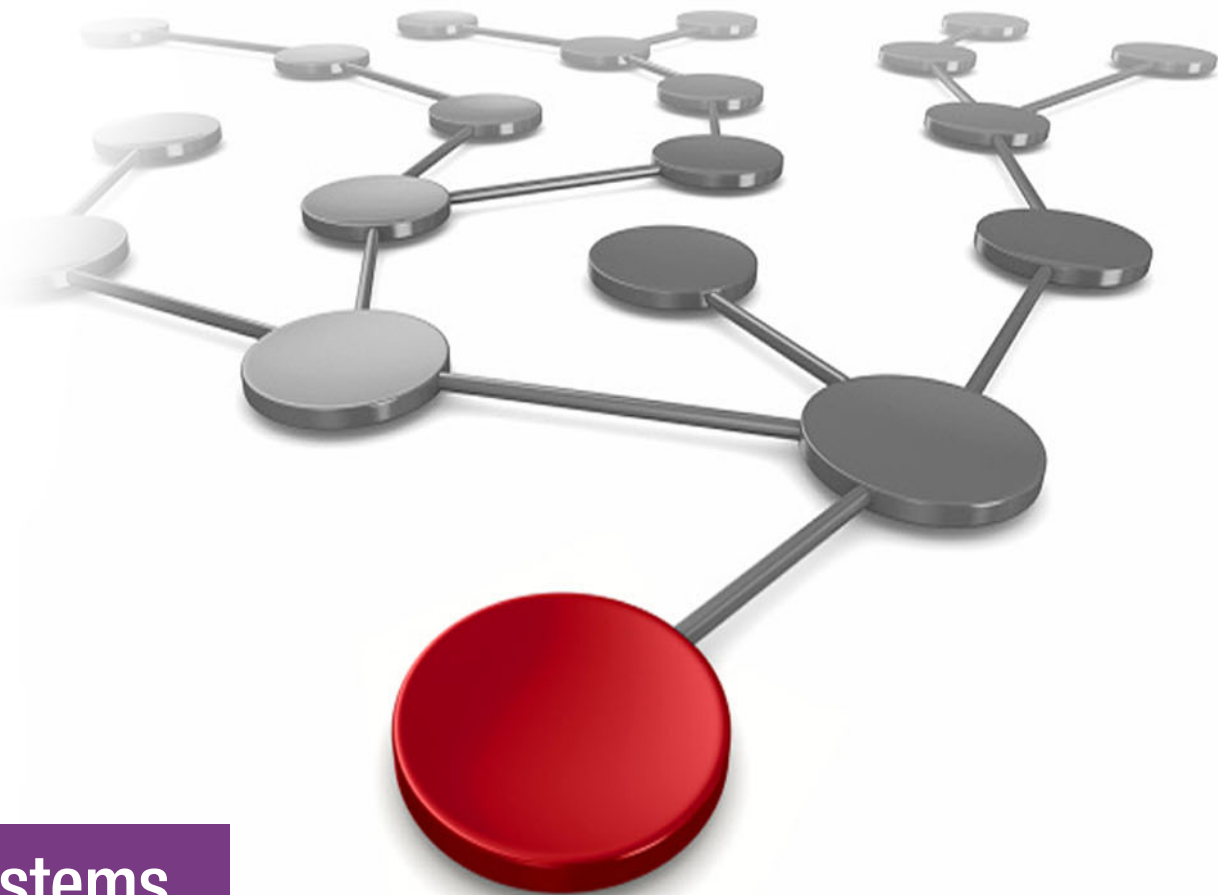


IBM Power System E950

Technical Overview and Introduction

James Cruickshank
Yongsheng Li (Victor)
Armin Röhl



Power Systems



International Technical Support Organization

IBM Power System E950: Technical Overview and Introduction

August 2018

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

First Edition (August 2018)

This edition applies to the IBM Power System E950 (9040-MR9) system.

Important: At time of publication, this book is based on a pre-GA version of a product. For the most up-to-date information regarding this product, consult the product documentation or subsequent updates of this book.

© Copyright International Business Machines Corporation 2018. All rights reserved.

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

Contents

Notices	vii
Trademarks	viii
Preface	ix
Authors	ix
Now you can become a published author, too!	x
Comments welcome	xi
Stay connected to IBM Redbooks	xi
Chapter 1. General description	1
1.1 System overview	3
1.2 Operating environment	7
1.3 Physical package	8
1.4 System features	9
1.4.1 Minimum configuration	10
1.4.2 Power supply features	11
1.4.3 Processor module features	11
1.4.4 Memory features	12
1.4.5 PCIe slots	14
1.4.6 USB	15
1.4.7 Disk and media features	16
1.5 I/O drawers	20
1.5.1 PCIe Gen3 I/O Expansion Drawer	21
1.5.2 I/O drawers and usable PCI slots	22
1.5.3 EXP24S SFF Gen2-bay Drawer	23
1.5.4 EXP24SX and EXP12SX SAS Storage Enclosures	24
1.6 System racks	25
1.6.1 New racks	26
1.6.2 IBM Enterprise 42U Slim Rack 7965-S42	26
1.6.3 IBM 7014 Model T42 rack	26
1.6.4 1.8 Meter Rack (#0551)	28
1.6.5 2.0 Meter Rack (#0553)	28
1.6.6 Rack (#ER05)	28
1.6.7 The AC power distribution unit and rack content	29
1.6.8 Rack-mounting rules	32
1.6.9 Useful rack additions	32
1.6.10 Original equipment manufacturer racks	35
1.7 Hardware Management Console	36
1.7.1 New features of the Hardware Management Console	36
1.7.2 Hardware Management Console overview	36
1.7.3 Hardware Management Console code level	38
1.7.4 Two architectures of Hardware Management Console	39
1.7.5 Connectivity to POWER9 processor-based systems	40
1.7.6 High availability Hardware Management Console configuration	41
Chapter 2. Architecture and technical overview	43
2.1 The IBM POWER9 processor	44
2.1.1 POWER9 processor overview	44
2.1.2 POWER9 processor core	46

2.1.3	Simultaneous multithreading	47
2.1.4	POWER9 compatibility modes	48
2.1.5	Processor feature codes	48
2.1.6	Memory access	50
2.1.7	On-chip L3 cache innovation and intelligent caching	51
2.1.8	Hardware transactional memory	52
2.1.9	POWER9 accelerator processor interfaces	52
2.1.10	Power and performance management	54
2.1.11	Comparison of the POWER9, POWER8, and POWER7+ processors	57
2.2	Memory subsystem	58
2.2.1	DIMM memory riser card	59
2.2.2	Memory placement rules	62
2.2.3	Memory activations	63
2.2.4	Memory throughput	64
2.2.5	Active Memory Mirroring	66
2.2.6	Memory error correction and recovery	68
2.2.7	Special Uncorrectable Error handling	68
2.3	Capacity on Demand	68
2.3.1	Capacity on Demand: New features	69
2.3.2	Capacity Upgrade on Demand	69
2.3.3	Processor activations	70
2.3.4	Elastic Capacity on Demand (Temporary)	71
2.3.5	Utility Capacity on Demand	72
2.3.6	Trial Capacity on Demand	73
2.3.7	Software licensing and CoD	73
2.3.8	Solution Edition for Healthcare	73
2.4	System buses	74
2.4.1	PCIe Gen4	74
2.5	PCIe adapters	76
2.5.1	New PCIe adapter features	76
2.5.2	PCIe	76
2.5.3	LAN adapters	77
2.5.4	Graphics adapters	77
2.5.5	SAS adapters	78
2.5.6	Fibre Channel adapters	78
2.5.7	USB ports	78
2.5.8	InfiniBand host channel adapters	79
2.5.9	Cryptographic Coprocessor	79
2.5.10	CAPI adapters	80
2.6	Internal storage	80
2.6.1	Backplane features	81
2.6.2	Internal NVMe SSD drives	83
2.6.3	Supported RAID functions	83
2.6.4	Easy Tier	84
2.6.5	Media drawers	86
2.6.6	External DVD drives	87
2.6.7	RDX removable disk drives	87
2.7	External I/O subsystems	88
2.7.1	PCIe Gen3 I/O Expansion Drawer	88
2.7.2	PCIe Gen3 I/O Expansion Drawer optical cabling	90
2.7.3	PCIe Gen3 I/O Expansion Drawer system power control network cabling	93
2.8	External disk subsystems	94
2.8.1	EXP24S SFF Gen2-bay Drawer	94

2.8.2	EXP24SX and EXP12SX SAS Storage Enclosure	103
2.8.3	IBM System Storage	105
2.9	Operating system support	106
2.9.1	AIX operating system	107
2.9.2	Linux operating system	108
2.9.3	Virtual I/O Server	108
Chapter 3.	Virtualization	109
3.1	IBM POWER Hypervisor	110
3.1.1	POWER processor modes	113
3.2	Active Memory Expansion	115
3.3	Single Root I/O Virtualization	115
3.4	PowerVM	116
3.4.1	Multiple shared processor pools	118
3.4.2	Virtual I/O Server	118
3.4.3	Live Partition Mobility	120
3.4.4	Active Memory Sharing	120
3.4.5	Active Memory Deduplication	120
3.4.6	Remote Restart	121
Chapter 4.	Reliability, availability, serviceability, and manageability	123
4.1	Power E950 specific RAS enhancements	124
4.2	Reliability	126
4.2.1	Designed for reliability	127
4.2.2	Placement of components	128
4.3	Processor RAS details	128
4.3.1	Correctable error introduction	128
4.3.2	Uncorrectable error introduction	129
4.3.3	Processor core/cache error handling	129
4.3.4	Cache uncorrectable error handling	130
4.3.5	Cyclic redundancy check and lane repair for processor fabric buses	130
4.3.6	Processor instruction retry and other try again techniques	131
4.3.7	Predictive processor deallocation	131
4.3.8	Core-contained checkstops and PowerVM handled errors	131
4.3.9	PCIe controller and enhanced error handling	132
4.3.10	Memory channel checkstops and hypervisor memory mirroring	132
4.3.11	Persistent guarding of failed elements	132
4.4	Memory RAS details	133
4.5	PCIe I/O subsystem RAS details	134
4.5.1	I/O subsystem availability and enhanced error handling	134
4.5.2	PCIe Gen3 I/O Expansion drawer RAS	135
4.6	Enterprise systems availability	137
4.7	Availability effects of a solution architecture	138
4.7.1	Clustering	138
4.7.2	Virtual I/O redundancy configurations	138
4.7.3	PowerVM Live Partition Mobility	139
4.8	Serviceability	141
4.8.1	Detecting errors	141
4.8.2	Error checkers, fault isolation registers, and first failure data capture	141
4.8.3	Service processor	142
4.8.4	Diagnosing	143
4.8.5	Reporting	145
4.8.6	Notifying	146

4.8.7 Locating and servicing	147
4.9 Manageability	150
4.9.1 Service user interfaces	150
4.9.2 Power Systems Firmware maintenance	155
4.9.3 Concurrent firmware maintenance improvements	158
4.9.4 Electronic Services and Electronic Service Agent	158
4.10 Selected POWER9 RAS capabilities by operating system	162
Related publications	165
IBM Redbooks	165
Online resources	165
Help from IBM	166

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

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

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

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

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

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

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

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


COPYRIGHT LICENSE:

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

Trademarks

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

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

Active Memory™	IBM Z®	POWER9™
AIX®	Micro-Partitioning®	PowerHA®
AIX 5L™	OpenCAPI™	PowerVM®
Db2®	POWER®	PowerVP™
DS8000®	Power Architecture®	Redbooks®
Easy Tier®	POWER Hypervisor™	Redpaper™
Electronic Service Agent™	Power Systems™	Redbooks (logo)  ®
EnergyScale™	POWER6®	RS/6000®
IBM®	POWER6+™	Storwize®
IBM FlashSystem®	POWER7®	System Storage®
IBM Spectrum™	POWER7+™	SystemMirror®
IBM Spectrum Accelerate™	POWER8®	XIV®

The following terms are trademarks of other companies:

Intel, Intel Xeon, 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.

LTO, Ultrium, the LTO Logo and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

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

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

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

Preface

This IBM® Redpaper™ publication gives a broad understanding of a new architecture of the IBM Power System E950 (9040-MR9) server that supports IBM AIX®, and Linux operating systems. The objective of this paper is to introduce the major innovative Power E950 offerings and relevant functions:

- ▶ The IBM POWER9™ processor, which is available at frequencies of 2.8 - 3.4 GHz.
- ▶ Significantly strengthened cores and larger caches.
- ▶ Supports up to 16 TB of memory, which is four times more than the IBM POWER8® processor-based IBM Power System E850 server.
- ▶ Integrated I/O subsystem and hot-pluggable Peripheral Component Interconnect Express (PCIe) Gen4 slots, which have double the bandwidth of Gen3 I/O slots.
- ▶ Supports EXP12SX and ESP24SX external disk drawers, which have 12 Gb Serial Attached SCSI (SAS) interfaces and support Active Optical Cables (AOCs) for greater distances and less cable bulk.
- ▶ New IBM EnergyScale™ technology offers new variable processor frequency modes that provide a significant performance boost beyond the static nominal frequency.

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 (ISVs)

This paper expands the current set of Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power E950 server.

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

Authors

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

James Cruickshank works in the Power Systems Client Technical Specialist team for IBM in the UK. He holds an honors degree in Mathematics from the University of Leeds. James has over 17 years experience working with IBM RS/6000®, IBM pSeries, IBM System p, and Power Systems products. James supports customers in the financial services sector in the UK.

Yongsheng Li (Victor) works in the Power Systems Level 2 Support team for IBM China. He holds a master's degree in Computer Application Technology from Graduate University of Chinese Academy of Sciences. Victor has 15 years of experience working on RS/6000, IBM System Storage®, AIX, Hardware Management Console (HMC), System p, and Power Systems products.

Armin Röhl works as a Power Systems IT specialist in Germany. He holds a degree in Experimental Physics from the University of Hamburg, Germany. Armin has 22 years of experience in Power Systems and AIX pre-sales technical support. He co-authored the AIX Version 4.3.3, the IBM AIX 5L™ Version 5.0, the AIX 5L Version 5.3, the AIX Version 6.1, and the AIX 7.1 Differences Guide IBM Redbooks® publications.

The project that produced this publication was managed by:

Scott Vetter
PMP, IBM US

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

Brian Allison, Joanna Z Bartz, Corine Diby, Arnold Flores, Nigel Griffiths, James Hermes, Daniel Henderson, Vic Mahaney, Charles Marino, Michael Mueller, Kaveh Naderi, Hoa Nguyen, Jeff Jajowka, Todd Rosedahl, Roxette Johnson, David Sheffield, Michelle Skibitzki, Alan Standridge, Bill Starke, Jeff Stuecheli, Brian M Werneke
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 by using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:

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

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:

<http://www.facebook.com/IBMRedbooks>

- ▶ Follow us on Twitter:

<http://twitter.com/ibmredbooks>

- ▶ Look for us on LinkedIn:

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

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

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

- ▶ Stay current on recent Redbooks publications with RSS Feeds:

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



General description

This chapter provides general description of the new IBM Power System E950 (9040-MR9) server, which is a logical follow-on to the IBM Power System E850 and IBM Power System E850C servers. Due to the new POWER9 processor-based architecture, the Power E950 provides improvements in the economics of application delivery and IT services through increased price/performance that is based on increased throughput, reduced response times, and increased memory and I/O bandwidth.

Figure 1-1 shows an exploded view of the Power E950 server.



Figure 1-1 Exploded view of a Power E950 server

1.1 System overview

The new Power E950 server is the successor of the Power E850 and Power E850C servers. It is the most agile, reliable, and high-performance system in the marketplace. It is ideal for cloud deployments because of its built-in virtualization, flexible capacity, and high usage.

The machine type model number of the Power E950 server is 9040-MR9. It is a single enclosure server that is four EIA units tall (4U). It can be configured with two or four processor modules.

Figure 1-2 shows the Power E950 server.



Figure 1-2 The Power E950 server

Figure 1-3 shows a top view of the Power E950 server with the top lid removed.

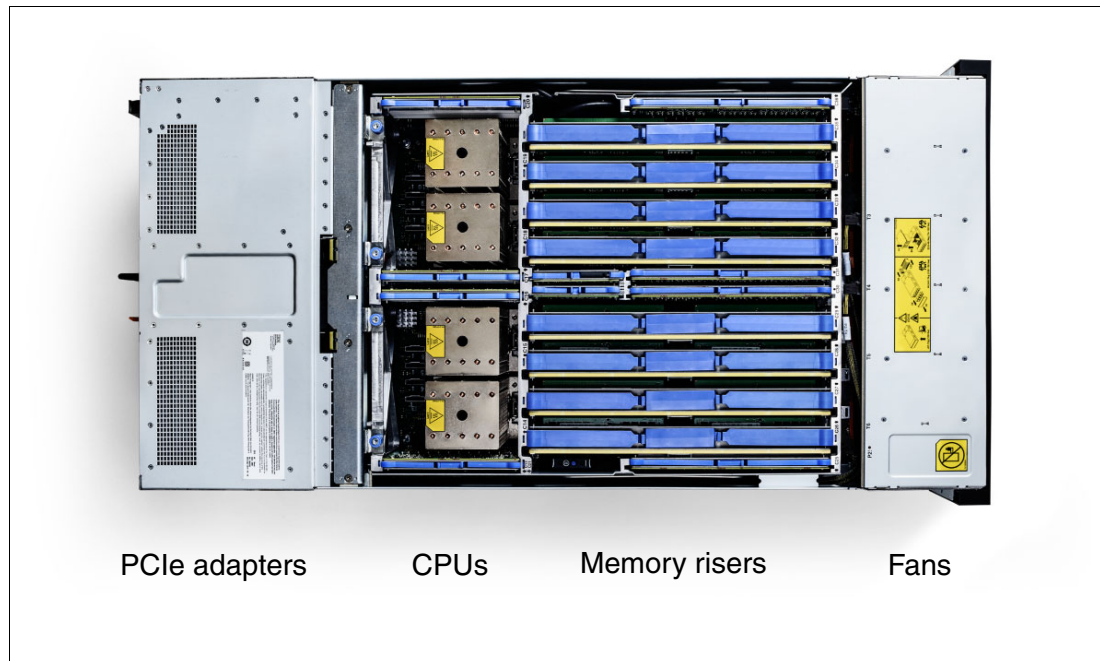


Figure 1-3 Top view of the Power E950 server with top lid removed

This scale-up server offers up to 48 processor cores with four processor modules. It supports up to 16 TB of memory, which is four times more than the Power E850 server.

Up to 11 Peripheral Component Interconnect Express (PCIe) slots are provided in the system unit. These slots include eight 16-lane PCIe Gen4 slots, two 8-lane Gen4 slots, and one 8-lane Gen3 slot. Up to 48 extra PCIe Gen3 slots can be added by using PCIe Gen3 I/O Expansion Drawers.

