual Network** wizard button in the HMC to add an existing virtual network or a new virtual network to the server.

The following tasks can be completed by using the Add Virtual Network wizard:

- Create internal or bridged networks
- Create tagged or untagged virtual networks
- Create a virtual network on an existing or a new virtual switch
- Create a load group or select an existing load group

Network Bridge Management

From a server that is managed by the HMC, it is possible to change the PowerVM virtual network bridge properties. Following are the changes that are allowed:

- Enable or disable network failover in the **Failover** field.
- Enable or disable load balancing in the **Load Balance** field.
- Change the primary VIOS and the physical adapter location from the table.
- Enable Jumbo Frame in the network bridge for the virtual Ethernet adapter to communicate to an external network.
- Enable QoS in the network bridge to check the priority value of all tagged packets and arrange those packets in the corresponding queue.

Link Aggregation Management

A link aggregation device can be added on the VIOS by using the Add Link Aggregation device wizard. The same wizard can be used to change a link aggregation device's properties or remove a link aggregation device.

PowerVM networking concepts review

PowerVM includes extensive and powerful networking tools and technologies, which can be used to enable more flexibility, better security, and enhanced usage of hardware resources. Some of these terms and concepts are unique to the Power Architecture. Table 3-4 introduces the PowerVM virtual networking technologies.

Table 3-4 PowerVM Network technologies

PowerVM technology	Definition
Virtual Network	Enables interpartition communication without assigning a physical network adapter to each partition. If the virtual network is bridged, partitions can communicate with external networks. A virtual network is identified by its name or VLAN ID and the associated virtual switch.
Virtual Ethernet adapter	Enables a client logical partition to send and receive network traffic without a physical Ethernet adapter.
Virtual switch	An in-memory, hypervisor implementation of a layer-2 switch.
Network bridge	A software adapter that bridges physical and virtual networks to enable communication. A network bridge can be configured for failover or load sharing.
Link aggregation device	A link aggregation (also known as Etherchannel) device is a network port-aggregation technology that allows several Ethernet adapters to be aggregated.

For more information about PowerVM Simplification Enhancements, see *IBM Power Systems Hardware Management Console Version 8 Release 8.1.0 Enhancements*, SG24-8232.

3.6 System Planning Tool

The IBM System Planning Tool (SPT) helps you design systems to be partitioned with logical partitions. You can also plan for and design non-partitioned systems by using the SPT. The resulting output of your design is called a *system plan*, which is stored in a `.sysplan` file. This file can contain plans for a single system or multiple systems. The `.sysplan` file can be used for the following reasons:

- To create reports
- As input to the IBM configuration tool (e-Config)
- To create and deploy partitions on your system (or systems) automatically

System plans that are generated by the SPT can be deployed on the system by the HMC.

Automatically deploy: Ask your IBM SSR or IBM Business Partner to use the Customer Specified Placement manufacturing option if you want to automatically deploy your partitioning environment on a new machine. SPT looks for the resource's allocation that is the same as that specified in your `.sysplan` file.

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

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

Integration between the System Planning Tool and both the Workload Estimator and IBM Performance Management allows you to create a system that is based on performance and capacity data from an existing system or that is based on new workloads that you specify.

You can use the SPT before you order a system to determine what you must order to support your workload. You can also use the SPT to determine how you can partition a system that you have.

Using the SPT is an effective way of documenting and backing up key system settings and partition definitions. With it, the user can create records of systems and export them to their personal workstation or backup system of choice. These same backups can then be imported back on to the same system when needed. This step can be useful when cloning systems, enabling the user to import the system plan to any system multiple times.

The SPT and its supporting documentation can be found at the IBM System Planning Tool website:

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

3.7 IBM Power Virtualization Center

IBM Power Virtualization Center (IBM PowerVC) is designed to simplify the management of virtual resources in your Power Systems environment.

Note: PowerVC uses the term *virtual machine* (VM) instead of LPAR.

After the product code is loaded, the IBM PowerVC no-menus interface guides you through three simple configuration steps to register physical hosts, storage providers, and network resources, and start capturing and intelligently deploying your VMs, among the other tasks shown in the following list:

- Create virtual machines and then resize and attach volumes to them.
- Import existing virtual machines and volumes so they can be managed by IBM PowerVC.
- Monitor the usage of the resources that are in your environment.
- Migrate virtual machines while they are running (hot migration).
- Deploy images quickly to create virtual machines that meet the demands of your ever-changing business needs.

IBM PowerVC is built on OpenStack. *OpenStack* is an open source software that controls large pools of server, storage, and networking resources throughout a data center. PowerVC can manage AIX, IBM i, and Linux VMs running under PowerVM virtualization and Linux VMs running under PowerKVM virtualization.

> **Note:** At the time of writing, IBM i is not supported on the Power E850.

IBM PowerVC is available as IBM Power Virtualization Center Standard Edition.

In April 2015 PowerVC V1.2.3 was announced. This new release supports all Power Systems servers that are built on IBM POWER8 technology. PowerVC includes the following features and benefits:

- Virtual machine image capture, deployment, resizing, and management
- Policy-based VM placement to help improve usage and reduce complexity
- VM Mobility with placement policies to help reduce the burden on IT staff in a simplified GUI
- A management system that manages existing virtualization deployments
- Integrated management of storage, network, and compute, which simplifies administration

For more information about IBM PowerVC, see *IBM PowerVC Version 1.2.1 Introduction and Configuration*, SG24-8199.

3.8 IBM Power Virtualization Performance

IBM Power Virtualization Performance (IBM PowerVP) for Power Systems is a new product that offers a performance view into an IBM PowerVM virtualized environment running on the latest firmware of IBM Power Systems. It can show which virtual workloads are using specific physical resources on an IBM Power Systems server.

IBM PowerVP helps reduce time and complexity to find and display performance bottlenecks through a simple dashboard that shows the performance health of the system. It can help simplify both prevention and troubleshooting and thus reduce the cost of performance management.

It assists you in the following ways:

- Shows workloads in real time, which highlights possible problems or bottlenecks (overcommitted resources)

- Helps better use virtualized IBM Power System servers by showing the distribution of workloads
- Can replay saved historical data
- Helps with the resolution of performance-related issues
- Helps to proactively address future issues that can affect performance

IBM PowerVP is integrated with the POWER Hypervisor and collects performance data directly from PowerVM, which offers the most accurate performance information about virtual machines running on IBM Power Systems. This performance information is displayed on a real-time, continuous GUI dashboard and is also available for historical review.

Features of IBM PowerVP include these:

- Real-time, continuous graphical monitor (dashboard) that delivers an easy-to-read display showing the overall performance health of the Power Systems server.
- Customizable performance thresholds that enable you to customize the dashboard to match your monitoring requirements.
- Historical statistics that enable you to go back in time and replay performance data sequences to discover performance bottlenecks.
- System-level performance views that show all LPARs (VMs) and how they are using real system resources.
- Virtual machine drill down, which gives you more performance details for each VM, displaying detailed information about various resources, such as CPU, memory, and disk activity.
- Support for all virtual machine types, including AIX, IBM i, and Linux.

  > **Note:** At the time of writing, the Power E850 does not support IBM i.

- Background data collection, which enables performance data to be collected when the GUI is not active.

IBM PowerVP even allows an administrator to drill down and view specific adapter, bus, or CPU usage. An administrator can see the hardware adapters and how much workload is placed on them. IBM PowerVP provides both an overall and detailed view of IBM Power Systems server hardware, so it is easy to see how virtual machines are consuming resources. For more information about this topic, see the following website:

http://www.ibm.com/systems/power/software/performance

The latest Power VP 1.1.3 release has been enhanced with these other new features and support:

- A new capability to export PowerVP performance data to an external repository
- Integration with the VIOS performance advisor
- New thresholds and alert
- The ability to run the PowerVM user interface in a browser
- Support for monitoring RHEL 7.1, SLES 12, and Ubuntu 15.04 guests running under PowerVM Little Endian (LE) mode

3.9 VIOS 2.2.3.51 features

The Power E850 systems require the VIOS version 2.2.3.51 with APAR IV68443 and IV68444. This release provides the same functionality as the previous release and is updated to support the Power E850 and other new hardware.

VIOS 2.2.3.51 features include these:

- Simplified Shared Ethernet Adapter Failover configuration setup.
- Shared Storage Pools enhancements.
- With VIOS version 2.2.3.51, or later, the maximum number of virtual I/O slots that are supported on AIX and Linux partitions is increased up to 32.
- VIOS support:
  - Supports IBM Power E850
  - Supports 8 GB quad port PCIe Gen2 Fibre Channels adapter
- There is support for Live Partition Mobility performance enhancements to better use the 10 Gb Ethernet Adapters with the mobility process, and PowerVM server evacuation function.
- Shared Ethernet Adapter by default uses the largesend attribute.

The VIOS update packages can be downloaded from the IBM Fix Central website:

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

3.10 Dynamic Partition Remote Restart

Dynamic Partition Remote Restart is only available on systems managed by HMC.

Partition Remote Restart is a function designed to enhance availability of a partition on another server when its original host server fails. This is a high availability function of PowerVM Enterprise Edition.

Starting from IBM Power Systems HMC V8.8.1.0, the requirement of enabling Remote Restart of an LPAR only at creation time has been removed. Dynamic Partition Remote Restart allows for the dynamic toggle of Remote Restart capability when an LPAR is deactivated.

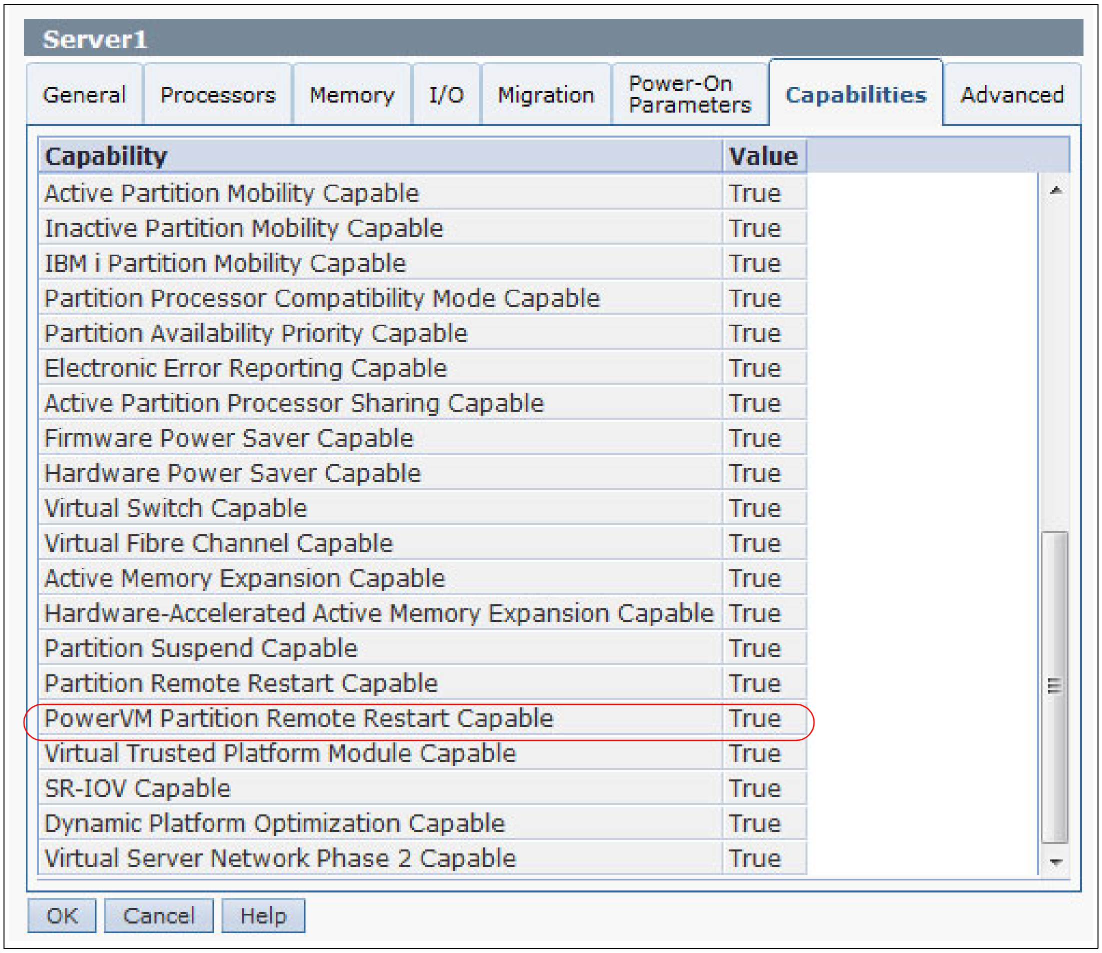
To verify that your managed system can support this capability, enter the following command at the HMC console as shown in Example 3-1.