The Power E950 server offers up to eight internal Storage Serial Attached SCSI (SAS) bays and four Non-Volatile Memory Express (NVMe) solid-state drives (SSDs) in 2.5" small form factor (SFF). Each NVMe drive contains its own controller and connects to the system by using PCIe Gen3 ports.

In addition to extensive hardware configuration flexibility, the Power E950 server offers Elastic Capacity on Demand (Temporary) (Elastic CoD) for both processor cores and memory, IBM Active Memory™ Expansion, and Active Memory Mirroring (AMM) for Hypervisor.

The Power E950 server also provides strong reliability, availability, and serviceability (RAS) characteristics, which include POWER9 chip capabilities, memory protection, multiple SAS storage protection options, hot-plug SAS bays, hot-plug NVMe bays, hot-plug PCIe slots, redundant and hot-plug power supplies and cooling fans, hot-plug redundant cooling fans, hot-plug Time of Day battery, and even highly resilient architecture for power regulators.

The Power E950 server is optimized for running AIX and Linux workloads. IBM i is not a supported operating system on the Power E950 server.

Table 1-1 shows a summary of features of the Power E950 server.

Table 1-1 Power E950 server feature summary

Feature	Comments
Processors	8, 10, 11, or 12 CPU cores per socket.
Sockets	Four sockets are available. 2 or 4 sockets may be populated.
Memory	Eight riser cards with 16 DDR4 dual inline memory modules (DIMMs).
	DIMM sizes 8, 16, 32, 64, and 128 GB.
	16 TB maximum memory.
Media bays	DVD through an external USB DVD.
Integrated PCIe	PCIe Gen 4: Eight 16-lane + two 8-lane slots (2 sockets = 4 x 16).
	PCIe Gen 3: One 8-lane slot. (The default is two 10 GB + two 1Gb Ethernet slots.)
	PCIe slots are full-high and half-length, and use blind-swap cassettes (BSCs).
Internal SAS bays	Up to eight SAS 2.5-inch drives by using 1 or 2 SAS adapters. Split disk-capable.
Internal Storage bays	USB 2.0: 1 - 4 NVMe slots for 1 - 4 NVMe devices. USB 3.0: Two front and two rear.
Maximum I/O drawers	Four PCIe Gen3 I/O drawers (EMX0).
External storage drawers	EXP12SX, ESP24SX, and EXP24S ^a .

a. The EXP24S drawer is support-only (at a time after the initial general availability (GA)) and cannot be ordered together with a Power E950 server.

Table 1-2 shows the major differences between the Power E850C and Power E950 servers.

Table 1-2 Comparison between Power E850C and Power E950 servers

Features	Power E850C server	Power E950 server
Processor	POWER8 (dual-chip module (DCM))	POWER9 (single-chip module (SCM))
Sockets	2 - 4	2 - 4
Cores	32, 40, or 48	32, 40, 44, or 48
Maximum memory	4 TB	16 TB
DIMM type/DIMM slots count	Up to 32 CDIMMs	Up to 128 industry-standard DIMMs (IS DIMMs)
L4 cache	Yes	Yes
Memory bandwidth	768 GBps	920 GBps
Memory DRAM spare	Yes	Yes
I/O expansion slots	Yes	Yes

Features	Power E850C server	Power E950 server
PCIe slots	11 (Eight Gen3 16-lane + three Gen3 8-lane slots)	11 (Eight Gen4 16-lane + two Gen4 8-lane + one Gen3 slots)
Acceleration ports	Yes (Coherent Accelerator Processor Interface (CAPI) 1.0)	Yes (CAPI 2.0 + IBM OpenCAPI™)
PCIe hot-plug support	Yes	Yes + Blind swap
IO bandwidth	315 GBps	630 GBps
Internal storage bays	12 SAS disk bays (eight 2.5" + four 1.8" drives)	12 (eight SAS + four NVMe drives)
Internal storage controllers	Integrated Non-concurrent/dedicated service	Optional Concurrently Maintainable service

Figure 1-4 shows the front view of a Power E950 server with the front bezel removed.



Figure 1-4 Front view of a Power E950 server

Figure 1-5 shows the rear view of the Power E950 server. The leftmost slot (C1) is the Flexible Service Processor (FSP), and then there are five PCIe slots. On the right side are six more adapter slots, and on the far right is the tunnel for SAS cables. In this system, there is a SAS adapter in slot C12 with cables routing to the internal disks.

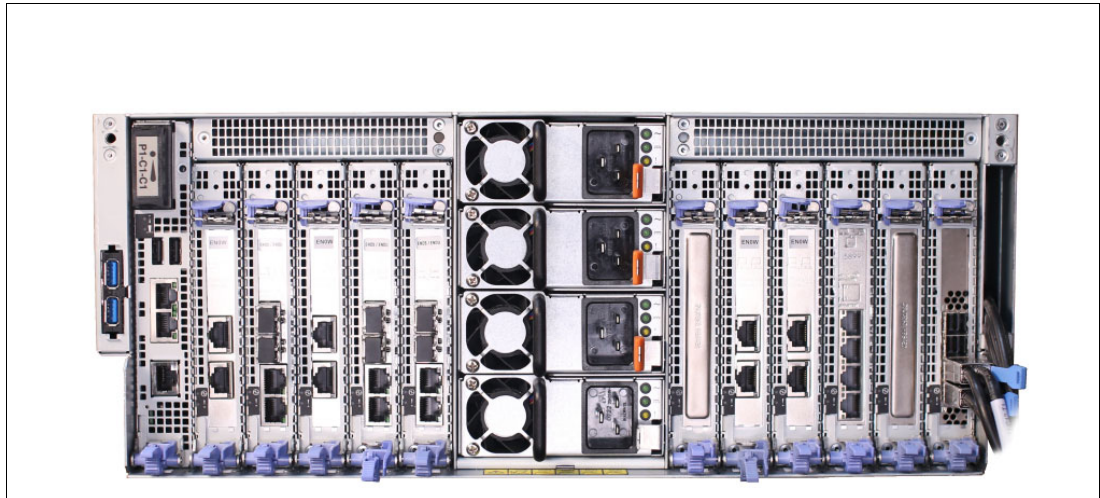


Figure 1-5 Rear view of the Power E950 server

1.2 Operating environment

Table 1-3 details the operating environment for the Power E950 server.

Table 1-3 Operating environment for the Power E950 server

Power E950 operating environment		
System	Power E950 server	
Item	Operating	Non-operating
Temperature	Recommended: 18 - 27 °C (64.4 - 80.6 °F)	5 - 45°C (41.0 - 113.0 °F)
	Allowable 10 - 35 °C (50.0 - 95.0 °F)	
Relative humidity	8 - 80%	8 - 85%
Maximum dew point	21 °C (69.8 °F)	27 °C (80.6 °F)
Operating voltage	200 - 240 V AC	N/A
Operating frequency	50 - 60 Hz +/- 3 Hz AC	N/A
Maximum power consumption	4,220 W maximum	N/A
Maximum power source loading	4.3 kVA maximum	N/A
Maximum thermal output	14,403 BTU/hour	N/A

Power E950 operating environment		
System	Power E950 server	
Maximum altitude	3,050 m (10,000 ft.)	N/A
Maximum noise level	8.8 bels LwAm (heavy workload on one maximally configured 4-socket enclosure, 25°C, 500 m)	

Environmental assessment: The [IBM Systems Energy Estimator tool](#) can provide more accurate information about the power consumption and thermal output of systems based on a specific configuration, including adapters and I/O expansion drawers.

The Power E950 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.

Government regulations, such as those prescribed by OSHA or European Community Directives, may govern noise level exposure in the workplace and might apply to you and your server installation. This IBM system is available with an optional acoustical door feature that can help reduce the noise that is emitted from this system.

The actual sound pressure levels in your installation depend upon various factors, including the number of racks in the installation, the size, materials, and configuration of the room where you install the racks, the noise levels from other equipment, the ambient room temperature, and employees' location in relation to the equipment.

Compliance with government regulations also depends on various other factors, including the duration of employees' exposure and whether employees wear hearing protection. IBM suggests that you consult with qualified experts in this field to determine whether you are in compliance with the applicable regulations.

1.3 Physical package

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

Table 1-4 lists the physical dimensions of the system node and the PCIe Gen3 I/O Expansion Drawer.

Table 1-4 Physical dimensions of the system node and the PCIe Gen3 I/O Expansion Drawer

Dimension	Power E950 system node	PCIe I/O expansion drawer
Width	448 mm (17.5 in.)	447.3 mm (17.61 in.)
Depth	902 mm (35.5 in.)	737 mm (29.0 in.)
Height	175 mm (6.9 in.) four EIA units	173 mm (6.8 in.) four EIA units
Weight	69 kg (152 lb)	54.4 kg (120 lb)

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

1.4 System features

The following conventions are used in the Order type column in all tables in this section.

Initial	Only available when ordered as part of a new system
MES	Only available as a Miscellaneous Equipment Specification (MES) upgrade
Both	Available with a new system or as part of an upgrade
Supported	Unavailable as a new purchase, but supported when migrated from another system or as part of a model conversion

You can order system features during the initial system order. You can also add or replace features later on as a MES. An MES is a hardware change that involves adding, removing, or changing features.

The following features are available on the Power E950 server:

- ▶ 16 - 48 processor cores with two or four POWER9 processor-based modules:
 - 8-core processor modules that provide up to 32 cores (#EPWR0).
 - 10-core processor modules that provide up to 40 cores (#EPWS).
 - 11-core processor modules that provide up to 44 cores (#EPWY).
 - 12-core processor modules that provide up to 48 cores (#EPWT).
- ▶ 128 GB - 16 TB high-performance DDR4 memory with L4 cache:
 - 8 GB DIMM Memory (#EM6A).
 - 16 GB DIMM Memory (#EM6B).
 - 32 GB DIMM Memory (#EM6C).
 - 64 GB DIMM Memory (#EM6D).
 - 128 GB DIMM Memory (#EM6E).
 - Optional Active Memory Expansion (AME) (#EMAM).
- ▶ Choice of three storage backplane features with different SAS RAID controller options. All backplane options have four NVMe SSD bays.
 - Zero DASD Backplane (#EJ0B).
 - Base DASD Backplane (single SAS controller) must be in slot C12 (#EJBB).
 - Split Disk Backplane (two SAS controllers) must be in slots C9 and C12 (#EJSB).
- ▶ Up to 11 hot-swap PCIe slots in the system unit:
 - One PCIe Gen3 8-lane slot is the default for the base Ethernet adapter.
 - Two PCIe Gen4 8-lane full-height, half-length slots.
 - Four, six, or eight PCIe Gen4 16-lane full-height, half-length slots.
 - With two processor modules, there are seven PCIe slots. With three modules, there are nine PCIe slots. With four modules, there are 11 PCIe slots in the system unit.
- ▶ The PCIe I/O Expansion Drawer (#EMX0) expands the number of full-high, hot-swap slots:
 - Up to two PCIe drawers with two processor modules (maximum of 27 slots on the server).
 - Up to four PCIe drawers with four processor modules (maximum of 55 slots on the server).

- ▶ Up to 64 EXP24SX SFF Drawers (#ESLS) can be attached, providing up to 1,536 SAS bays for disks or SSDs.
- ▶ System unit I/O (integrated I/O):
 - USB ports: Four 3.0 (two front) for general use and 2.0 (rear) for limited use.
 - HMC ports: Two 1 GbE RJ45 port.
 - System (serial) port: One RJ45 port.
- ▶ Four hot-plug and redundant power supplies 2000 W (200 - 240 V AC) (#EB3M).
- ▶ System unit only 4U in a 19-inch rack-mount hardware.
- ▶ Primary operating systems:
 - AIX (#2146).
 - Linux (#2147): Red Hat Enterprise Linux, SUSE Linux Enterprise Server, and Ubuntu.

Three processor modules: The Power E950 server supports the installation of two, three, or four processors. At initial availability, only two and four processor configurations are supported.

1.4.1 Minimum configuration

The minimum configuration is the 2-socket, 8-core processor, 128 GB of memory, four power supplies with four power cords, an operating system indicator, a language group specify, a disk backplane, and an internal disk or SAN load source.

Table 1-5 shows the minimum configuration of the Power E950 server.

Table 1-5 Minimum configuration of the Power E950 server

Components	Feature code	Number	Comments
Processor with eight cores		2	
Memory riser card	EM03	2	One riser card per processor socket.
DIMM (Minimum size is 8 GB.)	EM6A	16 x 8 GB = 128 GB	Eight DIMMs per riser card.
Direct access storage device (DASD) backplane with one 800 GB NVMe drive with no SAS drive	EJ0B and EC5J	1	
Or			
DASD backplane with one 300 GB hard disk drive (HDD) and no NVMe drive	EJBB and ESNK	1	Requires EJ0K in slot C12.
PCIe Quad Ethernet 1Gb adapter	EN0S	1	
Power supply		4	

Components	Feature code	Number	Comments
Power cords		4	
Language group specify			
AIX or Linux			

1.4.2 Power supply features

Here are the key power supply features:

- ▶ AC power supply - 2000 W for Server (200-240V AC) (#EB3M):
 - Four power supplies are always required. The power supplies are N+1 redundancy and have dual AC power cord tolerance.
 - If there is a power supply failure, the failed power supply can be exchanged without interrupting the operation of the system.
 - Requires a server line core C19/C20 AC connector to power distribution unit (PDU) and PDU to AC source connector (#4558).
- ▶ AC power supply Conduit for optional PCIe3 Expansion Drawer (#EMXA): Provides two 320-C14 inlet electrical connections for two separately ordered AC power cords with C13 connector plugs. The conduit provides electrical power connection between two power supplies that are at the front of a PCIe Gen3 I/O Expansion Drawer (#EMX0) and two power cords that connect at the rear of the PCIe Gen3 I/O Expansion Drawer.
- ▶ Specify AC power supply for optional EXP12SX/EXP24SX Storage Enclosure (#ESLA): The power supply has a 320-C14 inlet electrical connection for a separately ordered power cord.

1.4.3 Processor module features

The Power E950 server has the following processor module features:

- ▶ The Power E950 server supports 16 - 48 processor cores:
 - 8-core typical 3.6 - 3.8 GHz (maximum) POWER9 Processor (#EPWR)
 - 10-core typical 3.4 - 3.8 GHz (maximum) POWER9 Processor (#EPWS)
 - 11-core typical 3.2 - 3.8 GHz (maximum) POWER9 Processor (#EPWY)
 - 12-core typical 3.15 - 3.8 GHz (maximum) POWER9 Processor (#EPWT)
- ▶ A minimum of two and a maximum of four processor modules are required for each system. If fewer than the maximum number of modules are initially installed, more modules can be added to the system later through an MES order, which requires a scheduled downtime for the installation process.
- ▶ All processor modules in one server must be the same processor module feature. They cannot be mixed.
- ▶ Permanent CoD processor core activations are required for the first processor module in the configuration and are optional for the second, third, and fourth modules.

Table 1-6 details the minimum processor activation per system for the available processor options.

Table 1-6 Minimum permanent activations per system

Processor module	Sockets populated	Minimum permanent activations for system
8 cores (EPWR)	2 or 4	8
10 cores (EPWS)	2 or 4	10
11 cores (EPWY)	2 or 4	11
12 cores (EPWT)	2 or 4	12

- ▶ Elastic CoD (Temporary) 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 #EPWR (#EPN0, #EPN1).
 - 1 and 100 Processor Day Elastic CoD billing for #EPWS (#EPN5, #EPN6).
 - 1 and 100 Processor Day Elastic CoD billing for #EPWY (#EPN8, #EPN9).
 - 1 and 100 Processor Day Elastic CoD billing for #EPWT (#EPNK, #EPNL).
 - 100 Processor-minutes Utility CoD billing: for #EPWR (#EPN2), for #EPWS (#EPN7), for #EPWY (#EPNN), or for #EPWT (#EPNM).
 - An HMC is required for Elastic CoD (Temporary).

1.4.4 Memory features

Here are the memory features for the Power E950 server:

- ▶ 128 GB - 16 TB high-performance 1600 MHz DDR4 error correction code (ECC) memory with L4 cache:
 - 8 GB IS DIMM Memory (#EM6A).
 - 16 GB IS DIMM Memory (#EM6B).
 - 32 GB IS DIMM Memory (#EM6C).
 - 64 GB IS DIMM Memory (#EM6D).
 - 128 GB IS DIMM Memory (#EM6E).
- ▶ As the client memory requirements increase, the system capabilities are increased as follows:
 - With two processor modules that are installed, 64 IS DIMM slots are available. The minimum memory is 128 GB that is composed of sixteen 8 GB IS DIMMs.
 - With four processor modules, 128 IS DIMM slots are available. The minimum memory is 256 GB that is composed of thirty-two 8 GB IS DIMMs.
 - The more IS DIMM slots that are filled, the larger the available bandwidth to the server.

- ▶ Elastic CoD (Temporary) for memory is available for memory capacity that is not permanently activated:
 - 90 Days Elastic CoD Memory Enablement (#EM9U).
 - 8 GB-Day billing for Elastic CoD memory (#EMJE).
 - 800 GB-Day billing for Elastic CoD memory (#EMJF).
 - An HMC is required.
- ▶ Elastic CoD (Temporary) Memory Days can also be acquired through IBM Marketplace after system installation. For more information about new Elastic CoD features, see the [IBM Digital MarketPlace](#).

Memory subsystem highlights

Here are the memory subsystem highlights for the Power E950 server:

- ▶ Up to 16 TB of memory.
- ▶ Up to eight riser cards, each with 16 DIMM slots and four memory buffer chips.
- ▶ Each processor socket controls two memory riser cards.
- ▶ Up to 128 DIMMs slots. Memory is installed in increments of 8 DIMMs.
- ▶ Memory buffer on each riser card to improve performance.

Figure 1-6 shows the memory access path.

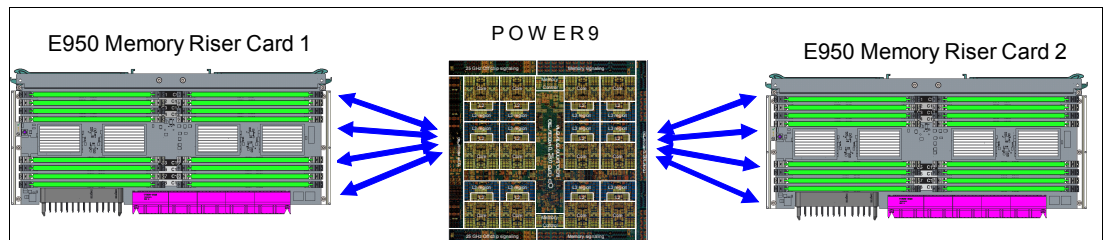


Figure 1-6 Memory access path

Figure 1-7 shows the memory DIMMs and buffer chips on a memory riser card.