Example 3-1 PowerVM remote restart capable

```
hscroot@slcb27a:~>lssyscfg -r sys -m Server1 -F capabilities
"active_lpar_mobility_capable,inactive_lpar_mobility_capable,os400_lpar_mobility_c
apable,active_lpar_share_idle_procs_capable,active_mem_dedup_capable,active_mem_ex
pansion_capable,hardware_active_mem_expansion_capable,active_mem_mirroring_hypervi
sor_capable,active_mem_sharing_capable,autorecovery_power_on_capable,bsr_capable,c
od_mem_capable,cod_proc_capable,custom_mac_addr_capable,dynamic_platform_optimizat
ion_capable,dynamic_platform_optimization_lpar_score_capable,electronic_err_report
ing_capable,firmware_power_saver_capable,hardware_power_saver_capable,hardware_dis
covery_capable,hardware_encryption_capable,hca_capable,huge_page_mem_capable,lpar_
affinity_group_capable,lpar_avail_priority_capable,lpar_proc_compat_mode_capable,l
par_remote_restart_capable,powervm_lpar_remote_restart_capable,lpar_suspend_capabl
e,os400_lpar_suspend_capable,micro_lpar_capable,os400_capable,5250_application_cap
able,os400_net_install_capable,os400_restricted_io_mode_capable,redundant_err_path
_reporting_capable,shared_eth_auto_control_channel_capable,shared_eth_failover_cap
able,sp_failover_capable,sriov_capable,vet_activation_capable,virtual_eth_disable_
capable,virtual_eth_dlpar_capable,virtual_eth_qos_capable,virtual_fc_capable,virtu
al_io_server_capable,virtual_switch_capable,vlan_stat_capable,vtpm_capable,vsi_on_
veth_capable,vsn_phase2_capable"
```

In Example 3-1, the highlighted text is indicating that the managed system is capable of remotely restarting a partition.

From the HMC, select the Managed System **Properties** → **Capabilities** tab to display all of the managed system capabilities as shown in Figure 3-17.



Capability	Value
Active Partition Mobility Capable	True
Inactive Partition Mobility Capable	True
IBM i Partition Mobility Capable	True
Partition Processor Compatibility Mode Capable	True
Partition Availability Priority Capable	True
Electronic Error Reporting Capable	True
Active Partition Processor Sharing Capable	True
Firmware Power Saver Capable	True
Hardware Power Saver Capable	True
Virtual Switch Capable	True
Virtual Fibre Channel Capable	True
Active Memory Expansion Capable	True
Hardware-Accelerated Active Memory Expansion Capable	True
Partition Suspend Capable	True
Partition Remote Restart Capable	True
PowerVM Partition Remote Restart Capable	True
Virtual Trusted Platform Module Capable	True
SR-IOV Capable	True
Dynamic Platform Optimization Capable	True
Virtual Server Network Phase 2 Capable	True

Figure 3-17 PowerVM Partition Remote Restart Capable

The capability is only displayed if the managed system supports it.

To activate a partition on a supported system to support Dynamic Partition Remote Restart, enter the following command:

```
chsyscfg -r lpar -m <ManagedSystemName> -i
"name=<PartitionName>,remote_restart_capable=1"
```

To use the Remote Restart feature, the following conditions need to be met:

- The managed system should support *toggle partition remote capability.*
- The partition should be in the inactive state.
- The partition type should be AIX, IBM i, or Linux.

> **Note:** At the time of writing, the Power E850 does not support IBM i.

- The reserved storage device pool exists.
- The partition should not own any of the below resources or settings:
  - BSR
  - Time Reference Partition
  - Service Partition
  - OptiConnect
  - HSL
  - Physical I/O
  - HEA
  - Error Reporting Partition
  - Is part of EWLM
  - Huge Page Allocation
  - Owns Virtual Serial Adapters
  - Belongs to I/O Fail Over Pool
  - SR-IOV logical port

For more information, including the use of Partition Remote Restart, see the following website:

http://www.ibm.com/support/knowledgecenter/POWER8/p8hat/p8hat_enadisremres.htm

4

Reliability, availability, and serviceability

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

The elements of RAS can be described as follows:

Reliability Indicates how infrequently a defect or fault in a server occurs

Availability Indicates how infrequently the functioning of a system or application is impacted by a fault or defect

Serviceability Indicates how well faults and their effects are communicated to system managers and how efficiently and nondisruptively the faults are repaired

4.1 Introduction

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

- Scale-out systems: For environments consisting of multiple systems working in concert. In such environments, application availability is enhanced by the superior availability characteristics of each system.
- Enterprise systems: For environments requiring systems with increased availability. In such environments, mission-critical applications can take full advantage of the scale-up characteristics, increased performance, flexibility to upgrade, and enterprise availability characteristics.

One key differentiator of the IBM POWER8 processor-based servers running PowerVM is that they leverage all the advanced RAS characteristics of the POWER8 processor design through the whole portfolio, offering reliability and availability features that are often not seen in other servers. Some of these features are improvements for POWER8 or features that were found previously only in higher-end Power Systems, which are now available across the entire range.

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 faults or soft faults. They use advanced technology, including stacked latches and Silicon-On-Insulator (SOI) technology, to reduce susceptibility to soft errors, and advanced design features within the processor for correction or try again after soft error events. In addition, the design incorporates spare capacity that is integrated into many elements to tolerate certain faults without requiring an outage or parts replacement. Advanced availability techniques are used to mitigate the impact of other faults that are not directly correctable in the hardware.

Features within the processor and throughout systems are incorporated to support design verification. During the design and development process, subsystems go through rigorous verification and integration testing processes by using these features. During system manufacturing, systems go through a thorough testing process to help ensure high product quality levels, again taking advantage of the designed ED/FI capabilities.

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

The Power E850 server has processor and memory upgrade capabilities characteristic of enterprise systems, and is designed to support higher levels of RAS than scale-out systems.

4.1.1 RAS enhancements of POWER8 processor-based servers

Several features were included in the whole portfolio of the POWER8 processor-based servers. Some of these features are improvements for POWER8 or features that were found previously only in higher-end Power Systems, leveraging a higher RAS even for scale-out equipment.

Here is a brief summary of these features:

- Processor enhancements

  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 substituted for a failing one dynamically during system operation.

A POWER8 processor module has improved performance compared to POWER7+, including support of a maximum of 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 additional communication fabrics and components.

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 can be used to attach external I/O drawers.

  For greater I/O capacity, the POWER8 processor-based Power E850 server also supports a PCIe switch to provide additional integrated I/O capacity.

- Memory subsystem

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

  After all self-healing and other RAS-related features are implemented, the hypervisor may still detect that a DIMM has a substantial fault that when combined with a future fault could cause an outage. In such a case, the hypervisor attempts to migrate data from the failing memory to other available memory in the system, if any is available. This feature is intended to further reduce the chances of an unplanned outage, and can take advantage of any deallocated memory including memory reserved for Capacity on Demand capabilities.

- Power distribution and temperature monitoring

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

4.1.2 RAS enhancements for enterprise servers

Following are RAS enhancements for enterprise servers:

- Memory Subsystem

  The Power E850 server has the option of mirroring the memory used by the hypervisor. This reduces the risk of system outage linked to memory faults, as the hypervisor memory is stored in two distinct memory CDIMMs. The Active Memory Mirroring feature is only available on enterprise systems.

- Power Distribution and Temperature Monitoring

  All systems make use of voltage converters that transform the voltage level provided by the power supply to the voltage level needed for the various components within the system. The Power E850 server has redundant or spare voltage converters for each voltage level provided to any given processor or memory CDIMM.

Converters 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 where when a converter fails, the system continues operation with another converter without generating a service event or the need to take any sort of outage for repair.

As with the Power E870 and Power E880 servers, the Power E850 uses triple redundant ambient temperature sensors.

4.2 Reliability

The reliability of systems starts with components, devices, and subsystems that are designed to be highly reliable. On IBM POWER processor-based systems, this basic principle is expanded upon with a clear design for reliability architecture and methodology. A concentrated, systematic, and architecture-based approach is designed to improve overall system reliability with each successive generation of system offerings. Reliability can be improved in primarily three ways:

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

In the POWER8 systems, elements of all three are used to improve system reliability.

During the design and development process, subsystems go through rigorous verification and integration testing processes. During system manufacturing, systems go through a thorough testing process to help ensure the highest level of product quality.

4.2.1 Designed for reliability

Systems that are designed with fewer components and interconnects have fewer opportunities to fail. Simple design choices, such as integrating processor cores on a single POWER chip, can reduce the opportunity for system failures. The POWER8 chip has more cores per processor module, and the I/O Hub Controller function is integrated in the processor module, which generates a PCIe BUS directly from the processor module. Parts selection also plays a critical role in overall system reliability.

IBM uses stringent design criteria to select server grade components that are extensively tested and qualified to meet and exceed a minimum design life of seven years. By selecting higher reliability grade components, the frequency of all failures is lowered, and wear-out is not expected within the operating system life. Component failure rates can be further improved by burning in select components or running the system before shipping it to the client. This period of high stress removes the weaker components with higher failure rates, that is, it cuts off the front end of the traditional failure rate bathtub curve (see Figure 4-1).

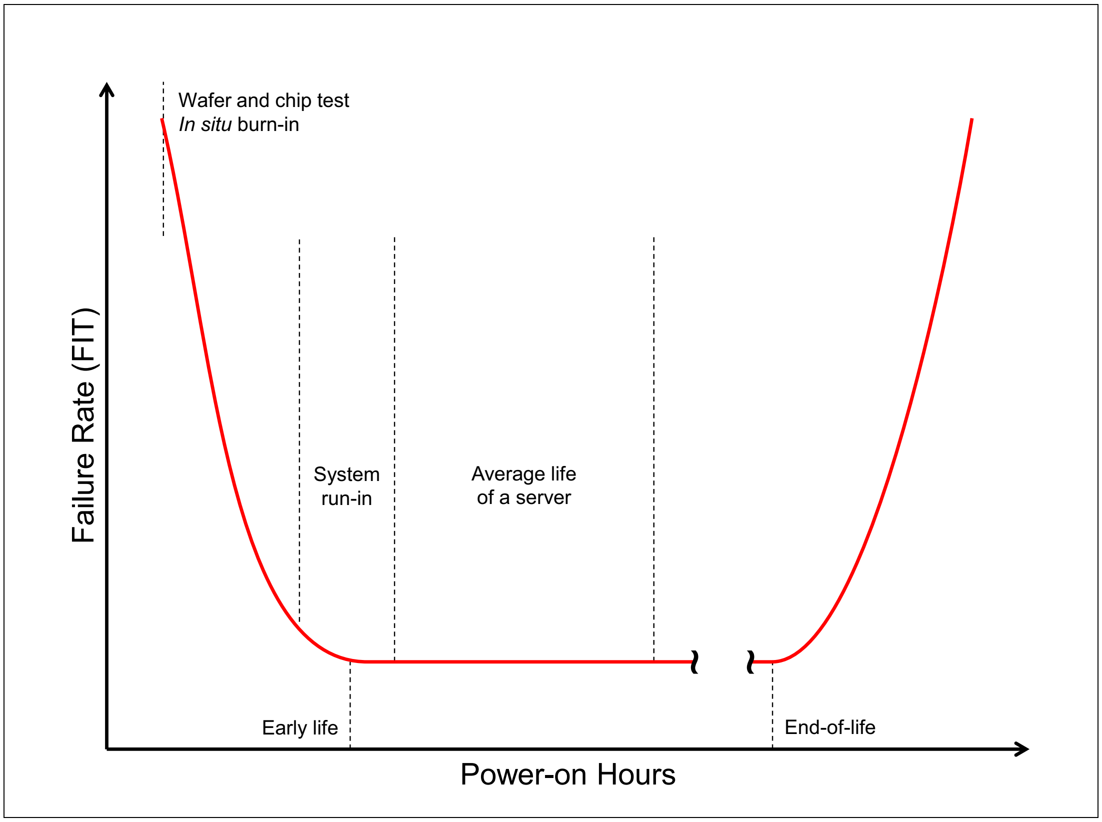


Figure 4-1 Failure rate bathtub curve

4.2.2 Placement of components

Packaging is designed to deliver both high performance and high reliability. For example, the reliability of electronic components is directly related to their thermal environment. Large decreases in component reliability are directly correlated to relatively small increases in temperature. All POWER processor-based systems are packaged to ensure adequate cooling. Critical system components, such as the POWER8 processor chips, are positioned on the system board so that they receive clear air flow during operation. POWER8 systems use a premium fan with an extended life to further reduce overall system failure rate and provide adequate cooling for the critical system components.