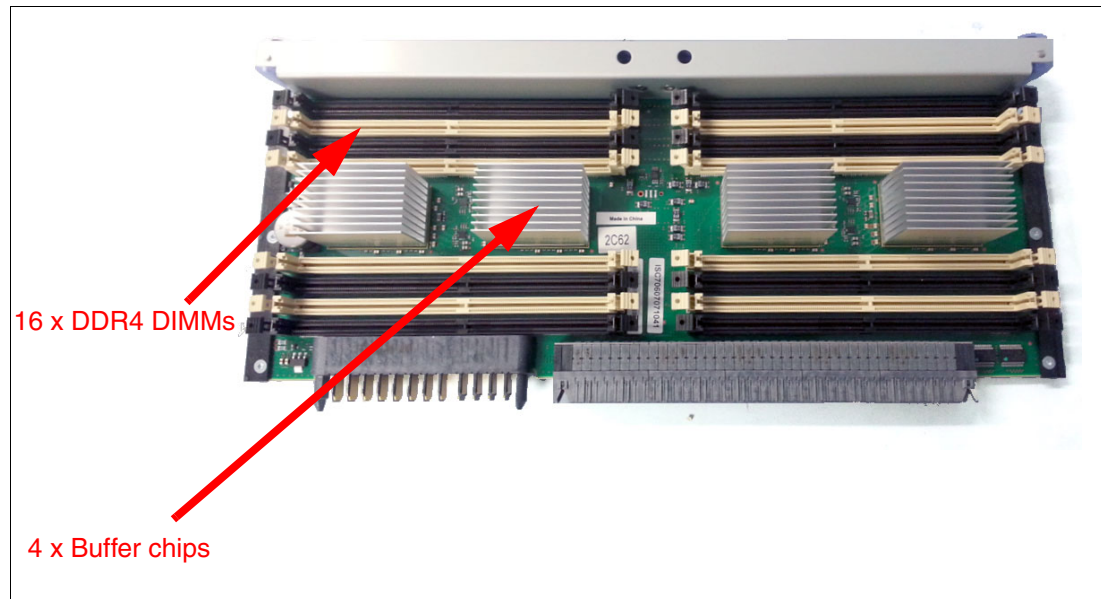


Figure 1-7 Memory riser card

Memory rules:

- ▶ Memory is ordered in group of eight IS DIMMs of the same memory feature codes (FCs). Each is made up of four identical IS DIMMs.
- ▶ All IS DIMMs on one riser card must be the same FC.
- ▶ The two riser cards of each processor socket can have different sizes, but it is preferable that the two riser cards for each processor are the same size.
- ▶ For optimal performance, the amount of memory per processor module should be the same.
- ▶ A minimum of one memory riser with eight DIMMs for each installed POWER9 processor is required.
- ▶ The minimum memory that is supported by two POWER9 processors that are installed is 128 GB.
- ▶ The minimum memory that is supported by four POWER9 processors that are installed is 256 GB.
- ▶ Permanent CoD memory activations are required for at least 50% of the physically installed memory or 128 GB of activations, whichever is larger. Use 1 GB activation (#EMAP) and 100 GB activation (#EMAQ) FCs to order permanent memory activations.

1.4.5 PCIe slots

The Power E950 server has up to 11 PCIe hot-plug slots, which provide excellent configuration flexibility and expandability. Eight adapter slots are PCIe Gen4 16-lane, two adapter slots are PCIe Gen4 8-lane, and one adapter slot is PCIe Gen3 8-lane (the default for the base Ethernet adapter). All adapter slots are full-height, half-length. BSCs are used for adapters in the system unit so that adapters can be installed, removed, and serviced from the rear of the system. All PCIe slots in the system unit are Single Root I/O Virtualization (SR-IOV) capable.

The server has a full set of BSCs, even if the BSCs are empty.

The 16-lane slots can provide up to twice the bandwidth of the 8-lane slots because they offer twice as many PCIe lanes. PCIe Gen4 slots can support up to twice the bandwidth of a PCIe Gen3 slot and up to four times the bandwidth of a PCIe Gen2 slot, assuming an equivalent number of PCIe lanes. PCIe Gen1, PCIe Gen2, PCIe Gen3, and PCIe Gen4 adapters can be plugged into a PCIe Gen4 slot, if that adapter is supported. PCIe Gen4 16-lane slots can be used to attach PCIe Gen3 or PCI Gen3 I/O expansion drawers.

Table 1-7 shows the number of slots that is supported by the number of processor modules.

Table 1-7 Available PCIe slots

PCIe slot type	2 sockets that are populated	4 sockets that are populated
X16 Gen4 slots (CAPI-capable)	4	8
X8 Gen4 slots	2	2
X8 Gen3 slots	1	1

Note:

- ▶ The PCIe Gen3 8-lane slot (C6) is the default for the base Ethernet adapter to ensure proper manufacture and test of the server.
- ▶ This server has an energy-efficient design for cooling the PCIe adapter environment. The server can sense which IBM PCIe adapters are installed in their PCIe slots. If an adapter is known to require higher levels of cooling, the server automatically speeds up fans to increase airflow across the PCIe adapters.
- ▶ PCIe slots C12 and C9 are used for controlling the internal SAS drive bays, depending on which DASD backplane feature is installed.

1.4.6 USB

The backplane in the Power E950 server provides four system USB 3.0 ports. Two USB ports are available at the front of the system and two at the rear. The two front USB ports are rated at 1.30 A, and the two rear ones are rated at 1 A. The external USB DVD is plugged into the front ports only.

Table 1-8 lists the USB-attached media options that are available for the Power E950 server.

Table 1-8 USB-attached external media

Feature code	CCIN	Description	Maximum	OS support	Order type ^a
EUA5	63BD	Stand-alone USB DVD drive w/cable	1	AIX and Linux	Both
EUA4		RDX USB External Docking Station ^b	6	AIX and Linux	Both

a. For more information about order types, see 1.4.7, "Disk and media features" on page 16.

b. #EUA4 is available to purchase only in the United States.

1.4.7 Disk and media features

The Power E950 server supports up to four internal NVMe SSD drives and up to eight SAS drives. One backplane is required in each Power E950 server. The backplanes that are available on the Power E950 server are shown in Table 1-9.

Table 1-9 Available backplanes

Feature code	CCIN	Description	Drive support	PCIe SAS adapters required
EJOB	2D37	DASD Backplane with no HDD/SDD drive selected. No PCIe SAS adapter is required.	4 NVMe 0 SAS	0
EJBB	2D37	Base DASD backplane together with one SAS PCIe adapter and select SAS drives.	4 NVMe 8 SAS	1 in slot C12
EJSB	2D37	Split DASD backplane together with two SAS PCIe adapters and select SAS drives.	4 NVMe 4 + 4 SAS	1 in slot C12 1 in slot C9

SFF-2 and SFF-3: Internal HDD/SSD bays are SFF-3. HDD/SSD bays in the optional EXP24SX disk expansion drawer are SFF-2. Be careful when ordering to ensure that the correct type of disk feature is ordered for the correct disk bay.

HDDs/SSDs cannot be moved between internal disk slots and external disk slots.

Each NVMe device is a separate PCIe endpoint, which means that each NVMe device can be assigned to a different logical partition (LPAR) or Virtual I/O Server (VIOS).

If #EJBB is selected, a SAS adapter feature (#EJ0K) must be installed in slot C12, which drives all eight SAS drives. If #EJSB is selected, two SAS adapter must be installed, one in slot C9 and a second in slot C12. Each adapter drives four SAS drives.

Figure 1-8 shows a picture of two NVMe devices.



Figure 1-8 Two NVMe devices

Table 1-10 lists the available NVMe features for the Power E950 server.

Table 1-10 NVMe features for the Power E950 server

Feature code	CCIN	Description	Maximum	OS support	Order type ^a
EC5J	59B4	Mainstream 800 GB SSD NVMe U.2 module	4	AIX and Linux	Both
EC5K	59B5	Mainstream 1.6 TB SSD NVMe U.2 module	4	AIX and Linux	Both
EC5L	59B6	Mainstream 3.2 TB SSD NVMe U.2 module	4	AIX and Linux	Both

a. For more information about order types, see 1.4.7, “Disk and media features” on page 16.

Table 1-11 lists disk features that are available for the Power E950 server.

Table 1-11 Disk features for the Power E950 server

Feature code	CCIN	Description	Maximum	OS support	Order type ^a
ESF3	59DA	1.2 TB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4096	1536	AIX and Linux	Both
ESF9	59DB	1.2 TB 10K RPM SAS SFF-3 Disk Drive 4K Block - 4096	8	AIX and Linux	Both
ESGP	5B12	1.55 TB Enterprise SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESGR	5B15	1.55 TB Enterprise SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ES8F	5B12	1.55 TB SFF-2 SSD 4K eMLC4 for AIX/Linux	768	AIX and Linux	Supported
ES8V	5B15	1.55 TB SFF-3 SSD 4K eMLC4 for AIX/Linux	8	AIX and Linux	Supported
ESFT	59DD	1.8 TB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4096	1536	AIX and Linux	Both
ESFV	59DE	1.8 TB 10K RPM SAS SFF-3 Disk Drive 4K Block - 4096	8	AIX and Linux	Both
ES96	5B21	1.86 TB Mainstream SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESHL	5B21	1.86 TB Mainstream SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ES92	5B20	1.86 TB Mainstream SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ESHU	5B20	1.86 TB Mainstream SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ESE7	5B2D	3.72 TB Mainstream SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESM8	5B2D	3.72 TB Mainstream SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESE1	5B2C	3.72 TB Mainstream SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both

Feature code	CCIN	Description	Maximum	OS support	Order type ^a
ESMQ	5B2C	3.72 TB Mainstream SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ES62	5B1D	3.86 - 4.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)	768	AIX and Linux	Both
ESHN	5B2F	7.45 TB Mainstream SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESHW	5B2E	7.45 TB Mainstream SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ES64	5B1F	7.72 - 8.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)	768	AIX and Linux	Both
ESEZ	59C9	300 GB 15K RPM SAS SFF-2 4K Block - 4096 Disk Drive	1536	AIX and Linux	Both
ESNM	5B43	300 GB 15K RPM SAS SFF-2 4K Block Cached Disk Drive (AIX/Linux)	1536	AIX and Linux	Both
1953	19B1	300 GB 15k RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX and Linux	Both
ESFB	59E1	300 GB 15K RPM SAS SFF-3 4K Block - 4096 Disk Drive	8	AIX and Linux	Both
ESNK	5B41	300 GB 15K RPM SAS SFF-3 4K Block Cached Disk Drive (AIX/Linux)	8	AIX and Linux	Both
ESDB	59E0	300 GB 15K RPM SAS SFF-3 Disk Drive (AIX/Linux)	8	AIX and Linux	Both
ESGB	5B10	387 GB Enterprise SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESGD	5B13	387 GB Enterprise SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ESG5	5B16	387 GB Enterprise SAS 5xx SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESG9	5B19	387 GB Enterprise SAS 5xx SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ES85	5B10	387 GB SFF-2 SSD 4K eMLC4 for AIX/Linux	768	AIX and Linux	Supported
ES78	5B16	387 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	768	AIX and Linux	Supported
ES8N	5B13	387 GB SFF-3 SSD 4K eMLC4 for AIX/Linux	8	AIX and Linux	Supported
ES7K	5B19	387 GB SFF-3 SSD 5xx eMLC4 for AIX/Linux	8	AIX and Linux	Supported
1964	19B3	600 GB 10k RPM SAS SFF-2 Disk Drive (AIX/Linux)	1536	AIX and Linux	Both
ESEV	59D2	600 GB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4096	1536	AIX and Linux	Both
ESD5	59D0	600 GB 10K RPM SAS SFF-3 Disk Drive (AIX/Linux)	8	AIX and Linux	Both

Feature code	CCIN	Description	Maximum	OS support	Order type ^a
ESF5	59D3	600 GB 10K RPM SAS SFF-3 Disk Drive 4K Block - 4096	8	AIX and Linux	Both
ESFP	59CC	600 GB 15K RPM SAS SFF-2 4K Block - 4096 Disk Drive	1536	AIX and Linux	Both
ESNR	5B47	600 GB 15K RPM SAS SFF-2 4K Block Cached Disk Drive (AIX/Linux)	1536	AIX and Linux	Both
ESFF	59E5	600 GB 15K RPM SAS SFF-3 4K Block - 4096 Disk Drive	8	AIX and Linux	Both
ESNP	5B45	600 GB 15K RPM SAS SFF-3 4K Block Cached Disk Drive (AIX/Linux)	8	AIX and Linux	Both
ESGK	5B11	775 GB Enterprise SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESGM	5B14	775 GB Enterprise SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ESGF	5B17	775 GB Enterprise SAS 5xx SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESGH	5B1A	775 GB Enterprise SAS 5xx SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ES8C	5B11	775 GB SFF-2 SSD 4K eMLC4 for AIX/Linux	768	AIX and Linux	Supported
ES7E	5B17	775 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	768	AIX and Linux	Supported
ES8Q	5B14	775 GB SFF-3 SSD 4K eMLC4 for AIX/Linux	8	AIX and Linux	Supported
ES7P	5B1A	775 GB SFF-3 SSD 5xx eMLC4 for AIX/Linux	8	AIX and Linux	Supported
ES8Y	5B29	931 GB Mainstream SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ESHJ	5B29	931 GB Mainstream SAS 4K SFF-2 SSD for AIX/Linux	768	AIX and Linux	Both
ES83	5B2B	931 GB Mainstream SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both
ESHS	5B2B	931 GB Mainstream SAS 4K SFF-3 SSD for AIX/Linux	8	AIX and Linux	Both

a. For more information about order types, see 1.4.7, "Disk and media features" on page 16.

Table 1-12 shows disk and SSD features that are available for bulk ordering.

Table 1-12 Bulk disk and SSD features

Feature code	CCIN	Description	Maximum	OS support	Order type ^a
1929	19B1	Quantity 150 of 1953	10	AIX and Linux	Both
1818	19B3	Quantity 150 of 1964	10	AIX and Linux	Both
EQ62	5B1D	Quantity 150 of ES62 3.86 - 4.0 TB 7200 rpm 4K LFF-1 Disk	5	AIX and Linux	Both

Feature code	CCIN	Description	Maximum	OS support	Order type ^a
EQ64	5B1F	Quantity 150 of ES64 7.72 - 8.0 TB 7200 rpm 4K LFF-1 Disk	5	AIX and Linux	Both
EQ78	5B16	Quantity 150 of ES78 387 GB SFF-2 SSD 5xx	5	AIX and Linux	Supported
EQ7E	5B17	Quantity 150 of ES7E 775 GB SFF-2 SSD 5xx	5	AIX and Linux	Supported
EQ85	5B10	Quantity 150 of ES85 387 GB SFF-2 SSD 4K	5	AIX and Linux	Supported
EQ8C	5B11	Quantity 150 of ES8C 775 GB SFF-2 SSD 4K	5	AIX and Linux	Supported
EQ8F	5B12	Quantity 150 of ES8F 1.55 TB SFF-2 SSD 4K	5	AIX and Linux	Supported
EQ8Y	5B29	Quantity 150 of ES8Y 931 GB SFF-2 SSD 4K	5	AIX and Linux	Both
EQ96	5B21	Quantity 150 of ES96 1.86 TB SFF-2 SSD 4K	5	AIX and Linux	Both
EQE7	5B2D	Quantity 150 of ESE7 3.72 TB SFF-2 SSD 4K	5	AIX and Linux	Both
EQEV	59D2	Quantity 150 of ESEV (600 GB 10k SFF-2)	10	AIX and Linux	Both
EQEZ	59C9	Quantity 150 of ESEZ (300 GB SFF-2)	10	AIX and Linux	Both
EQF3	59DA	Quantity 150 of ESF3 (1.2 TB 10k SFF-2)	10	AIX and Linux	Both
EQFP	59CC	Quantity 150 of ESFP (600 GB SFF-2)	10	AIX and Linux	Both
EQFT	59DD	Quantity 150 of ESFT (1.8 TB 10k SFF-2)	10	AIX and Linux	Both
EQG5	5B16	Quantity 150 of ESG5 (387 GB SAS 5xx)	5	AIX and Linux	Both
EQGB	5B10	Quantity 150 of ESGB (387 GB SAS 4K)	5	AIX and Linux	Both
EQGF	5B17	Quantity 150 of ESGF (775 GB SAS 5xx)	5	AIX and Linux	Both
EQGK	5B11	Quantity 150 of ESGK (775 GB SAS 4K)	5	AIX and Linux	Both
EQGP	5B12	Quantity 150 of ESGP (1.55 TB SAS 4K)	5	AIX and Linux	Both
ERHJ	5B29	Quantity 150 of ESHJ 931 GB SSD 4K SFF-2	5	AIX and Linux	Both
ERHL	5B21	Quantity 150 of ESHL 1.86 TB SSD 4K SFF-2	5	AIX and Linux	Both
ERHN	5B2F	Quantity 150 of ESHN 7.45 TB SSD 4K SFF-2	5	AIX and Linux	Both
ERM8	5B2D	Quantity 150 of ESM8 3.72 TB SSD 4K SFF-2	5	AIX and Linux	Both
ESPM	5B43	Quantity 150 of ESNM (300 GB 15k SFF-2)	10	AIX and Linux	Both
ESPR	5B47	Quantity 150 of ESNR (600 GB 15k SFF-2)	10	AIX and Linux	Both

a. For more information about order types, see 1.4.7, "Disk and media features" on page 16.

1.5 I/O drawers

If more PCIe slots beyond the system node slots are required, the Power E950 server supports the addition of I/O expansion drawers. At initial availability in September 2018, the Power E950 server supports the attachment of zero or two PCIe Gen3 I/O Expansion Drawers to each system. In November 2018, zero, two, or four PCIe Gen3 I/O Expansion Drawers per system will be supported. To connect an I/O expansion drawer, a x16 PCIe slot is used to attach a 6-slot expansion module in the I/O drawer. A PCIe Gen3 I/O Expansion

Drawer (#EMX0) holds two expansion modules, which are attached to any two x16 PCIe slots.

To connect SAS disks, disk-only I/O drawers are available. The EXP24S, EXP24SX, and EXP12SX are supported. The EXP24S is support-only (at a time after the initial GA) and cannot be ordered together with #E950.

1.5.1 PCIe Gen3 I/O Expansion Drawer

The 19-inch 4 EIA (4U) PCIe Gen3 I/O Expansion Drawer (#EMX0) and two PCIe Fan Out Modules (#EMXG) provide 12 PCIe I/O full-length, full-height slots. One fan-out module provides six PCIe slots (C1 - C6). C1 and C4 are x16 slots, and C2, C3, C5, and C6 are x8 slots. PCIe Gen1, Gen2, and Gen3 full-high adapters are supported.

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

The drawer has a full set of BSCs, even if the BSCs are empty.

Note: The BSC that is used in the I/O expansion drawer is a different one than the one that used in the server.