The Power E850 has two cooling channels. The front fans provide cooling for the upper part of the chassis, covering the memory cards, processors, and PCIe cards. These fans in this assembly provide redundancy, and support concurrent maintenance. The lower system fans, which are in the internal fan assembly, provide air movement for the lower part of the chassis, including the disk backplane and RAID controllers. The fans in the internal fan assembly provide redundancy, and also contain multiple integrated spares.

4.3 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 them. These design capabilities extend availability beyond what can be obtained through the underlying reliability of the hardware.

This design for availability begins with implementing an architecture for ED/FI.

First-Failure Data Capture (FFDC) is the capability of IBM hardware and microcode to continuously monitor hardware functions. Within the processor and memory subsystem, detailed monitoring is done by circuits within the hardware components themselves. Fault information is gathered into fault isolation registers (FIRs) and reported to the appropriate components for handling.

Processor and memory errors that are recoverable in nature are typically reported to the dedicated service processor built into each system. The dedicated service processor then works with the hardware to determine the course of action to be taken for each fault.

4.3.1 Correctable error introduction

Intermittent or soft errors are typically tolerated within the hardware design by using error correction code or advanced techniques to try operations again after a fault.

Tolerating a correctable solid fault runs the risk that the fault aligns with a soft error and causes an uncorrectable error situation. There is also the risk that a correctable error is predictive of a fault that continues to worsen over time, resulting in an uncorrectable error condition.

You can predictively deallocate a component to prevent correctable errors from aligning with soft errors or other hardware faults and causing uncorrectable errors to avoid such situations. However, unconfiguring components, such as processor cores or entire caches in memory, can reduce the performance or capacity of a system. This in turn typically requires that the failing hardware is replaced in the system. The resulting service action can also temporarily impact system availability.

To avoid such situations in solid faults in POWER8, processors or memory might be candidates for correction by using the “self-healing” features built into the hardware, such as taking advantage of a spare DRAM module within a memory DIMM, a spare data lane on a processor or memory bus, or spare capacity within a cache module.

When such self-healing is successful, the need to replace any hardware for a solid correctable fault is avoided. The ability to predictively unconfigure a processor core is still available for faults that cannot be repaired by self-healing techniques or because the sparing or self-healing capacity is exhausted.

4.3.2 Uncorrectable error introduction

An uncorrectable error can be defined as a fault that can cause incorrect instruction execution within logic functions, or an uncorrectable error in data that is stored in caches, registers, or other data structures. In less sophisticated designs, a detected uncorrectable error nearly always results in the termination of an entire system. More advanced system designs in some cases might be able to terminate just the application by using the hardware that failed. Such designs might require that uncorrectable errors are detected by the hardware and reported to software layers, and the software layers must then be responsible for determining how to minimize the impact of faults.

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

Depending on the fault, such recovery may use the virtualization capabilities of PowerVM in such a way that the operating system or any applications that are running in the system are not impacted or must participate in the recovery.

4.3.3 Processor Core/Cache correctable error handling

Layer 2 (L2) and Layer 3 (L3) caches and directories can correct single bit errors and detect double bit errors (SEC/DED ECC). Soft errors that are detected in the level 1 caches are also correctable by a try again operation that is handled by the hardware. Internal and external processor “fabric” busses have SEC/DED ECC protection as well.

SEC/DED capabilities are also included in other data arrays that are not directly visible to customers.

Beyond soft error correction, the intent of the POWER8 design is to manage a solid correctable error in an L2 or L3 cache by using techniques to delete a cache line with a persistent issue, or to repair a column of an L3 cache dynamically by using spare capability.

Information about column and row repair operations is stored persistently for processors, so that more permanent repairs can be made during processor reinitialization (during system reboot, or individual Core Power on Reset using the Power On Reset Engine.)

4.3.4 Processor Instruction Retry and other try again techniques

Within the processor core, soft error events might occur that interfere with the various computation units. When such an event can be detected before a failing instruction is completed, the processor hardware might be able to try the operation again by using the advanced RAS feature that is known as *Processor Instruction Retry*.

Processor Instruction Retry allows the system to recover from soft faults that otherwise result in an outage of applications or the entire server.

Try again techniques are used in other parts of the system as well. Faults that are detected on the memory bus that connects processor memory controllers to DIMMs can be tried again. In POWER8 systems, the memory controller is designed with a replay buffer that allows memory transactions to be tried again after certain faults internal to the memory controller faults are detected. This complements the try again abilities of the memory buffer module.

4.3.5 Alternative processor recovery and Partition Availability Priority

If Processor Instruction Retry for a fault within a core occurs multiple times without success, the fault is considered to be a solid failure. In some instances, PowerVM can work with the processor hardware to migrate a workload running on the failing processor to a spare or alternative processor. This migration is accomplished by migrating the pertinent processor core state from one core to another with the new core taking over at the instruction that failed on the faulty core. Successful migration keeps the application running during the migration without needing to terminate the failing application.

Successful migration requires that there is sufficient spare capacity that is available to reduce the overall processing capacity within the system by one processor core. Typically, in highly virtualized environments, the requirements of partitions can be reduced to accomplish this task without any further impact to running applications.

In systems without sufficient reserve capacity, it might be necessary to terminate at least one partition to free resources for the migration. In advance, PowerVM users can identify which partitions have the highest priority and, which do not. When you use this Partition Priority feature of PowerVM, if a partition must be terminated for alternative processor recovery to complete, the system can terminate lower priority partitions to keep the higher priority partitions up and running, even when an unrecoverable error occurred on a core running the highest priority workload.

Partition Availability Priority is assigned to partitions by using a weight value or integer rating. The lowest priority partition is rated at zero and the highest priority partition is rated at 255. The default value is set to 127 for standard partitions and 192 for Virtual I/O Server (VIOS) partitions. Priorities can be modified through the Hardware Management Console (HMC).

4.3.6 Core Contained Checkstops and other PowerVM error recovery

PowerVM can handle certain other hardware faults without terminating applications, such as an error in certain data structures (faults in translation tables or lookaside buffers).

Other core hardware faults that alternative processor recovery or Processor Instruction Retry cannot contain might be handled in PowerVM by a technique called *Core Contained Checkstops*. This technique allows PowerVM to be signaled when such faults occur and terminate code in use by the failing processor core (typically just a partition, although potentially PowerVM itself if the failing instruction were in a critical area of PowerVM code).

Processor designs without Processor Instruction Retry typically must resort to such techniques for all faults that can be contained to an instruction in a processor core.

4.3.7 Cache uncorrectable error handling

If a fault within a cache occurs that cannot be corrected with SEC/DED ECC, the faulty cache element is unconfigured from the system. Typically, this is done by purging and deleting a single cache line. Such purge and delete operations are contained within the hardware itself, and prevent a faulty cache line from being reused and causing multiple errors.

During the cache purge operation, the data that is stored in the cache line is corrected where possible. If correction is not possible, the associated cache line is marked with a special ECC code that indicates that the cache line itself has bad data.

Nothing within the system terminates just because such an event is encountered. Rather, the hardware monitors the usage of pages with marks. If such data is never used, hardware replacement is requested, but nothing terminates as a result of the operation. Software layers are not required to handle such faults.

Only when data is loaded to be processed by a processor core, or sent out to an I/O adapter, is any further action needed. In such cases, if data is used as owned by a partition, the partition operating system might be responsible for terminating itself or just the program using the marked page. If data is owned by the hypervisor, the hypervisor might choose to terminate, resulting in a system-wide outage.

However, the exposure to such events is minimized because cache-lines can be deleted, which eliminates repetition of an uncorrectable fault that is in a particular cache-line.

4.3.8 Other processor chip functions

Within a processor chip, there are other functions besides just processor cores.

POWER8 processors have built-in accelerators that can be used as application resources to handle such functions as random number generation. POWER8 also introduces a controller for attaching cache-coherent adapters that are external to the processor module. The POWER8 design contains a function to “freeze” the function that is associated with some of these elements, without taking a system-wide checkstop. Depending on the code using these features, a “freeze” event might be handled without an application or partition outage.

As indicated elsewhere, single bit errors, even solid faults, within internal or external processor “fabric busses”, are corrected by the error correction code that is used. POWER8 processor-to-processor module fabric busses also use a spare data-lane so that a single failure can be repaired without calling for the replacement of hardware.

4.3.9 Other fault error handling

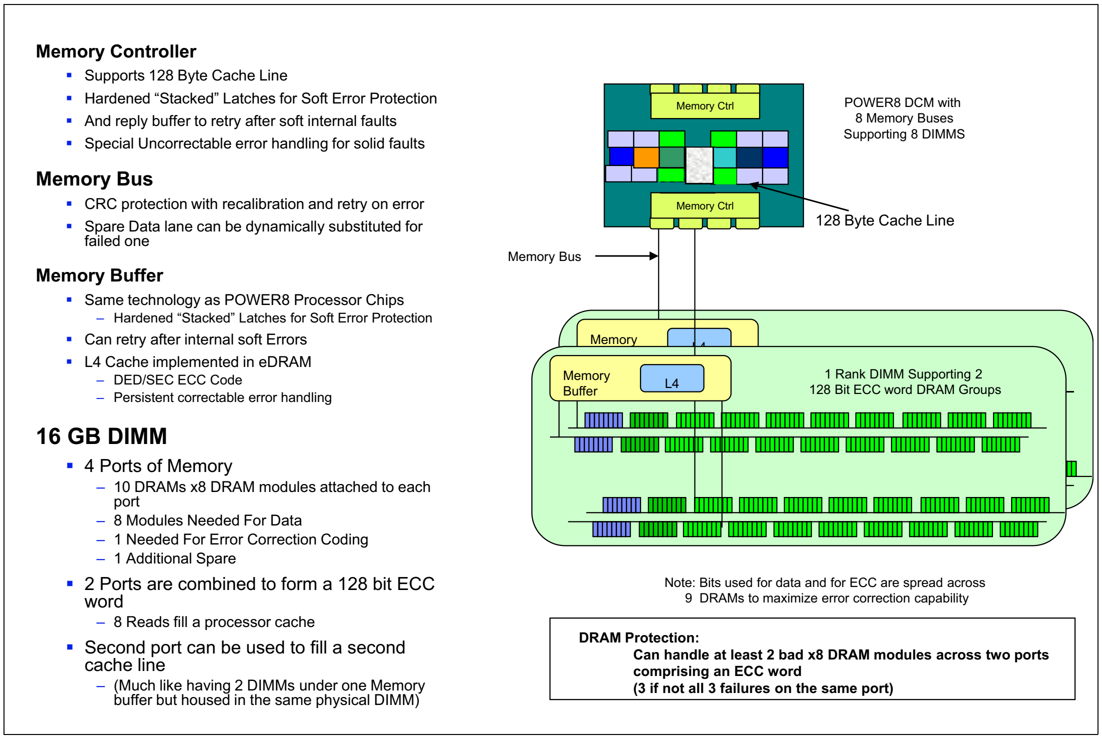
Not all processor module faults can be corrected by these techniques. Therefore, a provision is still made for some faults that cause a system-wide outage. In such a “platform” checkstop event, the ED/FI capabilities that are built in to the hardware and dedicated service processor work to isolate the root cause of the checkstop and unconfigure the faulty element 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, which communicate to buffer modules on memory DIMMS through memory channels and access the DRAM memory modules on DIMMs, as shown in Figure 4-2.