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

A PCIe CXP converter adapter and Active Optical Cables (AOCs) connect the system node to a PCIe fan-out module in the I/O expansion drawer. Each PCIe Gen3 I/O Expansion Drawer has two power supplies.

If CXP copper cable is used, A T42 rack requires the 8-inch rack extension to close the rear door.

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

Figure 1-9 shows a PCIe Gen3 I/O Expansion Drawer.



Figure 1-9 PCIe Gen3 I/O Expansion Drawer

1.5.2 I/O drawers and usable PCI slots

Figure 1-10 shows the rear view of the PCIe Gen3 I/O Expansion Drawer with the location codes for the PCIe adapter slots in the PCIe3 6-slot fan-out module.

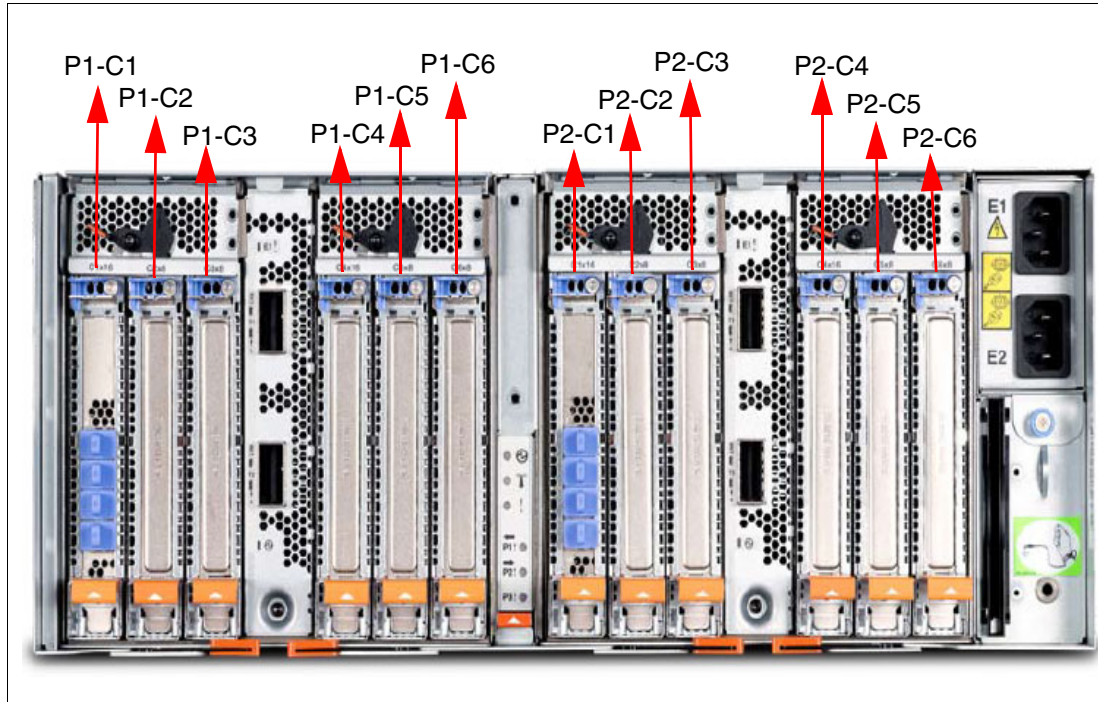


Figure 1-10 Rear view of a PCIe Gen3 I/O Expansion Drawer with PCIe slots location codes

Table 1-13 provides details about the PCI slots in the PCIe Gen3 I/O Expansion Drawer.

Table 1-13 PCIe slot locations and descriptions for the PCIe Gen3 I/O Expansion Drawer

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

In Table 1-13 on page 22:

- ▶ All slots support full-length, regular-height adapter or short (low-profile) adapters with a regular-height tailstock in single-wide, Gen3 BSCs.
- ▶ Slots C1 and C4 in each PCIe3 6-slot fan-out module are x16 PCIe3 buses, and slots C2, C3, C5, and C6 are x8 PCIe buses.
- ▶ All slots support enhanced error handling (EEH).
- ▶ All PCIe slots are hot swappable and support concurrent maintenance.

1.5.3 EXP24S SFF Gen2-bay Drawer

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

Note: The EXP24S is support-only (after the initial GA) and cannot be ordered together with #E950.

With AIX and 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 and cannot be changed in the field.

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

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

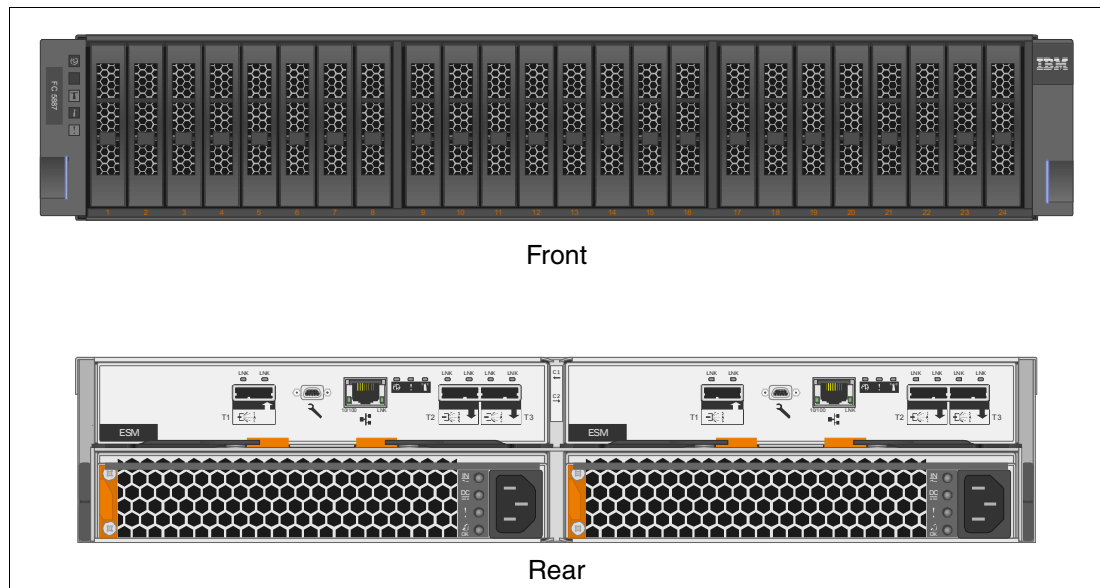


Figure 1-11 EXP24S SFF Gen2-bay drawer

1.5.4 EXP24SX and EXP12SX SAS Storage Enclosures

If you need more disks than are available with the internal disk bays, you can attach more external disk subsystems, such as an EXP24SX SAS Storage Enclosure (#ESLS) or EXP12SX SAS Storage Enclosure (#ESLL).

The EXP24SX is a storage expansion enclosure with twenty-four 2.5-inch SFF SAS bays. It supports up to 24 hot-plug HDDs or SSDs in only 2 EIA of space in a 19-inch rack. The EXP24SX SFF bays use SFF Gen2 (SFF-2) carriers or trays.

The EXP12SX is a storage expansion enclosure with twelve 3.5-inch large form factor (LFF) SAS bays. It supports up to 12 hot-plug HDDs in only 2 EIA of space in a 19-inch rack. The EXP12SX SFF bays use LFF Gen1 (LFF-1) carriers/trays. The 4 KB sector drives (#4096 or #4224) are supported. SSDs are not supported.

With AIX and Linux, and VIOS, the EXP24SX or the EXP12SX can be ordered with four sets of six bays (mode 4), two sets of 12 bays (mode 2), or one set of 24-four bays (mode 1).

Important: When changing modes, a skilled, technically qualified person should follow the special documented procedures. Improperly changing modes can potentially destroy existing RAID sets, prevent access to existing data, or allow other partitions to access another partition's existing data.

The attachment between the EXP24SX or EXP12SX and the PCIe3 SAS adapters or integrated SAS controllers is through SAS YO12 or X12 cables. All ends of the YO12 and X12 cables have mini-SAS HD narrow connectors. The PCIe Gen3 SAS adapters support 6 Gb throughput. The EXP24SX and EXP12SX can support up to 12 Gb throughput if future SAS adapters support that capability.

The EXP24SX or EXP12SX includes redundant AC power supplies and two power cords.

Figure 1-12 shows the EXP24SX drawer.



Figure 1-12 The EXP24SX drawer

Figure 1-13 shows the EXP12SX drawer.



Figure 1-13 The EXP12SX drawer

Note:

- ▶ For the EXP24SX, A maximum of twenty-four 2.5-inch SSDs or 2.5-inch HDDs are supported in the #ESLS 24 SAS bays. There can be no mixing of HDDs and SSDs in the same mode-1 drawer. HDDs and SSDs can be mixed in a mode-2 or mode-4 drawer, but they cannot be mixed within a logical split of the drawer. For example, in a mode-2 drawer with two sets of 12 bays, one set can hold SSDs and one set can hold HDDs, but you cannot mix SSDs and HDDs in the same set of 12-bays.
- ▶ The EXP24S, EXP24SX, and EXP12SX drawers can be mixed on the same server and on the same PCIe3 adapters.
- ▶ The EXP12SX does not support SSD.

1.6 System racks

The Power E950 server is designed to fit a standard 19 inch rack. The server has been certified and tested in IBM Enterprise racks (7965-S42, 7014-T42, #0553, or #0551). Clients can choose to place the server in other racks if they are confident that those racks have the strength, rigidity, depth, and hole pattern characteristics that are needed. Clients should work with IBM Service to determine the appropriateness of other racks. The Power E950 rails can adjust their depth to fit a rack that is 57.8 - 77.5 cm (22.75 - 30.5 inches) in depth.

Order information: It is a preferred practice that the Power E950 server be ordered with an IBM 42U Enterprise rack (7014-T42 or 7965-S42) to provide a complete and higher-quality environment for IBM Manufacturing system assembly and testing, and provide a complete package.

If a system will be installed in a rack or cabinet that is not from IBM, ensure that the rack meets the requirements that are described in 1.6.10, “Original equipment manufacturer racks” on page 35.

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

1.6.1 New racks

The following new racks are offered by IBM:

- ▶ The new IBM Enterprise 42U Slim Rack 7965-S42 (or #ER05) is an available offering of 42 EIA units (U) of space in a slim footprint.
- ▶ The T00 rack is no longer available to purchase with a Power E950 server. Installing a Power E950 server in an existing T00 rack is still supported.

1.6.2 IBM Enterprise 42U Slim Rack 7965-S42

The 2.0-meter (79-inch) Model 7965-S42 is compatible with past and present 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. It can be easily placed on standard 24-inch floor tiles.

The S42 is the default rack in eConfig.

Compared to the 7965-94Y Slim Rack, the Enterprise Slim Rack provides more strength and shipping/installation flexibility.

The 7965-S42 rack has space for up to four PDUs in side pockets. More PDUs beyond four are mounted horizontally and use 1U of rack space.

The Enterprise Slim Rack front door, which can be Basic Black/Flat (#ECRM), High-End appearance (#ECRF), or original equipment manufacturer (OEM) black (#ECRE), has perforated steel, providing ventilation, physical security, and visibility of indicator lights in the installed equipment within. It comes standard with a lock that is identical to the locks in the rear doors. The door (#ECRG and #ECRE only) can be hinged on either the left or right side.

Orientation: #ECRF should not be flipped because the IBM logo will be upside down.

1.6.3 IBM 7014 Model T42 rack

The 2.0-meter (79.3-inch) Model T42 is compatible with past and present Power Systems servers. The features of the T42 rack are as follows:

- ▶ Has 42U (EIA units) of usable space.
- ▶ Has optional removable side panels.
- ▶ Has optional side-to-side mounting hardware for joining multiple racks.
- ▶ Has increased power distribution and weight capacity.
- ▶ Supports both AC power only.
- ▶ Up to four PDUs can be mounted in the PDU bays (see Figure 1-15 on page 30), but others can fit inside the rack. For more information, see 1.6.7, “The AC power distribution unit and rack content” on page 29.

- ▶ **Ruggedized Rack Feature**

For enhanced rigidity and stability of the rack, the optional Ruggedized Rack Feature (#6080) provides more hardware that reinforces the rack and anchors it to the floor. This hardware is designed primarily for use in locations where earthquakes are a concern. The feature includes a large steel brace or truss that bolts into the rear of the rack.

It is hinged on the left side so it can swing out of the way for easy access to the rack drawers when necessary. The Ruggedized Rack Feature also includes hardware for bolting the rack to a concrete floor or a similar surface, and bolt-in steel filler panels for any unoccupied spaces in the rack.

- ▶ **Weights are as follows:**

- T42 base empty rack: 261 kg (575 lb)
- T42 full rack: 930 kg (2045 lb)

Some of the available door options for the T42 rack are shown in Figure 1-14.



Figure 1-14 Door options for the T42 rack

The door options are explained in the following list:

- ▶ The 2.0 m Rack Trim Kit (#6272) is used if no front door is used in the rack.
- ▶ The Front Door for a 2.0 m Rack (#6069) is made of 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. This door is non-acoustic and has a depth of about 25 mm (1 in.).
- ▶ The 2.0 m Rack Acoustic Door (#6249) consists of a front door and a rear door to reduce noise by approximately 6 dB(A). It has a depth of approximately 191 mm (7.5 in.).

- ▶ The #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 non-acoustic door has a depth of about 134 mm (5.3 in.).
- ▶ If CXP copper cable is used for the I/O drawers, A T42 rack requires the 8-inch rack extension so that you can close the rear door.

Rear Door Heat Exchanger

To lead away more heat, a special door that is named the Rear Door Heat Exchanger (#6858) is available. This door replaces the standard rear door on the rack. Copper tubes that are attached to the rear door circulate chilled water that is provided by the customer. The chilled water removes heat from the exhaust air being blown through the servers and attachments that are mounted in the rack. With industry-standard quick couplings, the water lines in the door attach to the customer-supplied secondary water loop.

For more information about planning for the installation of the IBM Rear Door Heat Exchanger, see [IBM Knowledge Center](#).

1.6.4 1.8 Meter Rack (#0551)

The 1.8 Meter Rack (#0551) is a 36 EIA unit rack. The rack that is delivered as #0551 is the same rack that is delivered when you order the 7014-T00 rack. The included features might vary. Certain features that are delivered as part of the 7014-T00 must be ordered separately with #0551.

Order availability: The #0551 rack is available only when ordered as an MES order. It is not available as an initial order.

1.6.5 2.0 Meter Rack (#0553)

The 2.0 Meter Rack (#0553) is a 42 EIA unit rack. The rack that is delivered as #0553 is the same rack that is delivered when you order the 7014-T42 rack. The included features might vary. Certain features that are delivered as part of the 7014-T42 must be ordered separately with #0553.

Order availability: The #0553 rack is available only when ordered as an MES order. It is not available as an initial order.

1.6.6 Rack (#ER05)

The #ER05 provides a 19-inch, 2.0-meter high rack with 42 EIA units of total space for installing a rack-mounted central electronics complex or expansion units. The 600 mm wide rack fits within a data center's 24-inch floor tiles and provides better thermal and cable management capabilities. The following features are required on #ER05:

- ▶ Front door (#EC01)
- ▶ Rear door (#EC02) or Rear Door Heat Exchanger (RDHX) indicator (#EC05)

PDUs on the rack are optional. Each #7196 and #7189 PDU uses one of six vertical mounting bays. Each PDU beyond four uses 1U of rack space.

If ordering Power Systems equipment in an MES order, use the equivalent rack (#ER05) instead of 7965-94Y so IBM Manufacturing can include the hardware in the rack.

1.6.7 The AC power distribution unit and rack content

Using previously provided IBM PDU features #7188, #7109, and #7196 reduces the number of Power E950 servers and other equipment that can be held most efficiently in a rack. The high-function PDUs provide more electrical power per PDU and offer better “PDU footprint” efficiency. In addition, they are intelligent PDUs that provide insight to actual power usage by receptacle and also provide remote power on/off capability for easier support by individual receptacle. The new PDUs are #EPTJ, #EPTL, #EPTN, and #EPTQ.

IBM Manufacturing integrates only the newer PDUs with the Power E950 server. IBM Manufacturing does not support integrating earlier PDUs, such as #7188, #7109, or #7196. Clients can choose to use older IBM PDUs in their racks, but must install those earlier PDUs at their site.

PDU FCs are shown in Table 1-14.

Table 1-14 PDUs available

PDUs	1-phase or 3-phase depending on country wiring standards	3-phase 208 V depending on country wiring standards
Nine C19 receptacles ^a	EPTJ	EPTL
Twelve C13 receptacles	EPTN	EPTQ

a. The Power E950 server has an AC power supply with a C19/C20 connector.

Power sockets: The Power E950 server takes IEC 60320 C19/C20 mains power and not C13. Ensure that the correct power cords and PDUs are ordered or available in the rack.

Four PDUs can be mounted vertically in the back of the S42, T00, and T42 racks. Figure 1-15 shows the placement of the four vertically mounted PDUs. In the rear of the rack, two more PDUs can be installed horizontally in the T00 rack and three in the S42 and T42 rack. The four vertical mounting locations are filled first in the T00, S42, and T42 racks.

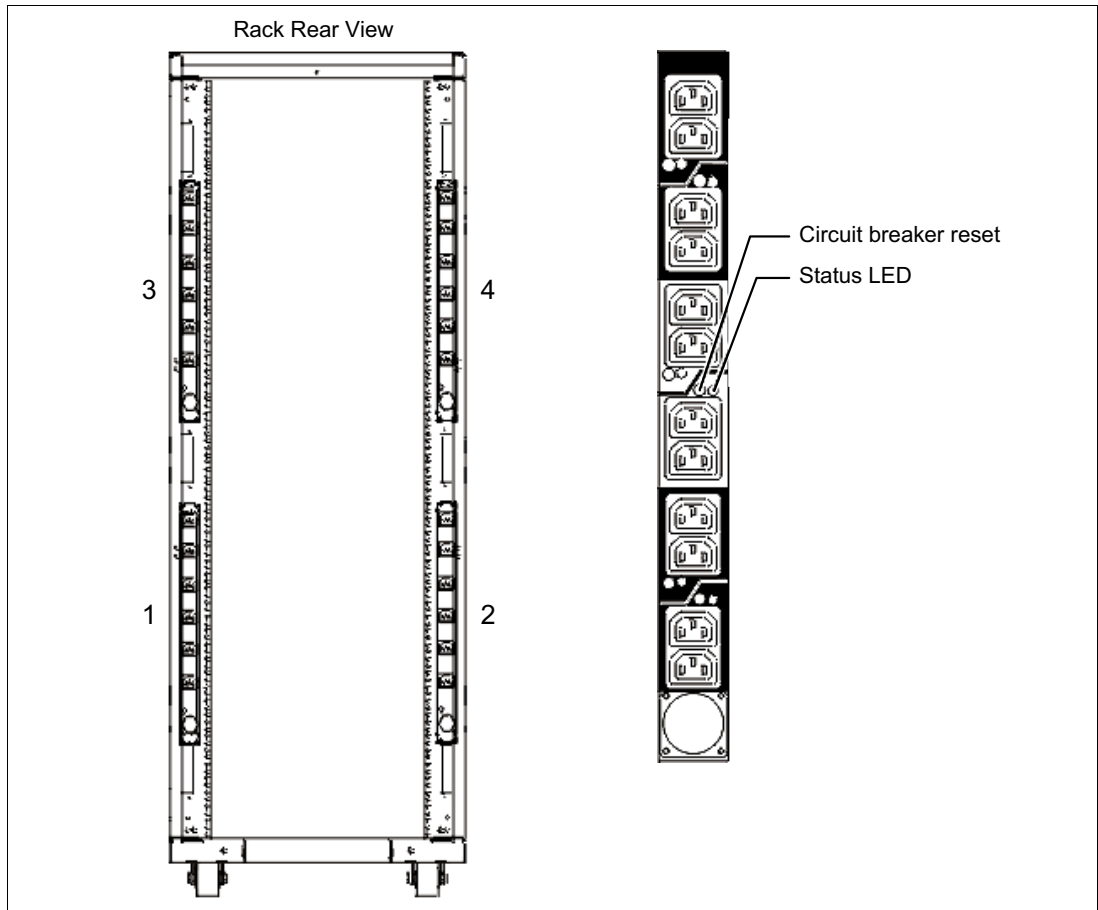


Figure 1-15 Power distribution unit placement and view

Mounting PDUs horizontally uses 1U per PDU and reduces the space that is available for other racked components. When mounting PDUs horizontally, the preferred approach is to use fillers in the EIA units that are occupied by these PDUs to facilitate proper air-flow and ventilation in the rack.

The PDU receives power through a UTG0247 power-line connector. Each PDU requires one PDU-to-wall power cord. Various power cord features are available for various countries and applications by varying the PDU-to-wall power cord, which must be ordered separately. Each power cord provides the unique design characteristics for the specific power requirements. To match new power requirements and save previous investments, these power cords can be requested with an initial order of the rack or with a later upgrade of the rack features.

Table 1-15 shows the available wall power cord options for the PDU and intelligent power distribution unit (iPDU) features, which must be ordered separately.

Table 1-15 Wall power cord options for the power distribution unit and iPDU features

Feature code	Wall plug	Rated voltage (V AC)	Phase	Rated amperage	Geography
6653	IEC 309, 3P+N+G, 16 A	230	3	16 amps/phase	Internationally available
6489	IEC309 3P+N+G, 32 A	230	3	32 amps/phase	EMEA
6654	NEMA L6-30	200 - 208, 240	1	24 amps	US, Canada, LA, and Japan
6655	RS 3750DP (watertight)	200 - 208, 240	1	24 amps	US, Canada, LA, and Japan
6656	IEC 309, P+N+G, 32 A	230	1	24 amps	EMEA
6657	PDL	230 - 240	1	32 amps	Australia, New Zealand
6658	Korean plug	220	1	30 amps	North and South Korea
6492	IEC 309, 2P+G, 60 A	200 - 208, 240	1	48 amps	US, Canada, LA, and Japan
6491	IEC 309, P+N+G, 63 A	230	1	63 amps	EMEA

Notes: Ensure that the appropriate power cord feature is configured to support the power that is being supplied. Based on the power cord that is used, the PDU can supply 4.8 - 19.2 kVA. The power of all the drawers that are plugged into the PDU must not exceed the power cord limitation.

The Universal PDUs are compatible with previous models.

To better enable electrical redundancy, each server has four power supplies that must be connected to separate PDUs, which are not included in the base order.

For maximum availability, a preferred approach is to connect power cords from the same system to two separate PDUs in the rack, and to connect each PDU to independent power sources.

For detailed power requirements and power cord details about the 7014 racks, see [IBM Knowledge Center](#).

For detailed power requirements and power cord details about the 7965-94Y rack, see [IBM Knowledge Center](#).

1.6.8 Rack-mounting rules

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

- ▶ The system is 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.

Order information: The rack for the initial order must be either a 7014-T42 or 7965-S42. If an extra rack is required for I/O expansion drawers as an MES to an existing system, either an #0551, #0553, or #ER05 rack must be ordered.

If you install the Power E950 server in a T42 rack, 2U of space must be left for cable routing. For the bottom cable exit, the 2U must be left at the bottom of the rack. For the top cable exit, the 2U must be left at the top of the rack.

If you install the Power E950 server in an S42 rack, no space is required for cable routing.

1.6.9 Useful rack additions

This section highlights several rack addition solutions for Power Systems rack-based systems.

IBM System Storage 7226 Model 1U3 Multi-Media Enclosure

The IBM System Storage 7226 Model 1U3 Multi-Media Enclosure can accommodate up to two tape drives, two RDX removable disk drive docking stations, or up to four DVD-RAM drives.

The IBM System Storage 7226 Multi-Media Enclosure supports LTO Ultrium and DAT160 Tape technology, DVD-RAM, and RDX removable storage requirements on the following IBM systems:

- ▶ IBM POWER6® processor-based systems
- ▶ IBM POWER7® processor-based systems
- ▶ IBM POWER8 processor-based systems
- ▶ IBM POWER9 processor-based systems

The IBM System Storage 7226 Multi-Media Enclosure offers an expansive list of drive feature options, as shown in Table 1-16.

Table 1-16 Supported drive features for the 7226-1U3

Feature code	Description	Status
5619	DAT160 SAS Tape Drive	Available
EU16	DAT160 USB Tape Drive	Available
1420	DVD-RAM SAS Optical Drive	Available
1422	DVD-RAM Slim SAS Optical Drive	Available
5762	DVD-RAM USB Optical Drive	Available

Feature code	Description	Status
5763	DVD Front USB Port Sled with DVD-RAM USB Drive	Available
5757	DVD RAM Slim USB Optical Drive	Available
8248	LTO Ultrium 5 Half High Fibre Tape Drive	Available
8241	LTO Ultrium 5 Half High SAS Tape Drive	Available
8348	LTO Ultrium 6 Half High Fibre Tape Drive	Available
8341	LTO Ultrium 6 Half High SAS Tape Drive	Available
EU03	RDX 3.0 Removable Disk Docking Station	Available

Here are the option descriptions:

- ▶ **DAT160 160 GB Tape Drives:** With SAS or USB interface options and a data transfer rate up to 12 MBps (assumes 2:1 compression), the DAT160 drive is read/write compatible with DAT160 and DDS4 data cartridges.
- ▶ **LTO Ultrium 5 Half-High 1.5 TB SAS and Fibre Channel Tape Drive:** With a data transfer rate up to 280 MBps (assuming a 2:1 compression), the LTO Ultrium 5 drive is read/write compatible with LTO Ultrium 5 and 4 data cartridges, and read-only compatible with Ultrium 3 data cartridges. By using data compression, an LTO-5 cartridge can store up to 3 TB of data.
- ▶ **LTO Ultrium 6 Half-High 2.5 TB SAS and Fibre Channel Tape Drive:** With a data transfer rate up to 320 MBps (assuming a 2.5:1 compression), the LTO Ultrium 6 drive is read/write compatible with LTO Ultrium 6 and 5 media, and read-only compatibility with LTO Ultrium 4. By using data compression, an LTO-6 cartridge can store up to 6.25 TB of data.
- ▶ **DVD-RAM:** The 9.4 GB SAS Slim Optical Drive with an SAS and USB interface option is compatible with most standard DVD disks.
- ▶ **RDX removable disk drives:** The RDX USB docking station is compatible with most RDX removable disk drive cartridges when it is used in the same operating system. The 7226 offers the following RDX removable drive capacity options:
 - 500 GB (#1107)
 - 1.0 TB (#EU01)
 - 2.0 TB (#EU2T)

Removable RDX drives are in a rugged cartridge that inserts in to an RDX removable (USB) disk docking station (#1103 or #EU03). RDX drives are compatible with docking stations, which are installed internally in IBM POWER6, IBM POWER6+™, POWER7, IBM POWER7+™, POWER8, and POWER9 processor-based servers, where applicable.

Media that is used in the 7226 DAT160 SAS and USB tape drive features are compatible with DAT160 tape drives that are installed internally in IBM POWER6, POWER6+, POWER7, POWER7+, POWER8, and POWER9 processor-based servers.

Media that is used in LTO Ultrium 5 Half-High 1.5 TB tape drives are compatible with Half High LTO5 tape drives that are installed in the IBM TS2250 and TS2350 external tape drives, IBM LTO5 tape libraries, and half-high LTO5 tape drives that are installed internally in IBM POWER6, POWER6+, POWER7, POWER7+, POWER8, and POWER9 processor-based servers.

Figure 1-16 shows the IBM System Storage 7226 Multi-Media Enclosure.



Figure 1-16 IBM System Storage 7226 Multi-Media Enclosure

The IBM System Storage 7226 Multi-Media Enclosure offers customer-replaceable unit (CRU) maintenance service to help make the installation or replacement of new drives efficient. Other 7226 components are also designed for CRU maintenance.

The IBM System Storage 7226 Multi-Media Enclosure is compatible with most IBM POWER6, POWER6+, POWER7, POWER7+, POWER8, and POWER9 processor-based systems that offer current level AIX and Linux operating systems.

For a complete list of host software versions and release levels that support the IBM System Storage 7226 Multi-Media Enclosure, see [System Storage Interoperation Center \(SSIC\)](#).

Note: Any of the existing 7216-1U2, 7216-1U3, and 7214-1U2 multimedia drawers are also supported.

Flat panel display options

The IBM 7316 Model TF4 is a rack-mountable flat panel console kit that can also be configured with the tray pulled forward and the monitor folded up, providing full viewing and keying capability for the HMC operator.

The Model TF4 is a follow-on product to the Model TF3 and offers the following features:

- ▶ A slim, sleek, and lightweight monitor design that occupies only 1U (1.75 in.) in a 19-inch standard rack.
- ▶ A 18.5-inch (409.8 mm x 230.4 mm) flat panel TFT monitor with truly accurate images and virtually no distortion.
- ▶ The ability to mount the IBM Travel Keyboard in the 7316-TF4 rack keyboard tray.
- ▶ Support for the IBM 1x8 Rack Console Switch (#4283) IBM Keyboard/Video/Mouse (KVM) switches.

#4283 is a 1x8 Console Switch that fits in the 1U space behind the TF4. It is a CAT5-based switch containing eight rack interface (ARI) ports for connecting either PS/2 or USB console switch cables. It supports chaining of servers that use the IBM Conversion Options switch cable (#4269). This feature provides four cables that connect a KVM switch to a system, or can be used in a daisy-chain scenario to connect up to 128 systems to a single KVM switch. It also supports server-side USB attachments.

1.6.10 Original equipment manufacturer racks

The system can be installed in a suitable OEM rack if that the rack conforms to the EIA-310-D standard for 19-inch racks. This standard is published by the Electrical Industries Alliance. For more information, see [IBM Knowledge Center](#).

The website mentions the following key points:

- ▶ The front rack opening must be 451 mm wide ± 0.75 mm (17.75 in. ± 0.03 in.), and the rail-mounting holes must be 465 mm ± 0.8 mm (18.3 in. ± 0.03 in.) apart on-center (that is, the horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges). Figure 1-17 is a top view showing the specification dimensions.

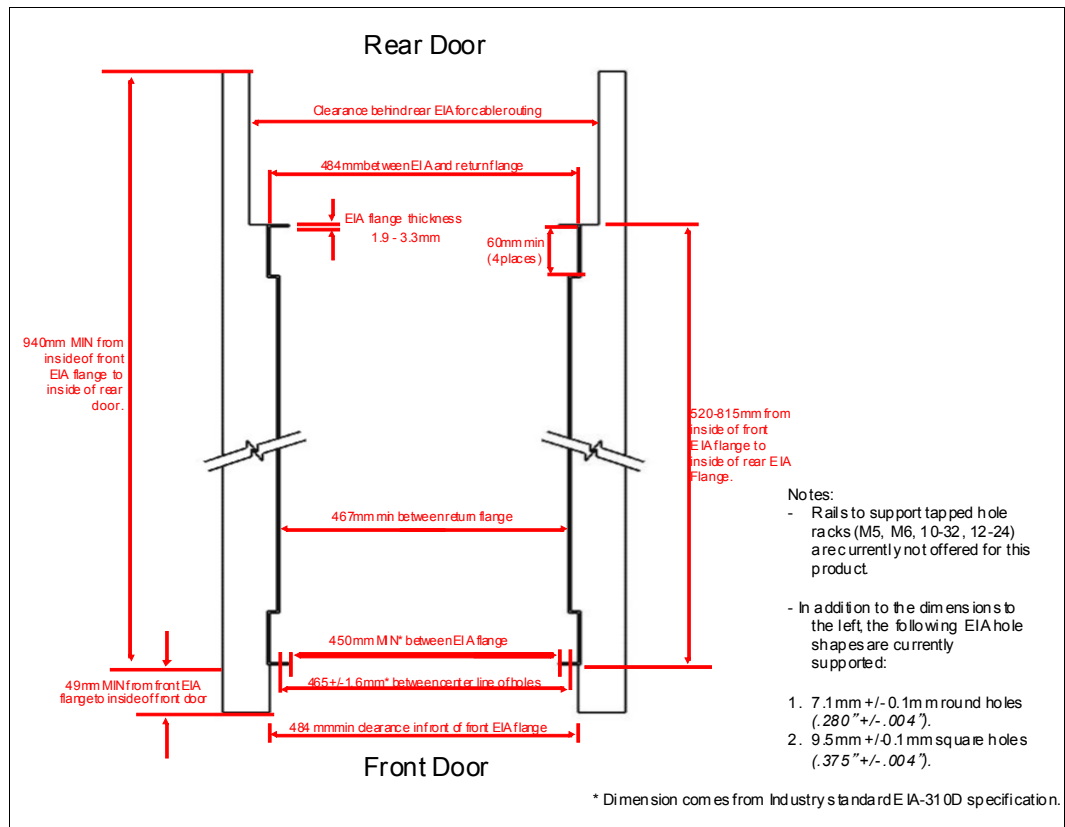


Figure 1-17 Top view of the rack specification dimensions (not specific to IBM)

- ▶ The vertical distance between the mounting holes must consist of sets of three holes spaced (from bottom to top) 15.9 mm (0.625 in.), 15.9 mm (0.625 in.), and 12.67 mm (0.5 in.) on-center, making each three-hole set of vertical hole spacing 44.45 mm (1.75 in.) apart on center. Rail-mounting holes must be 7.1 mm ± 0.1 mm (0.28 in. ± 0.004 in.) in diameter. Figure 1-18 shows the top front specification dimensions.

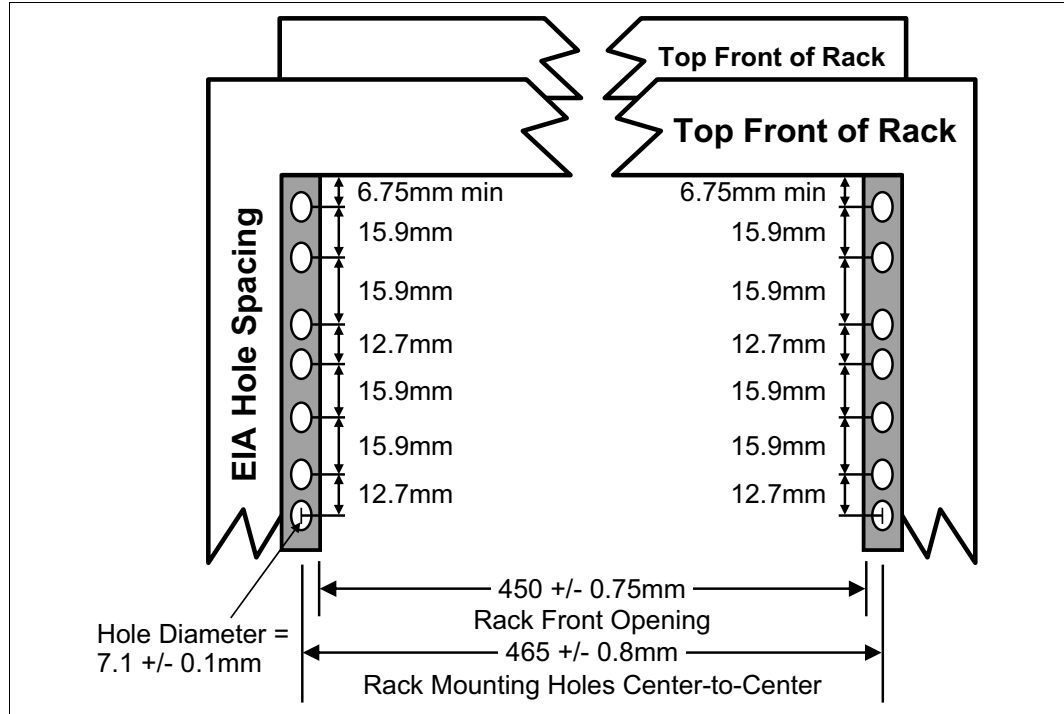


Figure 1-18 Rack specification dimensions: Top front view

1.7 Hardware Management Console

This section describes the HMCs that are available for use with Power Systems servers.

1.7.1 New features of the Hardware Management Console

Here are the new HMC features:

- ▶ New HMCs are now based on systems with IBM POWER® processors.
- ▶ Intel x86-based HMCs are supported but no longer available.
- ▶ Virtual HMCs (vHMC) are available for x86 and Power virtual environments.

1.7.2 Hardware Management Console overview

Administrators can use the HMC, which is a dedicated appliance, to configure and manage system resources on Power Systems servers. GUI, command-line interface (CLI), or REST API interfaces are available. The HMC provides basic virtualization management support for configuring LPARs and dynamic resource allocation, including processor and memory settings for selected Power Systems servers.

The HMC also supports advanced service functions, including guided repair and verification, concurrent firmware updates for managed systems, and around-the-clock error reporting through IBM Electronic Service Agent™ (ESA) for faster support.

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

The HMC is available as a hardware appliance or as a virtual appliance (vHMC). The Power E950 servers support an attachment to one or more HMCs or vHMCs. This is the default configuration for servers supporting multiple LPARs with dedicated resource or virtual I/O.

- ▶ X86 based HMCs: 7042-CR7, CR8, or CR9
- ▶ POWER based HMC: 7063-CR1
- ▶ vHMC on x86 or Power Systems LPAR

Hardware support for CRUs comes as standard with the HMC. In addition, users can upgrade this support level to IBM onsite support to be consistent with other Power Systems servers.

Note:

- ▶ An HMC or vHMC is required for the Power E950 server to support multiple LPARs.
- ▶ For a single or full partition system, an HMC or vHMC is optional.
- ▶ Integrated Virtual Manager (IVM) is no longer supported.

For more information about vHMC, see [Virtual HMC Appliance \(vHMC\) Overview](#).