Memory Controller

- Supports 128 Byte Cache Line
- Hardened “Stacked” Latches for Soft Error Protection
- And reply buffer to retry after soft internal faults
- Special Uncorrectable error handling for solid faults

Memory Bus

- CRC protection with recalibration and retry on error
- Spare Data lane can be dynamically substituted for failed one

Memory Buffer

- Same technology as POWER8 Processor Chips
  - Hardened “Stacked” Latches for Soft Error Protection
- Can retry after internal soft Errors
- L4 Cache implemented in eDRAM
  - DED/SEC ECC Code
  - Persistent correctable error handling

16 GB DIMM

- 4 Ports of Memory
  - 10 DRAMs x8 DRAM modules attached to each port
  - 8 Modules Needed For Data
  - 1 Needed For Error Correction Coding
  - 1 Additional Spare
- 2 Ports are combined to form a 128 bit ECC word
  - 8 Reads fill a processor cache
- Second port can be used to fill a second cache line
  - (Much like having 2 DIMMs under one Memory buffer but housed in the same physical DIMM)

Note: Bits used for data and for ECC are spread across 9 DRAMs to maximize error correction capability

DRAM Protection:
Can handle at least 2 bad x8 DRAM modules across two ports comprising an ECC word
(3 if not all 3 failures on the same port)

Figure 4-2 Memory protection features

The memory buffer chip is made by the same 22 nm technology that is used to make the POWER8 processor chip, and the memory buffer chip incorporates the same features in the technology to avoid soft errors. It implements a try again process for many internally detected faults. This function complements a replay buffer in the memory controller within the processor, which also handles internally detected soft errors.

The bus between a processor memory controller and a DIMM uses CRC error detection that is coupled with the ability to try soft errors again. The bus features dynamic recalibration capabilities plus a spare data lane that can be substituted for a failing bus lane through the recalibration process.

The buffer module implements an integrated L4 cache using eDRAM technology (with soft error hardening) and persistent error handling features.

The memory buffer on each DIMM has four ports for communicating with DRAM modules. The 16 GB DIMM, for example, has one rank that is composed of four ports of x8 DRAM modules, each port containing 10 DRAM modules.

For each such port, there are eight DRAM modules worth of data (64 bits) plus another DRAM module’s worth of error correction and other such data. There is also a spare DRAM module for each port that can be substituted for a failing port.

Two ports are combined into an ECC word and supply 128 bits of data. The ECC that is deployed can correct the result of an entire DRAM module that is faulty. This is also known as *Chipkill* correction. Then, it can correct at least an additional bit within the ECC word.

The additional spare DRAM modules are used so that when a DIMM experiences a Chipkill event within the DRAM modules under a port, the spare DRAM module can be substituted for a failing module, avoiding the need to replace the DIMM for a single Chipkill event.

Depending on how DRAM modules fail, it might be possible to tolerate up to four DRAM modules failing on a single DIMM without needing to replace the DIMM, and then still correct an additional DRAM module that is failing within the DIMM.

There are other DIMMs offered with these systems. A 32 GB DIMM has two ranks, where each rank is similar to the 16 GB DIMM with DRAM modules on four ports, and each port has 10 x8 DRAM modules.

In addition, there is a 64 GB DIMM that is offered through x4 DRAM modules that are organized in four ranks.

In addition to the protection that is provided by the ECC and sparing capabilities, the memory subsystem also implements scrubbing of memory to identify and correct single bit soft-errors. Hypervisors are informed of incidents of single-cell persistent (hard) faults for deallocation of associated pages. However, because of the ECC and sparing capabilities that are used, such memory page deallocation is not relied upon for repair of faulty hardware.

Should a more substantial fault persist after all the self-healing capabilities are utilized, the hypervisor also has the capability of dynamically moving logical memory blocks from faulty memory to unused memory blocks in other parts of the system. This feature can take advantage of memory otherwise reserved for Capacity on Demand capabilities.

Finally, should an uncorrectable error in data be encountered, the memory that is impacted is marked with a special uncorrectable error code and handled as described for cache uncorrectable errors.

4.3.11 I/O subsystem availability and Enhanced Error Handling

Usage of multi-path I/O and VIOS for I/O adapters and RAID for storage devices should be used to prevent application outages when I/O adapter faults occur.

To permit soft or intermittent faults to be recovered without failover to an alternative device or I/O path, Power Systems hardware supports *Enhanced Error Handling* (EEH) for I/O adapters and PCIe bus faults.

EEH allows EEH-aware device drivers to try again after certain non-fatal I/O events to avoid failover, especially in cases where a soft error is encountered. EEH also allows device drivers to terminate if there is an intermittent hard error or other unrecoverable errors, while protecting against reliance on data that cannot be corrected. This action typically is done by “freezing” access to the I/O subsystem with the fault. Freezing prevents data from flowing to and from an I/O adapter and causes the hardware/firmware to respond with a defined error signature whenever an attempt is made to access the device. If necessary, a special uncorrectable error code can be used to mark a section of data as bad when the freeze is first initiated.

In POWER8 processor-based systems, the external I/O hub and bridge adapters were eliminated in favor of a topology that integrates PCIe Host Bridges into the processor module itself. PCIe busses that are generated directly from a host bridge may drive individual I/O slots or a PCIe switch. The integrated PCIe controller supports try again (end-point error recovery) and freezing.

IBM device drivers under AIX are fully EEH-capable. For Linux under PowerVM, EEH support extends to many frequently used devices. There might be third-party PCI devices that do not provide native EEH support.

4.3.12 Remote Restart capability

The Power E850 server supports logical partitions that are capable of remote restarting. This option can be enabled for each partition individually. If a host system fails, the workload of the host system can be automatically restarted on an operational target system. To recover from unexpected failures and to improve the availability of the systems, workloads are restarted automatically.

The HMC can restart an AIX or Linux partition remotely if the partition supports an attribute that is called an *encapsulated state*. An encapsulated state partition is a partition in which the configuration information and the persistent data are stored external to the server on persistent storage.

This capability requires that an HMC manage the host system and the operational target system. The Remote Restart capability requires that the following requirements are met:

- The server supports the remote restart capability. This is the case for the Power E850.
- The partition must not have physical I/O adapters that are assigned to the partition.
- The partition must not be a full system partition, or a Virtual I/O Server.
- The partition must not be an alternative error-logging partition.
- The partition must not have a barrier-synchronization register (BSR).
- The partition must not have huge pages (applicable only if PowerVM Active Memory Sharing is enabled).
- The partition must not have its rootvg volume group on a logical volume or have any exported optical devices.

Remote Restart can be enabled for applicable partitions through the HMC interface.

4.4 Availability impacts of a solution architecture

Any given solution should not rely only on the hardware platform. Despite IBM Power Systems having superior RAS features to other comparable systems, it is advisable to design a redundant architecture surrounding the application in order to allow for easier maintenance tasks and greater flexibility.

When running in a redundant architecture, some tasks that would otherwise require that a given application be brought offline, can now be executed with the application running, allowing for even greater availability.

When determining a highly available architecture that fits your needs, the following topics are worth considering:

- Will I need to move my workload from an entire server during service or planned outages?
- If I use a clustering solution to move the workload, how will the failover time affect my service?
- If I use a server evacuation solution to move the workload, how long will it take to migrate all the partitions with my current server configuration?

4.4.1 Clustering

IBM Power Systems running under PowerVM, AIX, and Linux support a number of different clustering solutions. These solutions are designed to meet requirements not only for application availability in regard to server outages, but also data center disaster management, reliable data backups, and so forth. These offerings include distributed applications with IBM DB2® PureScale, HA solutions using clustering technology with IBM PowerHA SystemMirror®, and disaster management across geographies with PowerHA SystemMirror Enterprise Edition.

For more information, see the following references:

- *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`

4.4.2 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 terminates for any reason (hardware or software caused), all the partitions using that VIOS will terminate.

Dual VIOS partitions can only be created through an optional Hardware Management Console (HMC). A Power E850 server that is managed through the Integrated Virtualization Manager (IVM) is limited to a single VIOS partition.

Using redundant VIOS partitions would mitigate that risk. Maintaining redundancy of adapters within each VIOS, in addition to having redundant VIOS, will avoid most faults that keep a VIOS from running. Multiple paths to networks and SANs is therefore advised. Figure 4-3 shows a diagram of a partition accessing data from two distinct Virtual I/O Servers, each one with multiple network and SAN adapters to provide connectivity.

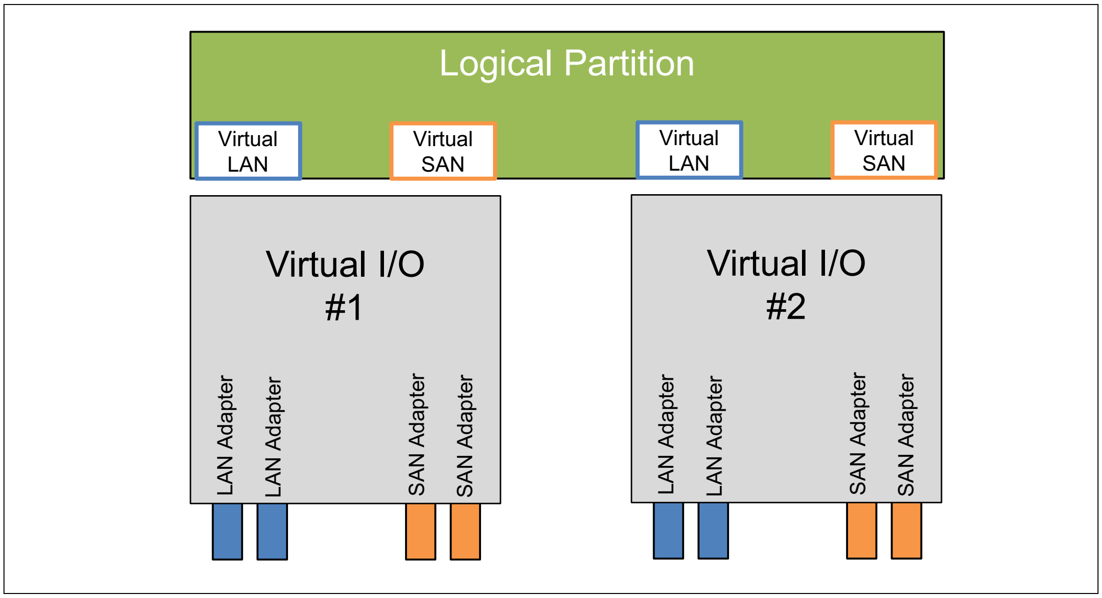


Figure 4-3 Partition utilizing dual redundant virtual I/O servers for connectivity

Since each VIOS can largely be considered as an AIX based partition, each VIOS also needs the ability to access a boot image, have paging space, and other functions that require a root volume group or rootvg. The rootvg can be accessed through a SAN, through storage that is locally attached to a server, or through internal hard disks or solid-state devices. For best availability, the rootvg for each VIOS should use mirrored or RAID protected drives with redundant access to the devices.

4.4.3 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility (LPM) allows you to move a running logical partition, including its operating system and running applications, from one system to another without any shutdown and without disrupting the operation of that logical partition. Inactive partition mobility allows you to move a powered-off logical partition from one system to another.

Live Partition Mobility provides systems management flexibility and improves system availability through the following functions:

- Avoid planned outages for hardware or firmware maintenance by moving logical partitions to another server and then performing the maintenance. Live Partition Mobility can help lead to zero downtime for maintenance because you can use it to work around scheduled maintenance activities.
- Avoid downtime for a server upgrade by moving logical partitions to another server and then performing the upgrade. This approach allows your users to continue their work without disruption.

- Avoid unplanned downtime. With preventive failure management, if a server indicates a potential failure, you can move its logical partitions to another server before the failure occurs. Partition mobility can help avoid unplanned downtime.
- Take advantage of server optimization:
  - Consolidation: You can consolidate workloads that run on several small, under utilized servers onto a single large server.
  - Optimized placement: 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.

Live Partition Mobility can be completed between two systems that are managed by the same HMC. It is also possible to migrate partitions between two systems that are managed by different HMCs, which must be connected via a network.

Management through an HMC is optional for Power E850 servers. For systems managed by the Integrated Virtualization Manager (IVM), partitions can only be migrated to other servers that are managed through an IVM, and these systems must be connected via a network.

Server Evacuation: This PowerVM function allows you to perform a Server Evacuation operation. Server Evacuation is used to move all migration-capable LPARs from one system to another if there are no active migrations in progress on the source or the target servers.

With the Server Evacuation feature, multiple migrations can occur based on the concurrency setting of the HMC. Migrations are performed as sets, with the next set of migrations starting when the previous set completes. Any upgrade or maintenance operations can be performed after all the partitions are migrated and the source system is powered off.

You can migrate all the migration-capable AIX, IBM i, and Linux partitions from the source server to the destination server by running the following command from the HMC command line:

```
migrlpar -o m -m source_server -t target_server --all
```

Hardware and operating system requirements for Live Partition Mobility

Live Partition Mobility is supported with PowerVM Enterprise Edition in compliance with all operating systems that are compatible with POWER8 technology.