Figure 1-19 shows HMC model selections and tier updates.

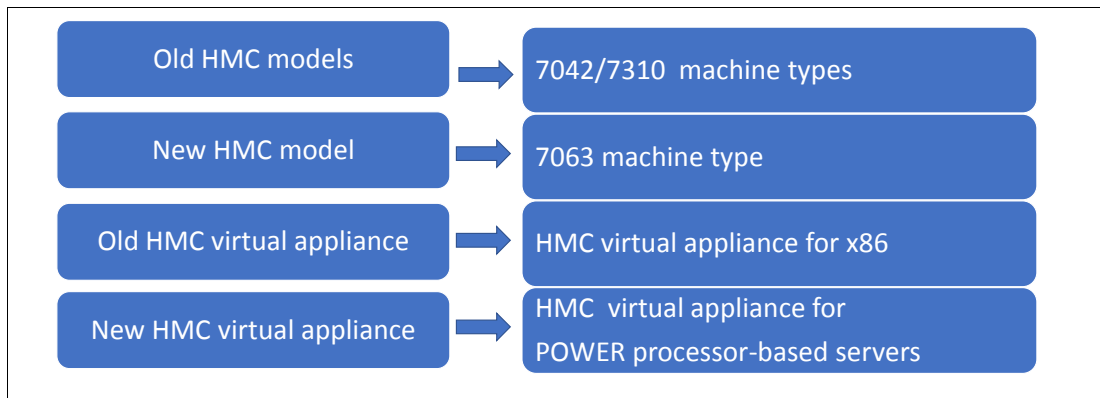


Figure 1-19 HMC selections

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.

Figure 1-20 shows several examples of HMC configurations.

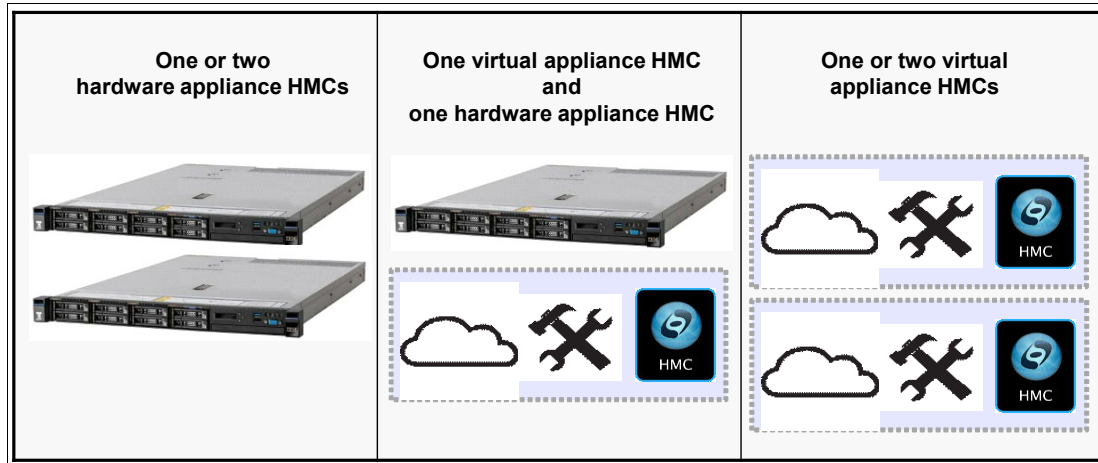


Figure 1-20 HMC configurations

1.7.3 Hardware Management Console code level

The HMC code must be running at Version 9 Release 1 Service Pack 920 or later when you use the HMC with the E950 server. You can check the version by looking in the HMC About window, as shown in Figure 1-21.

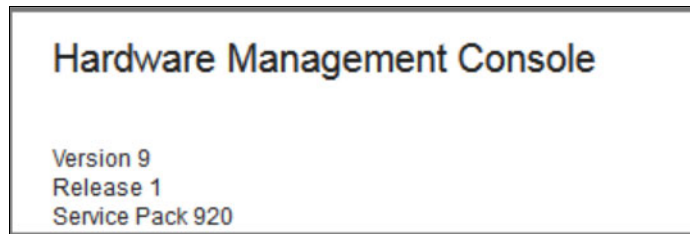


Figure 1-21 Hardware Management Console About window

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 [Fix Level Recommendation Tool \(FLRT\)](#) on or after the planned availability date for this product.

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

Note:

- ▶ Access to firmware and machine code updates is conditional on entitlement and license validation in accordance with IBM policy and practice. IBM might verify entitlement through customer number, serial number electronic restrictions, or any other means or methods that are employed by IBM at its discretion.
- ▶ HMC V9 supports only the Enhanced+ version of the GUI. The Classic version is no longer available.
- ▶ The HMC V9R1.911.0 release added support for managing IBM OpenPOWER systems. The same HMC that is used to manage FSP-based enterprise systems can now also manage the baseboard management controller (BMC)-based AC/LC servers. This provides a consistent and consolidated hardware management solution.
- ▶ HMC V9 supports connections to servers that are based on IBM servers that are based on POWER9, POWER8, and POWER7 processors. There is no support in this release for servers that are based on POWER6 processors or earlier.

1.7.4 Two architectures of Hardware Management Console

There are now two options for the HMC hardware: The earlier Intel-based HMCs, and the newer HMCs that are based on an IBM POWER8 processor. x86-based HMCs are no longer available to order, but are supported as an option for managing the Power E980 server.

You may use either architecture to manage the servers. You also may use one Intel-based HMC and one POWER8 based HMC if the software is at the same level.

The preferred practice is to use the new POWER8 processor-based consoles for server management.

Intel-based HMCs

HMCs that are based on Intel processors that support V9 code are:

- ▶ 7042-CR9
- ▶ 7042-CR8
- ▶ 7042-CR7

7042-CR6 and earlier HMCs are not supported for use with the Power E980 server.

The 7042-CR9 has the following specifications:

- ▶ 2.4 GHz Intel Xeon Processor E5-2620 V3
- ▶ 16 GB (1 x 16 GB) of 2.133 GHz DDR4 system memory
- ▶ 500 GB SATA SFF HDD
- ▶ SATA CD/RW - DVD-RAM
- ▶ Four Ethernet ports
- ▶ Six USB ports (two front and four rear)
- ▶ One PCIe slot

POWER8 processor-based HMC

The POWER processor-based HMC is machine type and model 7063-CR1. It has the following specifications:

- ▶ 1U base configuration
- ▶ IBM POWER8 120 W 6-core CPU
- ▶ 32 GB (4 x 8 GB) of DDR4 system memory

- ▶ 2 x 2 TB SATA LFF 3.5-inch HDD RAID 1
- ▶ Rail bracket option for round hole rack mounts
- ▶ Two USB 3.0 hub ports in the front of the server
- ▶ Two USB 3.0 hub ports in the rear of the server
- ▶ Redundant 1 kW power supplies
- ▶ 4 x 10 Gb Ethernet Ports (RJ-45) (10 Gb/1 Gb/100 Mb)
- ▶ 1 x 1 Gb Ethernet Port for Management (BMC)

All future HMC development will be done for the POWER8 processor-based 7063-CR1 and its successors.

Note: System administrators can remotely start or stop a 7063-CR1 HMC by running `ipmitool` or by using the WebUI.

1.7.5 Connectivity to POWER9 processor-based systems

POWER9 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. Additionally, to perform an operation on an LPAR, initiate Live Partition Mobility (LPM), or do PowerVM Active Memory Sharing operations, an Ethernet link to the managed partitions is required. A minimum of two Ethernet ports are needed on the HMC to provide this connectivity.

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

Figure 1-22 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.

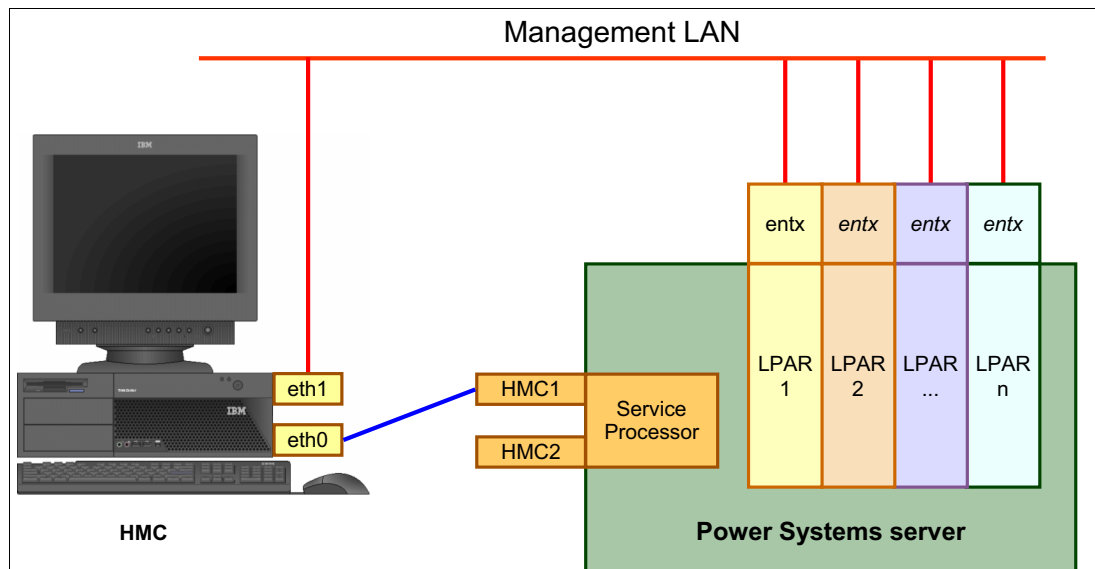


Figure 1-22 Network connections from the HMC to service processor and LPARs

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 FSP is allocated an IP address from a set of address ranges that is predefined in the HMC software.

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

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

- ▶ 1 Gbps connection speed.
- ▶ 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

1.7.6 High availability Hardware Management Console 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 that uses PowerVM LPM, or the creation of a partition, cannot be performed without the HMC. Power Systems servers support the capability to have two HMCs attached to a system, which provides redundancy in case one of the HMCs is unavailable.

To achieve HMC redundancy for a POWER9 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 functions.

Figure 1-23 shows one possible highly available HMC configuration that manages two servers. Each HMC is connected to one FSP port of each managed server.

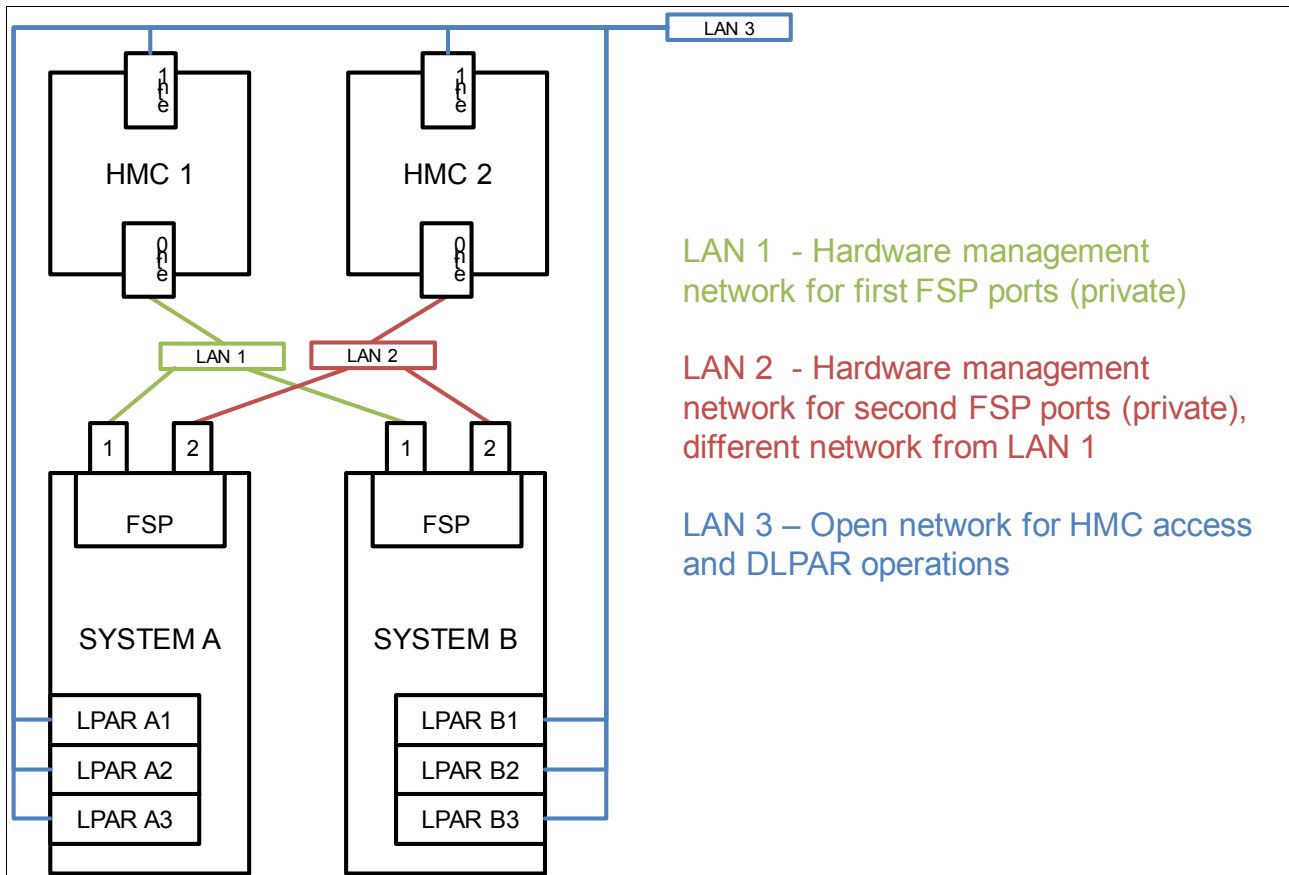


Figure 1-23 Highly available HMC networking example

For simplicity, only the hardware management networks (LAN1 and LAN2) are highly available. 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.



Architecture and technical overview

This chapter describes the overall system architecture for the IBM Power System E950 (9040-MR9) server. 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.

Figure 2-1 shows the logical system architecture of the Power E950 server.

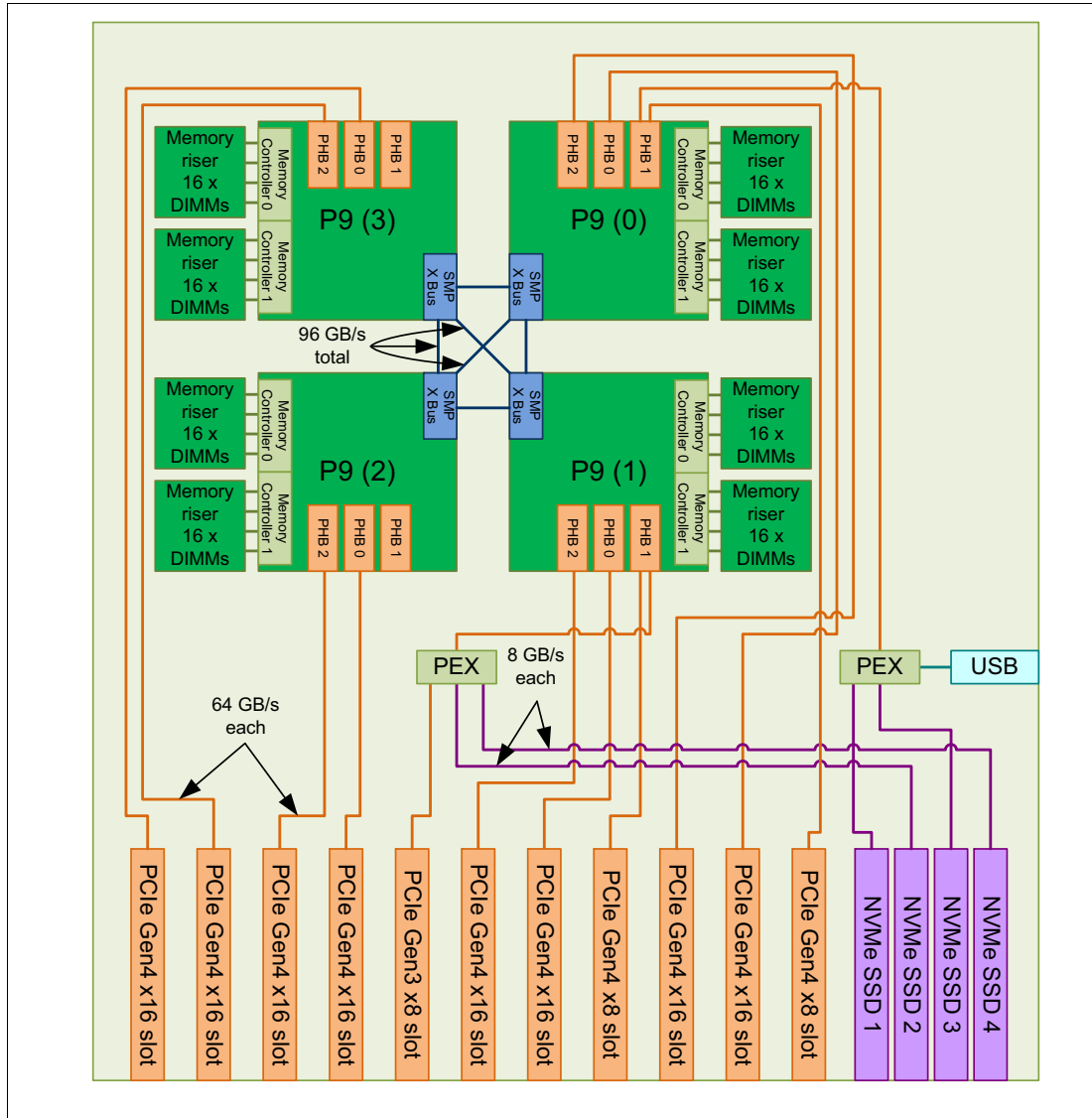


Figure 2-1 Power E950 logical system diagram

2.1 The IBM POWER9 processor

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

2.1.1 POWER9 processor overview

The POWER9 processor is composed of single-chip modules (SCMs) that are manufactured on the IBM 14-nm FinFET Silicon-On-Insulator (SOI) architecture. Each module is 68.5 mm x 68.5 mm and contains 8 billion transistors.

As shown in Figure 2-2, the chip contains 12 cores, two memory controllers, Peripheral Component Interconnect Express (PCIe) Gen4 I/O controllers, and an interconnection system that connects all components within the chip at 7 TBps. Each core has 512 KB of level 2 cache, and 10 MB of level 3 embedded DRAM (eDRAM) cache. The interconnect also extends through module and system board technology to other POWER9 processors in addition to memory and various I/O devices.