Logical partitions can only be relocated using Live Partition Mobility if they are running in a fully virtualized environment, utilizing external storage, which is accessible to both the exiting host server and the destination server.

VIOS partitions cannot be migrated.

For more information about Live Partition Mobility and how to implement it, see *IBM PowerVM Live Partition Mobility (Obsolete - See Abstract for Information)*, SG24-7460.

4.5 Serviceability

The purpose of serviceability is to repair or upgrade the system while attempting to minimize or eliminate service cost (within budget objectives) and maintaining application availability and high customer satisfaction. Serviceability includes system installation, miscellaneous equipment specification (MES) system upgrades and 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:

- A design for customer setup (CSU), customer installed features (CIF), and customer replaceable units (CRU)
- ED/FI incorporating FFDC
- Converged service approach across multiple IBM server platforms
- Concurrent Firmware Maintenance (CFM)

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.

4.5.1 Detecting errors

The first and most crucial component of a solid serviceability strategy is the ability to accurately and effectively detect errors when they occur.

Although not all errors are a threat to system availability, those that go undetected can cause problems because the system has no opportunity to evaluate and act if necessary. POWER processor-based systems employ IBM z Systems server-inspired error detection mechanisms, extending from processor cores and memory to power supplies and storage devices.

4.5.2 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 and memory subsystem error-checkers, error-check signals are captured and stored in hardware FIRs. The associated logic circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, runtime error diagnostic tests can be deterministic so that for every check station, the unique error domain for that checker is defined and mapped to field replaceable units (FRUs) that can be repaired when necessary.

Integral to the Power Systems design is the concept of FFDC. FFDC is a technique that involves sufficient error checking stations and coordination of fault reporting 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 the vast majority of faults, a good FFDC design means that the root cause is isolated at the time of the failure without intervention by a service representative. For all faults, good FFDC design still makes failure information available to the service representative, and this information can be used to confirm the automatic diagnosis. More detailed information can be collected by a service representative for rare cases where the automatic diagnosis is not adequate for fault isolation.

4.5.3 Service processor

In POWER8 processor-based systems 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 the Power E850 server, redundant connections to the service processor provide added reliability.

In addition to FFDC functions, the service processor performs many serviceability functions:

- 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 an integrated operating system interface, the service processor notifies the operating system of potential environmental related problems so that the system administrator can take appropriate corrective actions before a critical failure threshold is reached. The service processor can also post a warning and initiate an orderly system shutdown in the following circumstances:

  - The operating temperature exceeds the critical level (for example, failure of air conditioning or air circulation around the system).
  - Internal component temperatures reach or exceed critical levels.
  - 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.

- POWER Hypervisor (system firmware) and HMC connection surveillance

  The service processor monitors the operation of the firmware during the boot process, and also monitors the hypervisor for termination. The hypervisor monitors the service processor and can perform a reset and reload if it detects the loss of the service processor. If the reset or reload operation does not correct the problem with the service processor, the hypervisor notifies the operating system, and then the operating system can take appropriate action, including calling for service. The service processor also monitors the connection to an HMC and can report loss of connectivity to the operating system partitions for system administrator notification.

- Uncorrectable error recovery

  The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced (AC power) failure.

  The auto-restart (reboot) option must be enabled from the ASMI menu or from the operator panel on the front of the server.

- Concurrent access to the service processor menu of the ASMI

  The Advanced System Management Interface (ASMI) provides management functionality for the server hardware through the service processor. Access to these menus allows nondisruptive changes to system default parameters, interrogation of service processor progress and error logs, and the ability to set and reset service indicators (Light Path). You can also access all service processor functions without having to power down the system to the standby state. The administrator or service representative can dynamically access the menu and functionality from any web browser-enabled console that is attached to the Ethernet service network, concurrently with normal system operation. Some options, such as changing the hypervisor type, do not take effect until the next boot.

- Management of the interfaces for connecting uninterruptible power source systems to the POWER processor-based systems and performing timed power-on (TPO) sequences.

4.5.4 Diagnosing

General diagnostic objectives are to detect and identify problems so that they can be resolved quickly. The IBM diagnostic strategy includes the following elements:

- Provide a common error code format that is equivalent to a system reference code, system reference number, checkpoint, or firmware error code.
- Provide fault detection and problem isolation procedures. Support a remote connection ability that is used by the IBM Remote Support Center or IBM Designated Service.
- Provide interactive intelligence within the diagnostic tests with detailed online failure information while connected to IBM back-end systems.

Using the extensive network of advanced and complementary error detection logic that is built directly into hardware, firmware, and operating systems, the Power E850 server can perform considerable self-diagnosis.

Because of the FFDC technology that is designed in to all Power Systems servers, re-creating diagnostic tests for failures or requiring user intervention is not necessary. Solid and intermittent errors are designed to be correctly detected and isolated at the time that the failure occurs. Runtime and boot time diagnostic tests fall into this category.

Boot time

When an IBM Power Systems server powers up, the service processor initializes the system hardware. Boot-time diagnostic testing uses a 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, which is based on the way that the system was powered off, or on the boot-time selection menu.

Host Boot IPL

In POWER8, the initialization process during IPL changed. The Flexible Service Processor (FSP) is no longer the only instance that initializes and runs the boot process. With POWER8, the FSP initializes the boot processes, but on the POWER8 processor itself, one part of the firmware is running and performing the Central Electronics Complex chip initialization. A new component that is called the PNOR chip stores the Host Boot firmware and the Self Boot Engine (SBE) is an internal part of the POWER8 chip itself and is used to boot the chip.

With this Host Boot initialization, new progress codes are available. An example of an FSP progress code is C1009003. During the Host Boot IPL, progress codes, such as CC009344, appear.

If there is a failure during the Host Boot process, a new Host Boot System Dump is collected and stored. This type of memory dump includes Host Boot memory and is offloaded to the HMC when it is available.

Run time

All Power Systems servers can monitor critical system components during run time, and they can take corrective actions when recoverable faults occur. The hardware error-check architecture can report non-critical errors in the system in an *out-of-band* communications path to the service processor without affecting system performance.

A significant part of the runtime diagnostic capabilities originate with the service processor. Extensive diagnostic and fault analysis routines were developed and improved over many generations of POWER processor-based servers, and enable quick and accurate predefined responses to both actual and potential system problems.

The service processor correlates and processes runtime error information by using logic that is derived from IBM engineering expertise to count recoverable errors (called *thresholding*) and predict when corrective actions must be automatically initiated by the system. These actions can include the following items:

- Requests for a part to be replaced
- Dynamic invocation of built-in redundancy for automatic replacement of a failing part
- Dynamic deallocation of failing components so that system availability is maintained

Device drivers

In certain cases, diagnostic tests are best performed by operating system-specific drivers, most notably adapters or I/O devices that are owned directly by a logical partition. In these cases, the operating system device driver often works with I/O device microcode to isolate and recover from problems. Potential problems are reported to an operating system device driver, which logs the error. In non-HMC managed servers, the OS can start the Call Home application to report the service event to IBM. For HMC managed servers, the event is reported to the HMC, which can initiate the Call Home request to IBM. I/O devices can also include specific exercisers that can be started by the diagnostic facilities for problem re-creation (if required by service procedures).

4.5.5 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 can automatically send an alert or call for service if there is a critical system failure. A hardware fault also illuminates the amber system fault LED, which is on the system unit, to alert the user of an internal hardware problem.

On POWER8 processor-based servers, hardware and software failures are recorded in the system log. When a management console is attached, an ELA routine analyzes the error, forwards the event to the Service Focal Point™ (SFP) application running on the management console, and can notify the system administrator that it isolated a likely cause of the system problem. The service processor event log also records unrecoverable checkstop conditions, forwards them to the SFP application, and notifies the system administrator.

After the information is logged in the SFP application, if the system is correctly configured, a Call Home service request is initiated and the pertinent failure data with service parts information and part locations is sent to the IBM service organization. This information also contains the client contact information as defined in the IBM Electronic Service Agent (ESA) guided setup wizard. With the new HMC V8R8.1.0 a Serviceable Event Manager is available to block problems from being automatically transferred to IBM. For more information about this topic, see "Service Event Manager" on page 181 for more details.

Error logging and analysis

When the root cause of an error is identified by a fault isolation component, an error log entry is created with basic data, such as the following examples:

- An error code that uniquely describes the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data such as engineering and manufacturing levels
- Return codes
- Resource identifiers
- FFDC data

Data that contains information about the effect that the repair has on the system is also included. Error log routines in the operating system and FSP can then use this information and decide whether the fault is a Call Home candidate. If the fault requires support intervention, a call is placed with service and support, and a notification is sent to the contact that is defined in the ESA-guided setup wizard.

Remote support

The Remote Management and Control (RMC) subsystem is delivered as part of the base operating system, including the operating system that runs on the HMC. RMC provides a secure transport mechanism across the LAN interface between the operating system and the optional HMC and is used by the operating system diagnostic application for transmitting error information. It performs several other functions, but they are not used for the service infrastructure.

Service Focal Point application for partitioned systems

A critical requirement in a logically partitioned environment is to ensure that errors are not lost before being reported for service, and that an error should be reported only once, regardless of how many logical partitions experience the potential effect of the error. The SFP application on the management console or in the Integrated Virtualization Manager (IVM) is responsible for aggregating duplicate error reports, and ensures that all errors are recorded for review and management. The SFP application provides other service-related functions, such as controlling service indicators, setting up Call Home, and providing guided maintenance.

When a local or globally reported service request is made to the operating system, the operating system diagnostic subsystem uses the RMC subsystem to relay error information to the optional HMC. For global events (platform unrecoverable errors, for example), the service processor also forwards error notification of these events to the HMC, providing a redundant error-reporting path in case there are errors in the RMC subsystem network.

The first occurrence of each failure type is recorded in the Manage Serviceable Events task on the management console. This task then filters and maintains a history of duplicate reports from other logical partitions or from the service processor. It then looks at all active service event requests within a predefined timespan, analyzes the failure to ascertain the root cause and, if enabled, initiates a Call Home for service. This methodology ensures that all platform errors are reported through at least one functional path, ultimately resulting in a single notification for a single problem. Similar service functionality is provided through the SFP application on the IVM for providing service functions and interfaces on non-HMC partitioned servers.

Extended error data

Extended error data (EED) is additional data that is collected either automatically at the time of a failure or manually later. The data that is collected depends on the invocation method, but includes information such as firmware levels, operating system levels, additional fault isolation register values, recoverable error threshold register values, system status, and any other pertinent data.

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

System dump handling

In certain circumstances, an error might require a memory dump to be automatically or manually created. In this event, the memory dump can be offloaded to the optional HMC. Specific management console information is included as part of the information that optionally can be sent to IBM Support for analysis. If additional information that relates to the memory dump is required, or if viewing the memory dump remotely becomes necessary, the management console memory dump record notifies the IBM Support center of which management console the memory dump is on. If no management console is present, the memory dump might be either on the FSP or in the operating system, depending on the type of memory dump that was initiated and whether the operating system is operational.

4.5.6 Notifying

After a Power E850 server detects, diagnoses, and reports an error to an appropriate aggregation point, it then takes steps to notify the administrator and if necessary, the IBM Support organization. Depending on the assessed severity of the error and support agreement, this notification might range from a simple message to having field service personnel automatically dispatched to the client site with the correct replacement part.

Client Notify

When an event is important enough to report, but does not indicate the need for a repair action or the need to call home to IBM Support, it is classified as *Client Notify*. Clients are notified because these events might be of interest to an administrator. The event might be a symptom of an expected systemic change, such as a network reconfiguration or failover testing of redundant power or cooling systems. These events include the following examples:

- Network events, such as the loss of contact over a local area network (LAN)
- Environmental events, such as ambient temperature warnings
- Events that need further examination by the client (although these events do not necessarily require a part replacement or repair action)

Client Notify events are serviceable events because they indicate that something happened that requires client awareness, and the administrator might want to take further action. These events can be reported to IBM at the discretion of the administrator.

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 procedures within the service organization so that the appropriate service action can begin. Call Home can be done through the HMC if available, or directly from machines that are not managed by an HMC.

Although configuring a Call Home function is optional, clients are encouraged to implement this feature to obtain service enhancements, such as reduced time to problem determination and faster, more accurate transmission of error information. In general, using the Call Home feature can result in increased system availability. The ESA application can be configured for automated Call Home. For more information, see 4.6.4, “Electronic Services and Electronic Service Agent” on page 179.