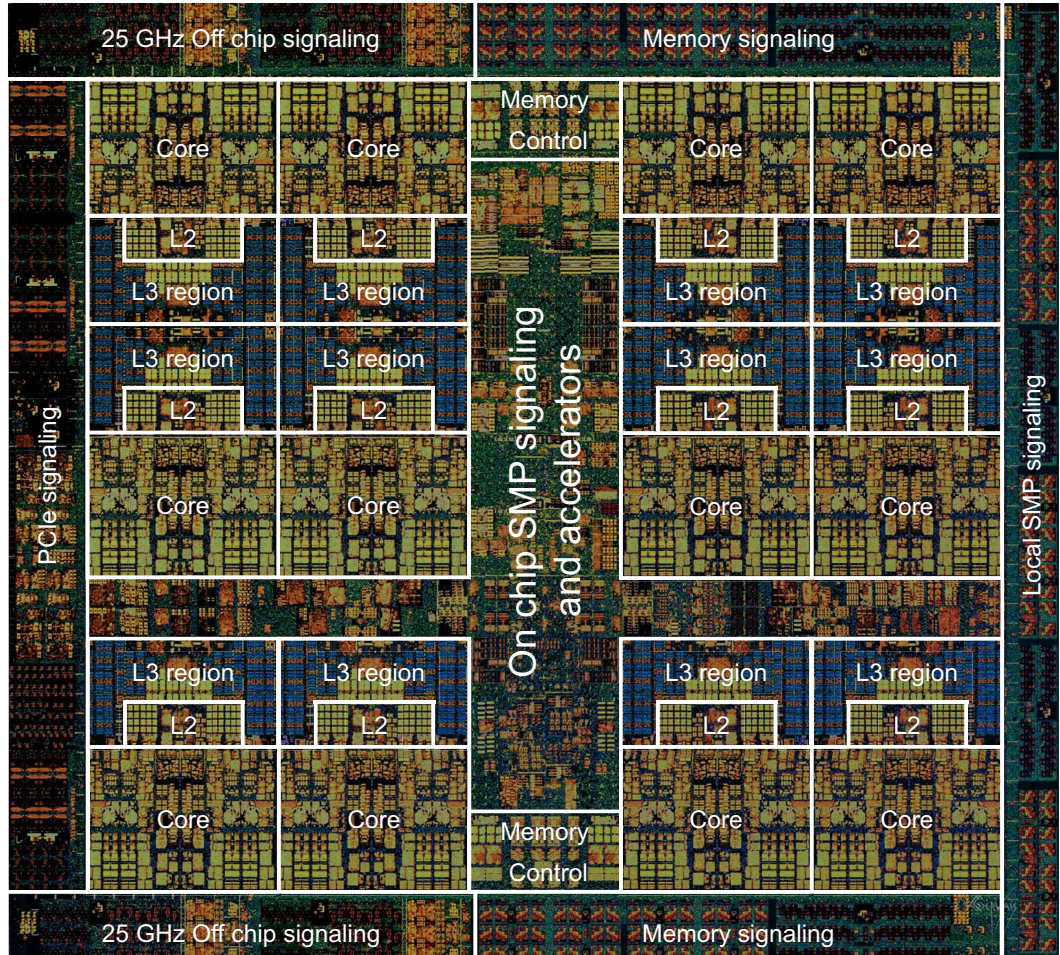


Figure 2-2 The POWER9 processor chip

The Power E950 server uses memory buffer chips to interface between the POWER9 processor and the 16 MB L4 cache or DDR4 memory. Each buffer chip includes the L4 cache to reduce the latency of local memory accesses.

The POWER9 chip provides an embedded algorithm for the following features:

- ▶ External Interrupt Virtualization Engine. Reduces the code impact/path length and improves performance compared to the previous architecture.
- ▶ File compression and decompression.
- ▶ PCIe Gen4 support.
- ▶ Two memory controllers that support buffered connection to memory.
- ▶ Cryptography: Advanced encryption standard (AES) engine.
- ▶ Random number generator (RNG).

- ▶ Secure Hash Algorithm (SHA) engine: SHA-1, SHA-256, and SHA-512, and Message Digest 5 (MD5).
- ▶ IBM Data Mover Tool.

Table 2-1 provides a summary of the POWER9 processor technology.

Note: The total values represent the maximum of 12 cores for the POWER9 architecture. The Power E950 server has options for 32, 40, 44, and 48 cores per node.

Table 2-1 Summary of the POWER9 processor technology

Technology	POWER9 processor
Module size	68.5 mm × 68.5 mm
Fabrication technology	<ul style="list-style-type: none"> ▶ 14-nm lithography ▶ Copper interconnect ▶ SOI ▶ eDRAM
Maximum processor cores	12
Maximum execution threads core/module	8/96
Maximum L2 cache core/module	512 KB/6 MB
Maximum On-chip L3 cache core/module	10 MB/120 MB
Number of transistors	8 billion
Compatibility	With prior generation of POWER processor

2.1.2 POWER9 processor core

The POWER9 processor core is a 64-bit implementation of the IBM Power Instruction Set Architecture (ISA) Version 3.0, and has the following features:

- ▶ Multi-threaded design, which is capable of up to eight-way simultaneous multithreading (SMT)
- ▶ 64 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 that uses 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, which is capable of up to 16 floating point operations per cycle (eight double precision or 16 single precision)
- ▶ In-core AES encryption capability

- ▶ Hardware data prefetching with 16 independent data streams and software control
- ▶ Hardware decimal floating point (DFP) capability

For more information about Power ISA Version 3.0, see [OpenPOWER: IBM Power ISA Version 3.0B](#).

Figure 2-3 shows a picture of the POWER9 core, with some of the functional units highlighted.

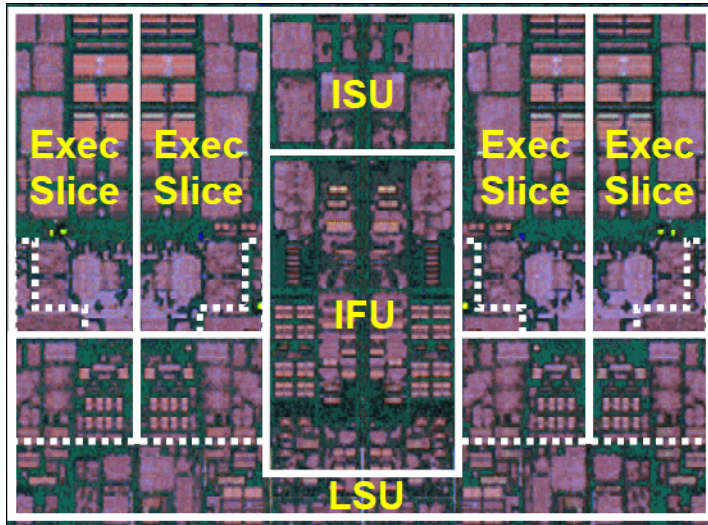


Figure 2-3 POWER9 processor chip

2.1.3 Simultaneous multithreading

POWER9 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 enables a single physical processor core to simultaneously dispatch instructions from more than one hardware thread context. With SMT, each POWER9 core can present eight hardware threads. Because there are multiple hardware threads per physical processor core, more instructions can run at the same time.

SMT is primarily beneficial in commercial environments where the speed of an individual transaction is not as critical as the total number of transactions that are performed. SMT typically increases the throughput of workloads with large or frequently changing working sets, such as database servers and web servers.

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

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

Technology	Cores/system	Maximum SMT mode	Maximum hardware threads per partition
IBM POWER4	32	Single thread	32
IBM POWER5	64	SMT2	128
IBM POWER6	64	SMT2	128

Technology	Cores/system	Maximum SMT mode	Maximum hardware threads per partition
IBM POWER7	256	SMT4	1024
IBM POWER8	192	SMT8	1536
IBM POWER9	48	SMT8	1536
IBM POWER9	192	SMT8	1536

2.1.4 POWER9 compatibility modes

The POWER9 processor has the ability to run in compatibility modes for previous POWER processor generations, which enables older operating systems to run on POWER9 systems. Compatibility modes also enable Live Partition Mobility (LPM) from systems that are based on previous generations of POWER processors. The POWER9 processor can run in the following compatibility modes:

- ▶ POWER7
- ▶ POWER8
- ▶ POWER9 Base

2.1.5 Processor feature codes

Each system enclosure in a Power E950 server has four sockets for processor modules. All sockets must be populated with a matching processor module. In servers with multiple system enclosures, all sockets on all enclosures must be populated with matching processor modules.

Table 2-3 shows the processor features that are available for the Power E950 server.

Table 2-3 Power E950 processor features

Feature code	CCIN	Description
EHC4	5C38	Solution Edition for Healthcare typical 3.15 - 3.8 GHz, 12-core Processor Module
EPWR	5C34	8-core typical 3.6 - 3.8 GHz (maximum) processor
EPWS	5C37	10-core 3.4 - 3.8 GHz (maximum) processor
EPWT	5C38	12-core 3.15 - 3.8 GHz (maximum) processor
EPWY	5C45	11-core 3.2 - 3.8 GHz (maximum) processor

Processors in the Power E950 system support Capacity on Demand (CoD). For more information about CoD, see 2.3, "Capacity on Demand" on page 68.

Processor-specific CoD features are shown in Table 2-4.

Table 2-4 Processor activation feature codes

Processor feature	Static activation feature	Static activation for Linux feature
EHC4 12-core module for Healthcare	ELAN ^a	N/A
EPWR 8-core module	EPWV	ELBG
EPWS 10-core module	EPWW	ELBP
EPWT 12-core module	EPWX	ELBH
EPWY 11-core module	EPN3	ELBR

a. ELAN activates 12 cores for each feature. It is mandatory to order ELAN with each EHC4 module.

CoD features that are independent of the processor feature are shown in Table 2-5.

Table 2-5 Processor-independent activation features

Feature code	Description
EP9T	90 Days Elastic CoD Processor Core Enablement
EPN0	1 Proc-day Elastic billing for EPWR/EPWK
EPN1	100 Proc-day Elastic CoD billing EPWR/EPWK AIX
EPN2	100 Proc-mins Utility CoD billing EPWR/EPWK
EPN5	1 Proc-day Elastic billing for EPWS/EPWL
EPN6	100 Proc-day Elastic CoD billing EPWS/EPWL AIX
EPN7	100 Proc-mins Utility CoD billing EPWS/EPWL
EPN8	1 Proc-day Elastic billing for EPWY/EPWZ
EPN9	100 Proc-day Elastic CoD billing EPWY/EPWZ AIX
EPNK	1 Proc-day Elastic billing for EPWT/EPWM
EPNL	100 Proc-day Elastic CoD billing EPWT/EPWM AIX
EPNM	100 Proc-mins Utility CoD billing EPWT/EPWM
EPNN	100 Proc-mins Utility CoD billing EPWY/EPWZ
MMCB	ECOD GB Memory Day - AIX/Linux
MMCY	ECOD Processor Day - AIX/Linux

For more information about CoD capabilities and features, see 2.3, “Capacity on Demand” on page 68.

2.1.6 Memory access

One POWER9 processor module of the Power E950 server provides two integrated memory controllers to facilitate access to the system's main memory. One memory controller drives four differential memory interface (DMI) channels with a maximum signaling rate of 9.6 GHz. This setup yields a sustained bandwidth of up to 28.8 GBps per memory channel or 230.4 GBps per processor module. Every DMI channel connects to one dedicated memory buffer chip. Each memory buffer chip in turn provides four DDR4 memory ports running at a 1,600 MHz signal rate and one 16 MB L4 cache. One DDR4 port connects to one industry-standard dual inline memory module (IS DIMM) slot that is populated with one DDR4 DIMM.

Four memory buffer chips are mounted with their 16 associated IS DIMM slots on one memory riser card. Every processor module of a Power E950 server uses either one or two memory riser cards. On a new system order, the DDR4 technology-based IS DIMMs are available with 8 GB, 16 GB, 32 GB, 64 GB, and 128 GB capacity. Due to the change in the memory access architecture, no DDR3 or DDR4 CDIMMs of the POWER8 processor-based predecessor Power E850 or Power E850C server models can be reused in the Power E950 server.

Figure 2-4 shows the POWER9 hierarchical memory subsystem of a Power E950 system.

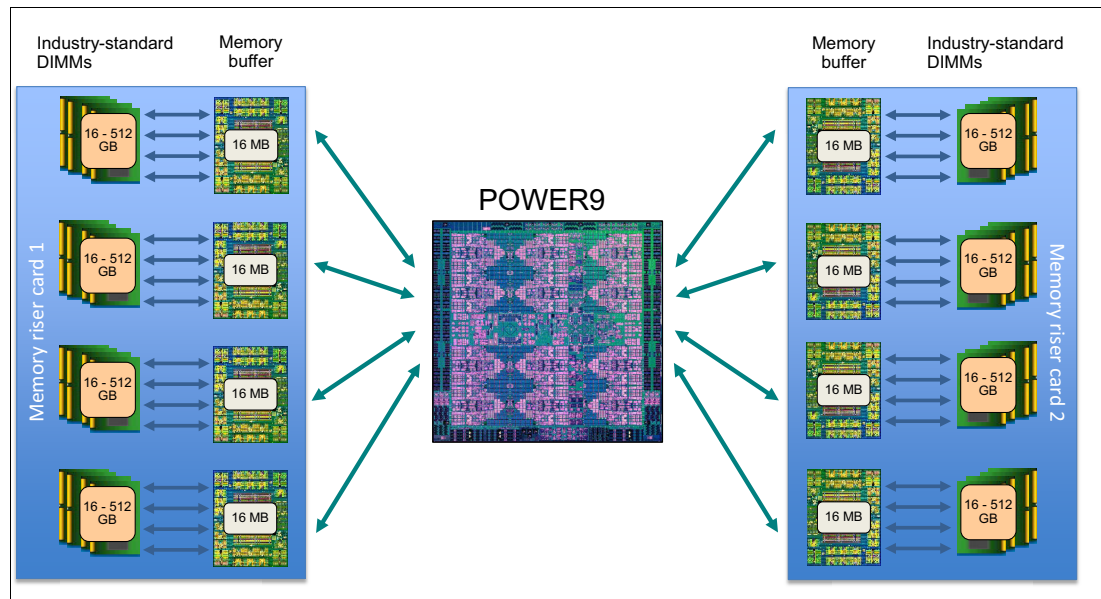


Figure 2-4 POWER9 hierarchical memory subsystem that uses memory buffers and IS DIMMs

The maximum supported memory capacity per processor module is 4 TB, which requires 128 GB IS DIMMs in all 16 DIMM slots in each of the two memory riser cards. A maximum of 16 TB of main memory can be made accessible to applications by a Power E950 server in a four-processor module configuration.

For more information about memory placement rules, memory bandwidth, and other topics that are related to the memory subsystem of the Power E950 server, see 2.2, "Memory subsystem" on page 58.

2.1.7 On-chip L3 cache innovation and intelligent caching

Similar to POWER8, the POWER9 processor uses a breakthrough in material engineering and microprocessor fabrication to implement the L3 cache in eDRAM technology and place it on the processor die. The 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 symmetric multiprocessor (SMP) interconnect.

Like the POWER8 processor, the POWER9 processor supports the same L3 non-uniform cache access (NUCA) architecture that provides mechanisms to distribute and share cache footprints across the chip. The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a fast 10 MB local region of L3 cache (FLR-L3), but also has access to other L3 cache regions as a 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 POWER9 processor to optimize the access to L3 cache lines and minimize overall cache latencies.

Regarding the POWER8 L3 implementation, the POWER9 L3 introduces an enhanced replacement algorithm with data type and reuse awareness that uses information from the core and L2 cache to manage cache replacement states. The L3 cache supports an array of prefetch requests from the core, including both instruction and data, and for different levels of urgency. Prefetch requests for the POWER9 processor include extra information exchange between the core, cache, and the memory controller to manage memory bandwidth and mitigate the prefetch based cache pollution.

The following list provides an overview of the features that are offered by the POWER9 L3 cache:

- ▶ Private 10 MB L3.0 cache/shared L3.1:
 - Victim cache for local L2 cache (L3.0)
 - Victim cache for other on-chip L3 caches (L3.1)
- ▶ 20-way set associative.
- ▶ 128-byte cache lines with 64-byte sector support.
- ▶ 10 EDRAM banks (interleaved for access overlapping).
- ▶ 64-byte wide data bus to L2 for reads.
- ▶ 64-byte wide data bus from L2 for L2 castouts.
- ▶ Eighty 1 Mb EDRAM macros that are configured in 10 banks, with each bank having a 64-byte wide data bus.
- ▶ All cache accesses have the same latency.
- ▶ 20-way directory that is organized as four banks, with up to four reads or two reads and two writes every two processor clock cycles to differing banks.

The L3 cache architecture of the 12-core POWER9 processor is identical to the 24-core POWER9 implementation. For more information about the L3 cache technology, see [POWER9 Processor User's Manual](#).

For more information about the L3 cache in the context of the POWER9 core architecture, see H. Le, et al., "IBM POWER9 processor core", *IBM Journal of Research & Development Volume 62 Number 4/5, July/September 2018* by searching at [IBM Journal of Research & Development](#).

2.1.8 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 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 enables programs to use a programming style that is close to coarse-grained locking to achieve performance that is close to fine-grained locking.

Most implementations of transactional memory are based on software. The POWER9 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 enabling the system to operate at maximum performance.

2.1.9 POWER9 accelerator processor interfaces

The POWER9 processor modules of the Power E950 server support two dedicated interfaces and protocols to attach advanced accelerator and future memory technologies:

- ▶ Open Coherent Accelerator Processor Interface (OpenCAPI)
- ▶ NVIDIA NVLink

On Power E950 systems, OpenCAPI-attached accelerators and devices and NVLink graphics processing units (GPUs) are supported by one bus per POWER9 processor module. Each bus provides two independently accessible groups of eight lanes for either OpenCAPI or NVLink protocol support at a signaling rate of 25.78 Gbps. One group of eight lanes is also commonly referred to as a *brick*. The connectors to the individual bricks are placed on the system board and are referenced by the location codes that are shown in Table 2-6.

Table 2-6 Power E950 internal accelerator interfaces for OpenCAPI and NVLink

	Processor module CP0	Processor module CP1	Processor module CP2	Processor module CP3
OpenCAPI / NVLink connector location codes	P1-T3 P1-T4	P1-T5 P1-T6	P1-T7 P1-T8	P1-T9 P1-T10

The maximum bandwidth of one brick that is used to attach a OpenCAPI accelerator or device or a NVLink GPU is 51.56 GBps:

$$1 \text{ brick} \times 8 \text{ lanes} \times 25.78 \text{ Gbps} \times 2 \text{ full duplex} = 51.56 \text{ GBps}$$

The OpenCAPI technology is developed and standardized by the OpenCAPI consortium. In-depth information about the consortium's mission and the OpenCAPI protocol specification can be found at [OpenCAPI](#).

The NVLink technology is developed by the NVIDIA corporation. For more information about the NVLink protocol, see [NVIDIA NVLink](#).