Vital product data and inventory management

Power Systems store vital product data (VPD) internally, which keeps a record of how much memory is installed, how many processors are installed, the manufacturing level of the parts, and similar system data. These records provide valuable information that can be used by remote support and service representatives, enabling the service representatives to assist in keeping the firmware and software current on the server.

IBM Service and Support Problem Management database

At the IBM Support center, historical problem data is entered into the IBM Service and Support Problem Management database. All of the information that is related to the error, along with any service actions that are taken by the service representative, are recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

4.5.7 Locating and servicing

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Packaging for service

The following service enhancements are included in the physical packaging of the systems to facilitate service:

- Color coding (touch points)

  Terracotta on the part or a release lever indicates the system might not be required to be powered off to perform service. This depends on system configuration and preparatory steps might be required before the service action is taken on the system. For any concurrent maintenance procedures, care should be taken to follow the steps that are indicated by the HMC or maintenance menus in the correct order.

  Blue on the part or on a release lever, latch, or thumb-screw indicates that the procedure may require the unit or system to be shut down before servicing or replacing the part. Check your service procedure before attempting repair, and ensure that you fully understand the process required before starting work.

- 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 becoming 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 of the Power E850 assists in repairing parts that can be replaced by the user. In the Light Path LED implementation, when a fault condition is detected on the Power E850 server, an amber fault LED is illuminated (turned on solid), which also illuminates the system fault LED. The Light Path system pinpoints the exact part by illuminating the amber fault LED that is associated with the part that needs to be replaced.

The service representative can clearly identify components for replacement by using specific component level identity LEDs. The system can also guide the service representative directly to the component by signaling (flashing) the component identity LED, and illuminating the blue enclosure identity LED to identify the server in a busy data center.

After the repair, the LEDs shut off automatically once the part is replaced. The Light Path LEDs are only visible while the system is powered on or has standby power connected.

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

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

The IBM Knowledge Center contains sections specific to each server model, and include detailed service procedures for a number of potential repair situations. The service procedure repository for a particular server model can be found in the “Troubleshooting, service and support” section.

The IBM Knowledge Center can be found online at the following link:

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

Service labels

Service representatives 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.

Some of these service labels and their purposes are described in the following list:

- *Location diagrams* are strategically positioned on the system hardware and relate information about the placement of hardware components. Location diagrams can include location codes, drawings of physical locations, concurrent maintenance status, or other data that is pertinent to a repair. Location diagrams are especially useful when multiple components are installed, such as DIMMs, sockets, processor cards, fans, adapters, LEDs, and power supplies.
- *Remove or replace procedure labels* contain procedures that are often found on the cover of the system or in other locations that are accessible to the service representative. These labels provide systematic procedures, including diagrams, detailing how to remove and replace certain serviceable hardware components.
- *Numbered arrows* are used to indicate the order of operation and serviceability direction of components. Various serviceable parts, such as latches, levers, and touch points, must be pulled or pushed in a certain direction and order so that the mechanical mechanisms can engage or disengage. Arrows generally improve the ease of serviceability.

QR code labels for servicing information

A label containing a QR code can be found on the top service cover of the Power E850. This can be scanned with an appropriate app on a mobile device to link to a number of sources of information that simplify the servicing of the system.

From this quick access link you can find information about topics including:

- Installing and configuring the system
- Troubleshooting and problem analysis
- Reference code lookup tables
- Part location guides
- Removing and replacing field replaceable units
- Video guides for removal and installation of customer replaceable units
- Warranty and maintenance contracts
- Full product documentation

The operator panel

The operator panel on a POWER processor-based system is an LCD display (two rows of 16 characters) that is used to present boot progress codes, indicating advancement through the system power-on and initialization processes. The operator panel is also used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, enabling a service representative or administrator 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 wear from repeated plugging and unplugging. For these reasons, these devices are designed to be concurrently maintainable when properly configured. This allows parts to be replaced while the system is fully running, without requiring any downtime to applications. Concurrent maintenance is facilitated by the redundant design for the power supplies, fans, and physical storage.

The following system parts allow for concurrent maintenance:

- Disk drives and solid-state devices (SSDs)
- DVD drive

- Front fans
- PCIe adapters (including PCIe optical cable adapter for I/O expansion drawer)
- Power supplies
- Time-of-day battery card
- Operator panel

Maintenance procedures: Concurrent maintenance functions need to be initiated through a service interface such as the HMC (if available) or the ASMI service functions menu. Some concurrent maintenance can be initiated through the host operating system. Attempting to replace parts without following the correct procedures can lead to further faults or system damage.

Concurrent maintenance of the PCIe optical cable adapter requires an HMC.

Repair and verify services

Repair and verify (R&V) services are automated service procedures that are used to guide a service representative, step-by-step, through the process of repairing a system and verifying that the problem was repaired. The steps are customized in an appropriate sequence for the particular repair for the specific system being serviced. The following scenarios are covered by R&V services:

- Replacing a defective Field Replaceable Unit or a Customer Replaceable Unit
- Reattaching a loose or disconnected component
- Correcting a configuration error
- Removing or replacing an incompatible Field Replaceable Unit
- Updating firmware, device drivers, operating systems, middleware components, and IBM applications after replacing a part

R&V procedures can be used by service representatives who are familiar with the task and those who are not. Education-on-demand content is placed in the procedure at the appropriate locations. Throughout the R&V procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event to ensure that the guided maintenance procedures are operating correctly.

Clients can subscribe through the subscription services on the IBM Support Portal to obtain notifications about the latest updates that are available for service-related documentation.

4.6 Manageability

Several functions and tools are available to help manage the Power E850 server. These allow you to efficiently and effectively manage your system alongside other Power Systems servers and other machines.

4.6.1 Service user interfaces

The service interface allows support personnel or the client to communicate with the service support applications in a server by using a console, interface, or terminal. Delivering a clear, concise view of available service applications, the service interface allows the support team to manage system resources and service information in an efficient and effective way. Applications that are available through the service interface are carefully configured and placed to give service providers access to important service functions.

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

- Light Path (See “Light Path” on page 169 and “Service labels” on page 169)
- Service processor through the ASMI menus
- Operator panel (See “The operator panel” on page 170)
- Operating system service menu
- Service Focal Point via the HMC
- Service Focal Point Lite via the IVM

Service processor

The service processor is a micro processor-based controller that runs its own operating system. It is a component of the service interface card.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface that is connected to the POWER8 processor. The service processor is always running, regardless of the main system unit’s state. The system unit can be in the following states:

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

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring that the connection to the management console for manageability purposes and accepting ASMI Secure Sockets Layer (SSL) network connections. The service processor can view and manage the machine-wide settings by using the ASMI, and enables complete system and partition management from the HMC.

> **Analyzing a system that does not boot:** The FSP can analyze a system that does not boot. Reference codes and detailed data are available in the ASMI and are transferred to the HMC.

The service processor uses two Ethernet ports that run at 1 Gbps speed. Consider the following information:

- Both Ethernet ports are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI. The ASMI options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- Both Ethernet ports support only auto-negotiation. Customer-selectable media speed and duplex settings are not available.
- Both Ethernet ports have a default IP address, as follows:
  - Service processor eth0 (HMC1 port) is configured as 169.254.2.147.
  - Service processor eth1 (HMC2 port) is configured as 169.254.3.147.

The following functions are available through the service processor:

- Call Home
- Advanced System Management Interface (ASMI)
- Error information (error code, part number, and location codes) menu
- View of guarded components
- Limited repair procedures
- Generate dump
- LED Management menu

- Remote view of ASMI menus
- Firmware update through a USB key

Advanced System Management Interface

ASMI is the interface to the service processor that enables you to manage the operation of the server, such as auto-power restart, and to view information about the server, such as the error log and VPD. Various repair procedures require that you have a connection to the ASMI.

The ASMI is accessible through the HMC if used to manage the server. It is also accessible by using a web browser on a system that is connected directly to the service processor (in this case, either a standard Ethernet cable or a crossed cable) or through an Ethernet network. ASMI can also be accessed from an ASCII terminal, but this is available only while the system is in the platform powered-off mode.

Use the ASMI to change the service processor IP addresses or to apply certain security policies and prevent access from unwanted IP addresses or ranges.

You might be able to use the service processor's default settings. In that case, accessing the ASMI is not necessary. To access ASMI, use one of the following methods:

- Use an HMC.

  If configured to do so, the HMC connects directly to the ASMI for a selected system from this task.

  To connect to the ASMI from an HMC, complete the following steps:

  a. Open **Systems Management** from the navigation pane.
  b. From the work window, select one of the managed systems.
  c. From the System Management tasks list, click **Operations** → **Launch Advanced System Management (ASM)**.

- Use a web browser.

  At the time of writing, supported web browsers are Microsoft Internet Explorer (Version 10.0.9200.16439), Mozilla Firefox ESR (Version 24), and Chrome (Version 30). Later versions of these browsers might work, but are not officially supported. The JavaScript language and cookies must be enabled and TLS 1.2 might need to be enabled.

  The web interface is available during all phases of system operation, including the initial program load (IPL) and run time. However, several of the menu options in the web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides an SSL web connection to the service processor. To establish an SSL connection, open your browser by using the following address:

  ```
  https://<ip_address_of_service_processor>
  ```

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

- Use an ASCII terminal.

  The ASMI on an ASCII terminal supports a subset of the functions that are provided by the web interface and is available only when the system is in the platform powered-off mode. The ASMI on an ASCII console is not available during several phases of system operation, such as the IPL and run time.

- Command-line start of the ASMI.

  Either on the HMC itself or when properly configured on a remote system, it is possible to start ASMI web interface from the HMC command line. Open a terminal window on the HMC or access the HMC with a terminal emulation and run the following command:

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

On the HMC itself, a browser window opens automatically with the ASMI window and, when configured properly, a browser window opens on a remote system when issued from there.

The operator panel

The service processor provides an interface to the operator panel, which is used to display system status and diagnostic information. The operator panel can be accessed in two ways:

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

Here are several of the operator panel features:

- A 2 x 16 character LCD display
- Reset, enter, power On/Off, increment, and decrement buttons
- Amber system information or attention LED, and a green Power LED
- Blue enclosure identify LED
- Altitude sensor
- USB port
- Speaker

The following functions are available through the operator panel:

- Error information
- Generate dump
- View machine type, model, and serial number
- View or change IP addresses of the service processor
- Limited set of repair functions

Operating system service menu

The system diagnostic tests consist of stand-alone diagnostic tests that are loaded from the DVD drive, and online diagnostic tests (available in AIX).

Online diagnostic tests, when installed, are a part of the AIX operating system on the server. They can be booted in single-user mode (service mode), run in maintenance mode, or run concurrently (concurrent mode) with other applications. They have access to the AIX error log and the AIX configuration data.

The modes are as follows:

- Service mode

  This mode requires a service mode boot of the system and enables the checking of system devices and features. Service mode provides the most complete self-check of the system resources. All system resources, except the SCSI adapter and the disk drives that are used for paging, can be tested.

- Concurrent mode

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

- Maintenance mode

  This mode enables the checking of most system resources. Maintenance mode provides the same test coverage as service mode. The difference between the two modes is the way that they are started. Maintenance mode requires that all activity on the operating system is stopped. Run `shutdown -m` to stop all activity on the operating system and put the operating system into maintenance mode.

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

The service processor's error log can be accessed on the ASMI menus. You can also access the system diagnostics from a Network Installation Management (NIM) server.

Alternative method: When you order a Power E850, a DVD-RAM drive is available as an option. An alternative method for maintaining and servicing the system must be available if you do not order the DVD-RAM drive.

Depending on the operating system, the following service-level functions are what you typically see when you use the operating system service menus:

- Product activity log
- Trace Licensed Internal Code
- Work with communications trace
- Display/Alter/Dump
- Licensed Internal Code log
- Main storage memory dump manager
- Hardware service manager
- Call Home/Customer Notification
- Error information menu
- LED management menu
- Concurrent/Non-concurrent maintenance (within scope of the OS)
- Managing firmware levels
  - Server
  - Adapter
- Remote support (access varies by OS)

Service Focal Point on the Hardware Management Console

Optional HMC: An HMC is not required to run and manage a Power E850 server. If you are running multiple Power Systems servers, we recommend using an HMC for management tasks to simplify the overall management structure.

Service strategies become more complicated in a partitioned environment. The Manage Serviceable Events task in the management console can help streamline this process.

Each logical partition reports errors that it detects and forwards the event to the SFP application that is running on the management console, without determining whether other logical partitions also detect and report the errors. For example, if one logical partition reports an error for a shared resource, such as a managed system power supply, other active logical partitions might report the same error.

By using the Manage Serviceable Events task in the management console, you can avoid long lists of repetitive Call Home information by recognizing that these are repeated errors and consolidating them into one error.

In addition, you can use the Manage Serviceable Events task to initiate service functions on systems and logical partitions, including the exchanging of parts, configuring connectivity, and managing memory dumps.

4.6.2 IBM Power Systems Firmware maintenance

The IBM Power Systems Client-Managed Microcode is a methodology that enables you to manage and install microcode updates on Power Systems and its associated I/O adapters.

Firmware entitlement

With the HMC Version V8R8.1.0.0 and POWER8 processor-based servers, the firmware installations are restricted to entitled servers. The customer must be registered with IBM and entitled with a service contract. During the initial machine warranty period, the access key is already installed in the machine by manufacturing. The key is valid for the regular warranty period plus some additional time. The Power Systems Firmware is relocated from the public repository to the access control repository. The I/O firmware remains on the public repository, but the server must be entitled for installation. When the **`lslic`** command is run to display the firmware levels, a new value, **`update_access_key_exp_date`**, is added. The HMC GUI and the ASMI menu show the Update access key expiration date.

When the system is no longer entitled, the firmware updates fail. Some new System Reference Code (SRC) packages are available:

- E302FA06: Acquisition entitlement check failed
- E302FA08: Installation entitlement check failed

Any firmware release that was made available during the entitled time frame can still be installed. For example, if the entitlement period ends on 31 December 2015, and a new firmware release is released before the end of that entitlement period, it can still be installed. If that firmware is downloaded after 31 December 2015, but it was made available before the end of the entitlement period, it can still be installed. Any newer release requires a new update access key.

Note: The update access key expiration date requires a valid entitlement of the system to perform firmware updates.

You can find an update access key at the IBM CoD Home website:

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

To access the IBM entitled Software Support page for further details, go to the following website:

http://www.ibm.com/servers/eserver/ess

Firmware updates

System firmware is delivered as a release level or a service pack. Release levels support the general availability (GA) of new functions or features, and new machine types or models. Upgrading to a higher release level can be disruptive to customer operations. IBM intends to introduce no more than two new release levels per year. These release levels will be supported by service packs. Service packs contain only firmware fixes and do not introduce new functions. A service pack is an update to an existing release level.

If the system is managed by a management console, you use the management console to perform system firmware updates. By using the management console, you can take advantage of the Concurrent Firmware Maintenance (CFM) option when concurrent service packs are available. CFM is the firmware update process that can be partially or wholly concurrent or nondisruptive. With the introduction of CFM, IBM is increasing its clients' opportunity to stay on a given release level for longer periods. Clients that want maximum stability can defer until there is a compelling reason to upgrade, such as the following reasons:

- A release level is approaching its end-of-service date (that is, it has been available for about a year, and soon service will not be supported).
- They want to move a system to a more standardized release level when there are multiple systems in an environment with similar hardware.
- A new release has a new function that is required in the environment.
- A scheduled maintenance action causes a platform reboot, which provides an opportunity to also upgrade to a new firmware release.

Any required security patches or firmware fixes will be incorporated into service packs for the life of a given release level. Customers are not required to upgrade to the latest release level to ensure security and stability of their systems.

Firmware can also be updated by using a running partition.

The updating and upgrading of system firmware depends on several factors, including the current firmware that is installed, and what operating systems are running on the system. These scenarios and the associated installation instructions are comprehensively outlined in the firmware section of Fix Central, found at the following website:

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

You might also want to review the preferred practice white papers that are found at the following website:

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

Firmware update steps

The system firmware consists of service processor microcode, Open Firmware microcode, and Systems Power Control Network (SPCN) microcode.

The firmware and microcode can be downloaded and installed either from the HMC, or from a running partition.

Power Systems has a permanent firmware boot side (A side) and a temporary firmware boot side (B side). New levels of firmware must be installed first on the temporary side to test the update's compatibility with existing applications. When the new level of firmware is approved, it can be copied to the permanent side.

For access to the initial websites that address this capability, see the POWER8 section on the IBM Support Portal:

https://www.ibm.com/support/entry/portal/product/power

For POWER8 based Power Systems, select the **POWER8 systems** link.

Within this section, search for **Firmware and HMC updates** to find the resources for keeping your system's firmware current.

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

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

- Installed level

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

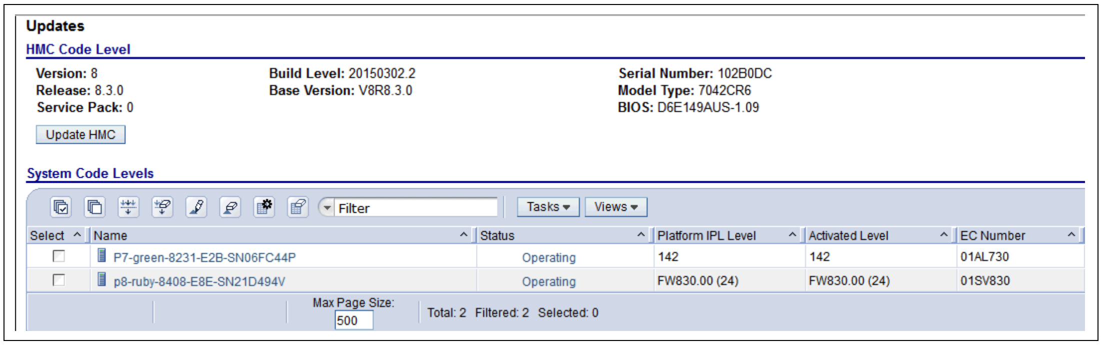
- Activated level

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

- Accepted level

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

Figure 4-4 shows the different levels as shown in the HMC.



Updates

HMC Code Level

Version: 8
Release: 8.3.0
Service Pack: 0

Build Level: 20150302.2
Base Version: V8R8.3.0

Serial Number: 102B0DC
Model Type: 7042CR6
BIOS: D6E149AUS-1.09

System Code Levels

Select	Name	Status	Platform IPL Level	Activated Level	EC Number
□	P7-green-8231-E2B-SN06FC44P	Operating	142	142	01AL730
□	p8-ruby-8408-E8E-SN21D494V	Operating	FW830.00 (24)	FW830.00 (24)	01SV830

Max Page Size: 500 Total: 2 Filtered: 2 Selected: 0

Figure 4-4 HMC and server firmware levels shown in the HMC

IBM provides the CFM function on the Power E850 model. This function supports applying nondisruptive system firmware service packs to the system concurrently (without requiring a reboot operation to activate changes).

The concurrent levels of system firmware can, on occasion, contain fixes that are known as *deferred*. These deferred fixes can be installed concurrently but are not activated until the next IPL. Deferred fixes, if any, are identified in the Firmware Update Descriptions table of the firmware document. For deferred fixes within a service pack, only the fixes in the service pack that cannot be concurrently activated are deferred.

Table 4-1 shows the file-naming convention for system firmware.

Table 4-1 Firmware naming convention

PPNNSSS_FFF_DDD			
PP	Package identifier	for example, 01	-
NN	Platform and class	for example, SV	Scale out systems
SSS	Release indicator		

PPNNSSS_FFF_DDD	
FFF	Current fix pack
DDD	Last disruptive fix pack

The following example uses the convention:

01SV830_010_010 = Firmware for 8208-E8E release 830 fix pack 10

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.

4.6.3 Concurrent firmware maintenance improvements

Since POWER6, firmware service packs are concurrently applied and take effect immediately. Occasionally, a service pack is shipped where most of the features can be concurrently applied, but because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during operation, a patch in this area required a system reboot for activation.

With the Power-On Reset Engine (PORE), the firmware can now dynamically power off processor components, change the registers, and reinitialize while the system is running, without discernible impact to any applications running on a processor. This allows concurrent firmware changes in POWER8, which in earlier designs required a reboot to take effect.

Activating new firmware functions requires installation of a higher firmware release level. This process is disruptive to server operations and requires a scheduled outage and full server reboot.

4.6.4 Electronic Services and Electronic Service Agent

IBM transformed its delivery of hardware and software support services to help you achieve higher system availability. Electronic Services is a web-enabled solution that offers an exclusive, no additional charge enhancement to the service and support that is available for IBM servers. These services provide the opportunity for greater system availability with faster problem resolution and preemptive monitoring. The Electronic Services solution consists of two separate, but complementary, elements:

- Electronic Services news page
- Electronic Service Agent

Electronic Services news page

The Electronic Services news page is a single Internet entry point that replaces the multiple entry points that are traditionally 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 runs on the 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 an IBM SSR when they help answer your questions or diagnose problems. Installation and use of ESA for problem reporting enables IBM to provide better support and service for your IBM server.

To learn how Electronic Services can work for you, see the following website (an IBM ID is required):

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

Here are some of the benefits of Electronic Services:

- Increased uptime

  The ESA tool enhances the warranty or maintenance agreement by providing faster hardware error reporting and uploading system information to IBM Support. This can translate to less time that is wasted monitoring the symptoms, diagnosing the error, and manually calling IBM Support to open a problem record.

  Its 24x7 monitoring and reporting mean no more dependence on human intervention or off-hours customer personnel when errors are encountered in the middle of the night.

- Security

  The ESA tool is designed to be secure in monitoring, reporting, and storing the data at IBM. The ESA tool securely transmits through the Internet (HTTPS or VPN), and can be configured to communicate securely through gateways to provide customers with a single point of exit from their site.

  Communication is one way. Activating ESA does not enable IBM to call into a customer's system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. It is viewable only by the customer and IBM. The customer's business applications or business data is never transmitted to IBM.

- More accurate reporting

  Because system information and error logs are automatically uploaded to the IBM Support center with the service request, customers are not required to find and send system information, decreasing the risk of misreported or misdiagnosed errors.

  When inside IBM, problem error data is run through a data knowledge management system and knowledge articles are appended to the problem record.

- Customized support

  By using the IBM ID that you enter during activation, you can view system and support information by selecting **My Systems** at the Electronic Support website:

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

  My Systems provides valuable reports of installed hardware and software, by using information that is collected from the systems by ESA. Reports are available for any system that is associated with the customer’s IBM ID. Premium Search combines the function of search and the value of ESA information, providing advanced search of the technical support knowledge base. Using Premium Search and the ESA information that was collected from your system, your clients can see search results that apply specifically to their systems.

For more information about how to use the power of IBM Electronic Services, contact your IBM SSR, or see the following website:

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

Service Event Manager

The Service Event Manager (SEM) allows the user to decide which of the Serviceable Events are called home with the ESA. It is possible to lock certain events. Some customers might not allow data to be transferred outside their company. After the SEM is enabled, the analysis of the possible problems might take longer.

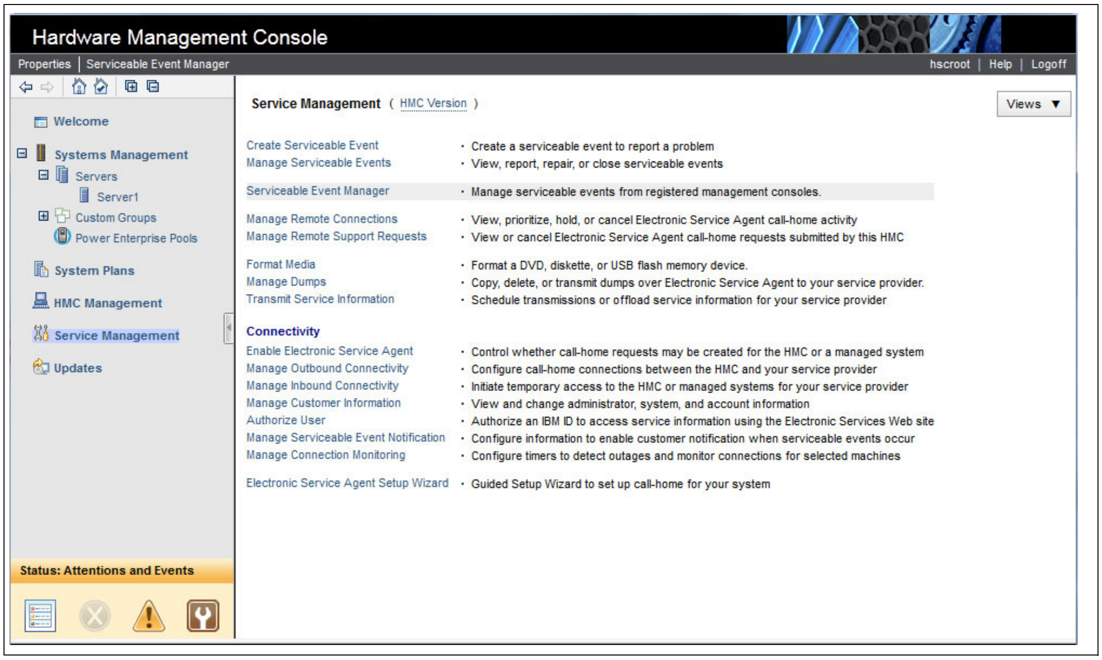
- The SEM can be enabled by running the following command:

```
chhmc -c sem -s enable
```

- You can disable SEM mode and specify what state in which to leave the Call Home feature by running the following commands:

```
chhmc -c sem -s disable --callhome disable
chhmc -c sem -s disable --callhome enable
```

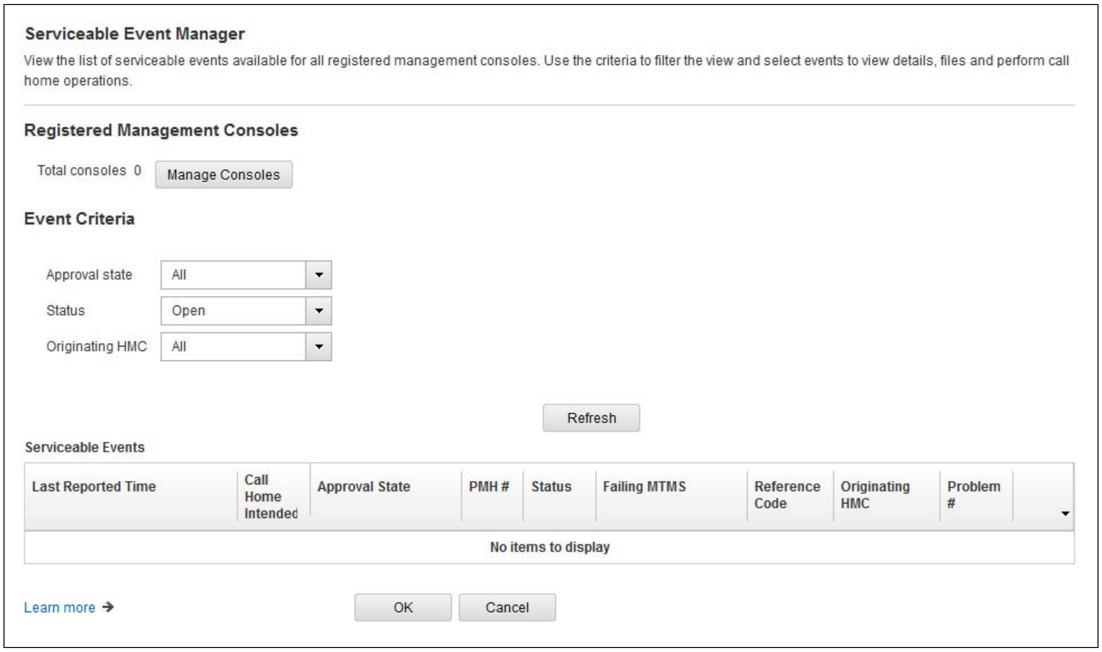
You can do the basic configuration of the SEM from the HMC GUI. After you select the Service Event Manager, as shown in Figure 4-5, you must add the HMC console.



Create Serviceable Event	• Create a serviceable event to report a problem
Manage Serviceable Events	• View, report, repair, or close serviceable events
Serviceable Event Manager	• Manage serviceable events from registered management consoles.
Manage Remote Connections	• View, prioritize, hold, or cancel Electronic Service Agent call-home activity
Manage Remote Support Requests	• View or cancel Electronic Service Agent call-home requests submitted by this HMC
Format Media	• Format a DVD, diskette, or USB flash memory device.
Manage Dumps	• Copy, delete, or transmit dumps over Electronic Service Agent to your service provider.
Transmit Service Information	• Schedule transmissions or offload service information for your service provider
Connectivity	
Enable Electronic Service Agent	• Control whether call-home requests may be created for the HMC or a managed system
Manage Outbound Connectivity	• Configure call-home connections between the HMC and your service provider
Manage Inbound Connectivity	• Initiate temporary access to the HMC or managed systems for your service provider
Manage Customer Information	• View and change administrator, system, and account information
Authorize User	• Authorize an IBM ID to access service information using the Electronic Services Web site
Manage Serviceable Event Notification	• Configure information to enable customer notification when serviceable events occur
Manage Connection Monitoring	• Configure timers to detect outages and monitor connections for selected machines
Electronic Service Agent Setup Wizard	• Guided Setup Wizard to set up call-home for your system

Figure 4-5 HMC selection for Service Event Manager

In the next window, you can configure the HMC that is used to manage the Serviceable Events and proceed with further configuration steps, as shown in Figure 4-6.



Serviceable Event Manager

View the list of serviceable events available for all registered management consoles. Use the criteria to filter the view and select events to view details, files and perform call home operations.

Registered Management Consoles

Total consoles 0

Event Criteria

Approval state: All

Status: Open

Originating HMC: All

Serviceable Events

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

Figure 4-6 Initial SEM window

Here are detailed descriptions of the different configurable options:

- Registered Management Consoles

  “Total consoles” lists the number of consoles that are registered. Select **Manage Consoles** to manage the list of RMCs.

- Event Criteria

  Select the filters for filtering the list of serviceable events that are shown. After the selections are made, click **Refresh** to refresh the list based on the filter values.

- Approval state

  Select the value for approval state to filter the list.

- Status

  Select the value for the status to filter the list.

- Originating HMC

  Select a single registered console or **All consoles** to filter the list.

- Serviceable Events

  The Serviceable Events table shows the list of events based on the filters that are selected. To refresh the list, click **Refresh**.

The following menu options are available when you select an event in the table:

- View Details...

  Shows the details of this event.

- View Files...

  Shows the files that are associated with this event.

- Approve Call Home

  Approves the Call Home of this event. This option is available only if the event is not approved already.

The Help/Learn more function can be used to get more information about the other available windows for the Serviceable Event Manager.

4.7 Selected POWER8 RAS capabilities by operating system

Table 4-2 provides a list of the Power Systems RAS capabilities by operating system. The HMC is an optional feature on the Power E850 server.

Table 4-2 Selected RAS features by operating system

RAS feature	**AIX** **V7.1 TL3 SP5** **V6.1 TL9 SP5**	**Linux** **RHEL6.6** **RHEL7.1** **SLES12** **Ubuntu 15.04**
Processor		
FFDC for fault detection/error isolation	X	X
Processor instruction retry	X	X
Dynamic processor deallocation	X	X
Core error recovery		
► Alternative processor recovery	X	X
► Partition core contained checkstop	X	X
I/O subsystem		
PCI Express bus enhanced error detection	X	X
PCI Express bus enhanced error recovery	X	X
PCI Express card hot-swap	X	X
Cache availability		
Cache line removal	X	X
Dynamic bit-line sparing	X	X
Special uncorrectable error handling	X	X
Memory availability		
Memory page deallocation	X	X
Dynamic DRAM sparing	X	X
Periodic memory scrubbing	X	X

RAS feature	AIX V7.1 TL3 SP5 V6.1 TL9 SP5	Linux RHEL6.6 RHEL7.1 SLES12 Ubuntu 15.04
Special uncorrectable error handling	X	X
Fault detection and isolation		
Storage Protection Keys	X	Not used by OS
Error log analysis	X	X
Serviceability		
Boot-time progress indicators	X	X
Firmware error codes	X	X
Operating system error codes	X	X
Inventory collection	X	X
Environmental and power warnings	X	X
Hot-swap DASD/media	X	X
Dual disk controllers/split backplane	X	X
Active-active dual disk controllers	optional	optional
EED collection	X	X
SP “Call Home” on non-HMC configurations	X	X
IO adapter/device stand-alone diagnostic tests with PowerVM	X	X
SP mutual surveillance with POWER Hypervisor	X	X
Concurrent firmware update with HMC	X	X
Service Agent Call Home Application	X	X
Service indicator LED support	X	X
System dump for memory, POWER Hypervisor, and SP	X	X
IBM Knowledge Center/IBM Systems Support Site service publications	X	X
System Support Site education	X	X
Operating system error reporting to HMC SFP application	X	X
RMC secure error transmission subsystem	X	X
Healthcheck scheduled operations with HMC	X	X
Operator panel (real or virtual)	X	X
Concurrent operator panel maintenance	X	X
Redundant HMCs	optional	optional
Automated server recovery/restart	optional	optional

RAS feature	**AIX** **V7.1 TL3 SP5** **V6.1 TL9 SP5**	**Linux** **RHEL6.6** **RHEL7.1** **SLES12** **Ubuntu 15.04**
High availability clustering support	X	X
Repair and Verify guided maintenance with HMC	X	X
PowerVM Live Partition/Live Application Mobility with PowerVM Enterprise Edition	X	X
Emergency power-off warning (EPOW)		
EPOW errors handling	X	X

Related publications

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

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Some publications referenced in this list might be available in softcopy only.

- *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491
- *IBM Power Systems S812L and S822L Technical Overview and Introduction*, REDP-5098
- *IBM Power Systems S814 and S824 Technical Overview and Introduction*, REDP-5097
- *IBM Power System S822 Technical Overview and Introduction*, REDP-5102
- *IBM Power System S824L Technical Overview and Introduction*, REDP-5139
- *IBM Power Systems E870 and E880 Technical Overview and Introduction*, REDP-5137
- *IBM Power Systems SR-IOV Technical Overview and Introduction*, REDP-5065
- *IBM PowerVM Best Practices*, SG24-8062
- *IBM PowerVM Enhancements What is New in VIOS 2.2.3*, 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

Other publications

These publications are also relevant as further information sources:

- *Active Memory Expansion: Overview and Usage Guide*

  http://www.ibm.com/systems/power/hardware/whitepapers/am_exp.html
- *IBM EnergyScale for POWER8 Processor-Based Systems* white paper:

  http://public.dhe.ibm.com/common/ssi/ecm/po/en/pow03125usen/POW03125USEN.PDF
- IBM Power Facts and Features: IBM Power Systems, IBM PureFlex System, and Power Blades

  http://www.ibm.com/systems/power/hardware/reports/factsfeatures.html
- IBM Power System S812L and S822L server specifications

  http://www.ibm.com/systems/power/hardware/s812l-s822l/specs.html

- IBM Power System S814 server specifications

  http://www.ibm.com/systems/power/hardware/s814/specs.html
- IBM Power System S822 server specifications

  http://www.ibm.com/systems/power/hardware/s822/specs.html
- IBM Power System S812L and S822L server specifications

  http://www.ibm.com/systems/power/hardware/s812l-s822l/specs.html
- IBM Power System S824 server specifications

  http://www.ibm.com/systems/power/hardware/s824/specs.html
- IBM Power System S824L server specifications:

  http://www.ibm.com/systems/power/hardware/s824l/specs.html
- IBM Power System E850 server specifications:

  http://www.ibm.com/systems/power/hardware/e850/specs.html
- IBM Power System E870 server specifications:

  http://www.ibm.com/systems/power/hardware/e870/specs.html
- Specific storage devices that are supported for Virtual I/O Server

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

Online resources

These websites are also relevant as further information sources:

- 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 Power8 systems information: IBM Knowledge Center

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

  http://www.ibm.com/systems/storage
- IBM System Planning Tool website

  http://www.ibm.com/systems/support/tools/systemplanningtool
- IBM Systems Energy Estimator

  http://www-912.ibm.com/see/EnergyEstimator
- Migration combinations of processor compatibility modes for active Partition Mobility

  http://www.ibm.com/support/knowledgecenter/POWER7/p7hc3/iphc3pcmcombosact.htm?cp=POWER7%2F1-8-3-7-2-0-4-3-0
- Power Systems Capacity on Demand website

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

- Support for IBM Systems website

  http://www.ibm.com/support/entry/portal/Overview?brandind=Hardware~Systems~Power

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5222-00

ISBN 0738454222

Printed in U.S.A.