Note: The Power E950 system is designed to support the OpenCAPI and the NVLink protocol, but at the time of writing OpenCAPI-attached accelerators or devices and NVLink-attached GPUs are *not* supported. The information pertaining to these technologies are included for future reference only.

Open Coherent Accelerator Processor Interface

In October 2016, AMD, Google, IBM, Mellanox Technologies, and Micron formed the OpenCAPI not-for-profit organization to create an open, coherent, high-performance bus interface that is based on a new bus standard that is called OpenCAPI, and grow the infrastructure that uses this interface. This initiative is being driven by the emerging accelerated computing and advanced memory/storage solutions that have introduced significant system bottlenecks in today's current open bus protocols, and requires a technical solution that is openly available.

Two major technology trends heavily impact the industry currently:

- ▶ Hardware acceleration will become commonplace as microprocessor technology and design continues to deliver far less than the historical rate of cost/performance improvement per generation.
- ▶ New advanced memory technologies will change the economics of computing.

Existing system interfaces are insufficient to address these disruptive forces. Traditional I/O architecture results in high processor usage when applications communicate with I/O or accelerator devices at the necessary performance levels. Also, they cannot integrate multiple memory technologies with different access methods and performance attributes.

These challenges are addressed by the OpenCAPI architecture in a way that allows full industry participation. Embracing an open architecture is fundamental to establish sufficient volume base to reduce cost and ensure the support of a broad infrastructure of software products and attached devices.

OpenCAPI is an open interface architecture that allows any microprocessor to attach to:

- ▶ Coherent user-level accelerators and I/O devices
- ▶ Advanced memories accessible through read/write or user-level direct memory access (DMA) semantics

OpenCAPI is neutral regarding processor architectures, and exhibits the following key attributes:

- ▶ A high-bandwidth, low-latency interface that is optimized to enable streamlined implementations of attached devices.
- ▶ A 25 Gbps signaling and protocol that is built to enable a low latency interface on processors and attached devices.
- ▶ Complexities of coherence and virtual addressing are implemented on the host microprocessor to simplify attached devices and facilitate interoperability across multiple CPU architectures.
- ▶ Attached devices operate natively within an application's user space and coherently with processors allowing attached devices to fully participate in applications without kernel involvement/overhead.

- ▶ Supports a wide range of use cases and access semantics:
 - Hardware accelerators
 - High-performance I/O devices
 - Advanced memories

2.1.10 Power and performance management

POWER9 processor-based scale-out and scale-up servers implement Workload Optimized Frequency (WOF) as a new feature of the power management EnergyScale technology. With EnergyScale in POWER9 processor-based servers, the POWER8 dynamic power saver (DPS) modes that either favor lower power consumption (DPS) or favor performance (DPS-FP) are replaced by two new power saver modes:

- ▶ Dynamic performance mode (DPM)
- ▶ Maximum performance mode (MPM)

Every POWER9 processor-based scale-out or scale-up system has either the DPM or the MPM enabled by default. Both modes dynamically adjust the processor frequency to maximize performance and enable a much higher processor frequency range in comparison to POWER8 servers. Each of the new power saver modes deliver consistent system performance without any variation if the nominal operating environment limits are met. For POWER9 processor-based systems that are under control of the PowerVM hypervisor, the DPM and MPM are a systemwide configuration setting, but each processor module frequency is optimized separately.

Several factors determine the maximum frequency a processor module can run at:

- ▶ Processor usage: Lighter workloads run at higher frequencies.
- ▶ Number of active cores: Fewer active cores run at higher frequencies.
- ▶ Environmental conditions: At lower ambient temperatures, the cores are enabled to run at higher frequencies.

The new power saver modes are defined as follows:

DPM In DPM, the workload is run at the highest frequency possible if the nominal power consumption limit of the processor modules is not exceeded. The frequency of the processor modules is always at the nominal frequency of the POWER9 processor-based system or above the nominal frequency up to the upper limit of the DPM frequency range. This DPM typical frequency range (DTFR) is published as part of the system specifications of a particular POWER9 system if it is running by default in the DPM.

The system performance is deterministic within the allowed operating environmental limits, so it does not depend on the ambient temperature if the temperature stays within the supported range. The idle power saver (IPS) function can be enabled or disabled. If IPS is enabled and all cores in a processor module are idle for hundreds of milliseconds, the frequency of the cores in the respective module drop to the predefined power save frequency.

MPM In MPM, the workload is run at the highest frequency possible, but unlike in DPM, the processor module may operate at a higher power consumption level. The higher power draw enables the processor modules to run in an MPM typical frequency range (MTFR) where the lower limit is above the nominal frequency and the upper limit is determined by the system's maximum frequency.

The MTFR is published as part of the system specifications of a particular POWER9 system if it is running by default in MPM. The higher power draw potentially increased the fan speed of the respective system node to meet the higher cooling requirements, which caused a higher noise emission level of up to 15 decibels. The processor frequency typically stays within the limits that are determined by the MTFR, but may be lowered to frequencies between the MTFR lower limit and the nominal frequency at high ambient temperatures above 25 °C (77 °F). If the data center ambient environment is less than 25 °C, the frequency in MPM is consistently in the upper range of the MTFR (roughly 10% - 20% better than nominal). At lower ambient temperatures (below 25 °C), MPM mode also provides deterministic performance. As the ambient temperature increases above 25 °C, determinism can no longer be guaranteed.

The IPS function can be enabled or disabled. If IPS is enabled, the frequency is dropped to the static power saver level if the entire system meets the configured idle conditions.

Figure 2-5 shows the frequency ranges for the POWER9 static nominal mode (all modes disabled), the DPM, and the MPM. The frequency adjustments for different workload characteristics, ambient conditions, and idle states are also indicated.

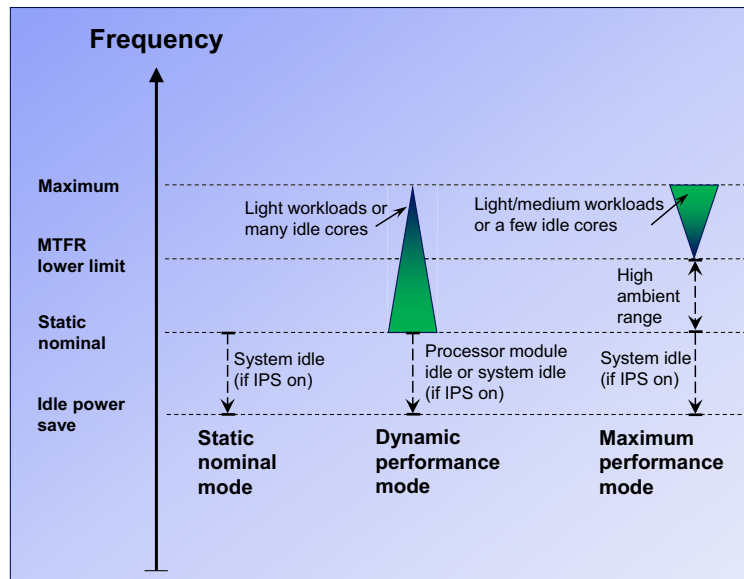


Figure 2-5 POWER9 power management modes and related frequency ranges

Table 2-7 shows the static nominal and the static power saver mode frequencies, and the frequency ranges of the DPM and the MPM for all four processor module types that are available for the Power E950 server.

Table 2-7 Characteristic frequencies and frequency ranges for Power E950 server

Feature code	Cores per SCM	Static nominal frequency [GHz]	Static power saver mode frequency [GHz]	DPM frequency range [GHz]	MPM frequency range [GHz]
EPWR	8	3.3	2.75	3.3 - 3.8	3.6 - 3.8
EPWS	10	3.0	2.75	3.0 - 3.8	3.4 - 3.8

Feature code	Cores per SCM	Static nominal frequency [GHz]	Static power saver mode frequency [GHz]	DPM frequency range [GHz]	MPM frequency range [GHz]
EPWY	11	2.85	2.75	2.85 - 3.8	3.2 - 3.8
EPWT	12	2.8	2.75	2.8 - 3.8	3.15 - 3.8

Figure 2-6 shows the POWER9 processor frequency as a function of power management mode and system utilization.

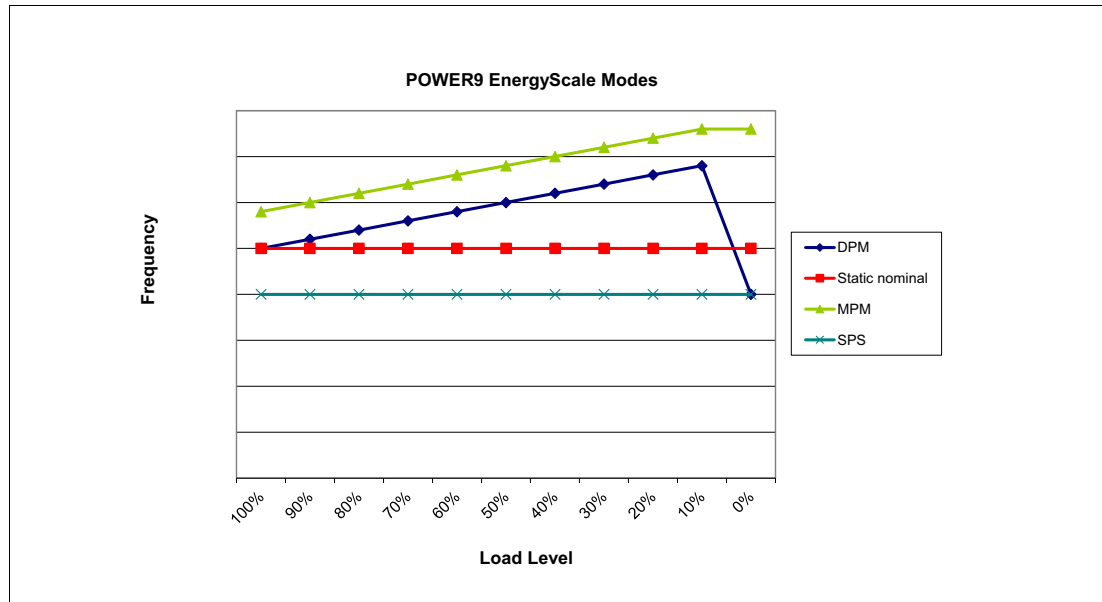


Figure 2-6 POWER9 processor frequency as a function of power management mode and system load

The default performance mode depends on the POWER9 processor-based server model. For Power E950 systems, MPM is enabled by default.

The controls for all power saver modes are available on the Advanced System Management Interface (ASMI) and can be dynamically modified, which includes enabling or disabling the IPS function and changing the EnergyScale tunable parameters. A system administrator may also use the Hardware Management Console (HMC) to disable all power saver modes or to enable one of the three available power and performance modes: static power saver mode, DPM, or MPM.

Figure 2-7 shows the ASMI menu for Power and Performance Mode Setup on a Power E950 server.

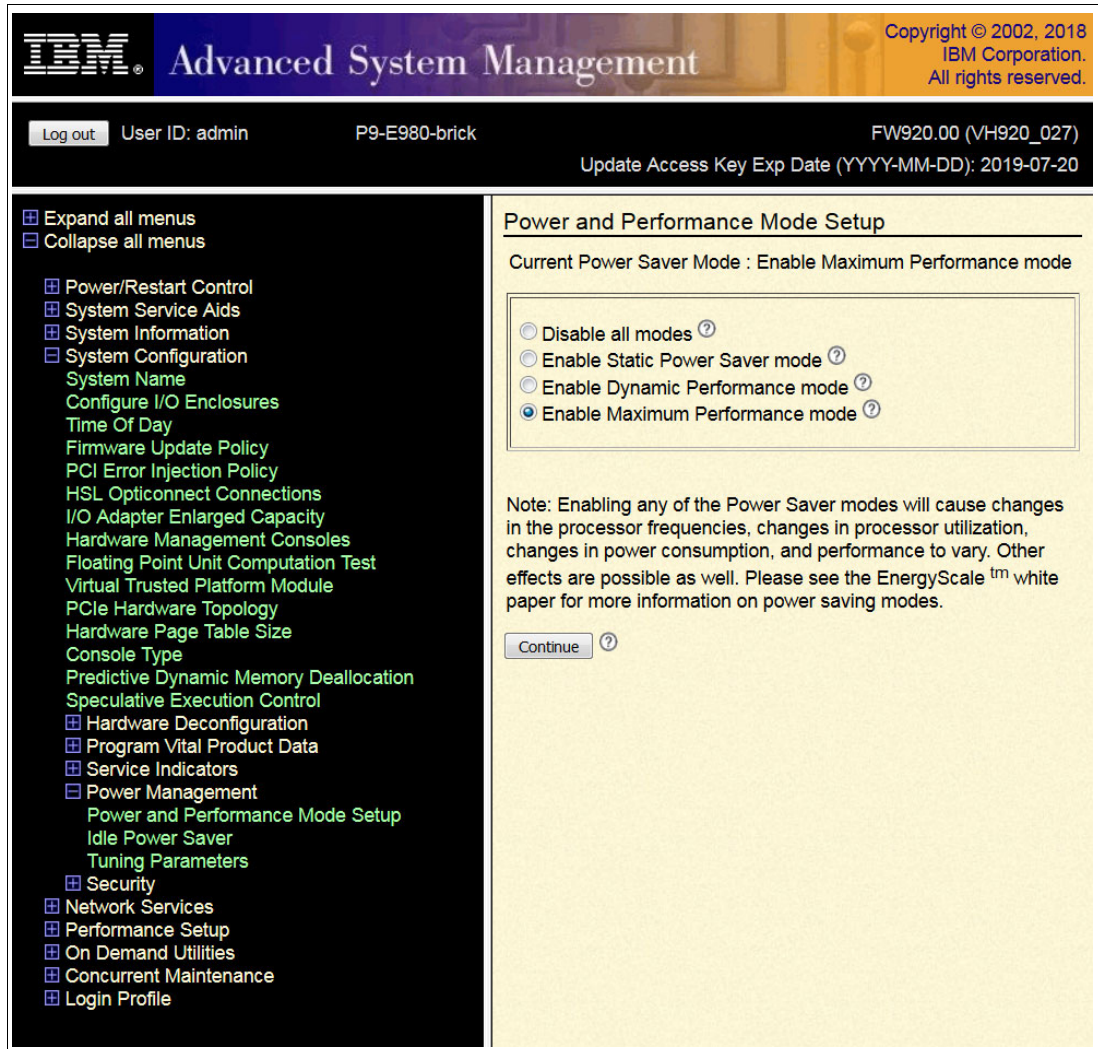


Figure 2-7 Power E950 ASMI menu for power and performance mode setup

For more information about the POWER9 EnergyScale technology, see [POWER9 EnergyScale Introduction](#).

2.1.11 Comparison of the POWER9, POWER8, and POWER7+ processors

The Power E950 server exclusively uses processor modules with 12 cores that can support eight SMT execution contexts. These 12-core modules are referred to as POWER9 *scale-up* processor modules. All other POWER9 processor-based systems, such as Power S914, Power LC921, Power S922, Power H922, Power L922, Power LC922, Power AC922, Power S924, and Power H924 servers, run processor modules with 24 cores that can support four SMT execution contexts. These 24-core modules are referred to as POWER9 *scale-out* processor modules.

Table 2-8 shows the key features and characteristics as compared between the POWER9 scale-up, POWER9 scale-out, POWER8, and POWER7+ processor implementations.

Table 2-8 Comparison of technology for the POWER9 processor and prior processor generations

Characteristics	POWER9 scale-up	POWER9 scale-out	POWER8	POWER7+
Technology	14 nm	14 nm	22 nm	32 nm
Module size	68.5 mm x 68.5 mm	68.5 mm x 68.5 mm	649 mm ²	567 mm ²
Die size	25.228 mm x 27.48416 mm (695 mm ²)	25.228 mm x 27.48416 mm (695 mm ²)		
Number of transistors	8 billion	8 billion	4.2 billion	2.1 billion
Maximum cores	12	24	12	8
Maximum SMT threads per core	8 threads	4 threads	8 threads	4 threads
Maximum frequency	3.9 - 4.0 GHz	3.8 - 4.0 GHz	4.15 GHz	4.4 GHz
L2 Cache	512 KB per core	256 KB per core	512 KB per core	256 KB per core
L3 Cache	10 MB of L3 cache per core with each core having access to the full 120 MB of L3 cache, on-chip eDRAM	10 MB of L3 cache that is shared by two cores with each core having access to the full 120 MB of L3 cache, on-chip eDRAM	8 MB of L3 cache per core with each core having access to the full 96 MB of L3 cache, on-chip eDRAM	10 MB of L3 cache per core with each core having access to the full 80 MB of L3 cache, on-chip eDRAM
Memory support	DDR4 and DDR3 ^a	DDR4	DDR3 and DDR4	DDR3
I/O bus	PCIe Gen4	PCIe Gen4	PCIe Gen3	GX++

a. Only DDR3 memory CDIMMs that are transferred in the context of a model upgrade from Power E880 or Power E880C systems to a Power E980 server are supported. DDR3 is not supported in the Power E950 server.

2.2 Memory subsystem

The Power E950 system uses one or two memory riser cards for each processor module to provide access to main memory. Four memory buffer chips are mounted on one memory riser card (#EM03), which facilitate access to 16 IS DIMM slots. The slots use the industry-standard form factor, but support only IBM supplied DDR4 DIMMs. The slots support DDR4 IS DIMMs with 8 GB, 16 GB, 32 GB, 64 GB, and 128 GB capacity.

- ▶ A two-processor module Power E950 configuration provides 64 IS DIMM slots through four memory riser cards with a maximum of 8 TB RAM.
- ▶ A four-processor module configuration provides 128 IS DIMM slots through eight memory riser cards and a server maximum of 16 TB RAM.

The Power E950 server does not support any DDR3 DIMM modules. Because of the change in the memory subsystem architecture, it is not possible to reuse older CDIMM memory cards because they are installed in POWER8 processor-based Power E850 and Power E850C systems.

The memory of Power E950 systems is CoD-capable, enabling the purchase of more physical memory capacity that can then be dynamically activated when needed. At least 50% of the installed memory capacity must be active.

The Power E980 server supports an optional feature that is called Active Memory Expansion (AME) (#EMAM). By using this FC, the effective maximum memory capacity can be much larger than the true physical memory. This FC uses a dedicated coprocessor on the POWER9 processor to compress memory pages as they are written to and decompress them as they are read from memory. This process can deliver memory expansion of up to 125%, depending on the workload type and its memory usage.¹

2.2.1 DIMM memory riser card

The Power E950 system has a three-part memory subsystem design. This design consists of two memory controllers in each processor module that communicate through eight DMI channels with dedicated memory buffer modules that are mounted on memory riser cards (#EM03) to access DRAM memory modules on IS DIMMs.

Figure 2-8 shows the system board location codes P1-C26 - P1-C29 and P1-C32 - P1-C35 for the memory riser card slots.

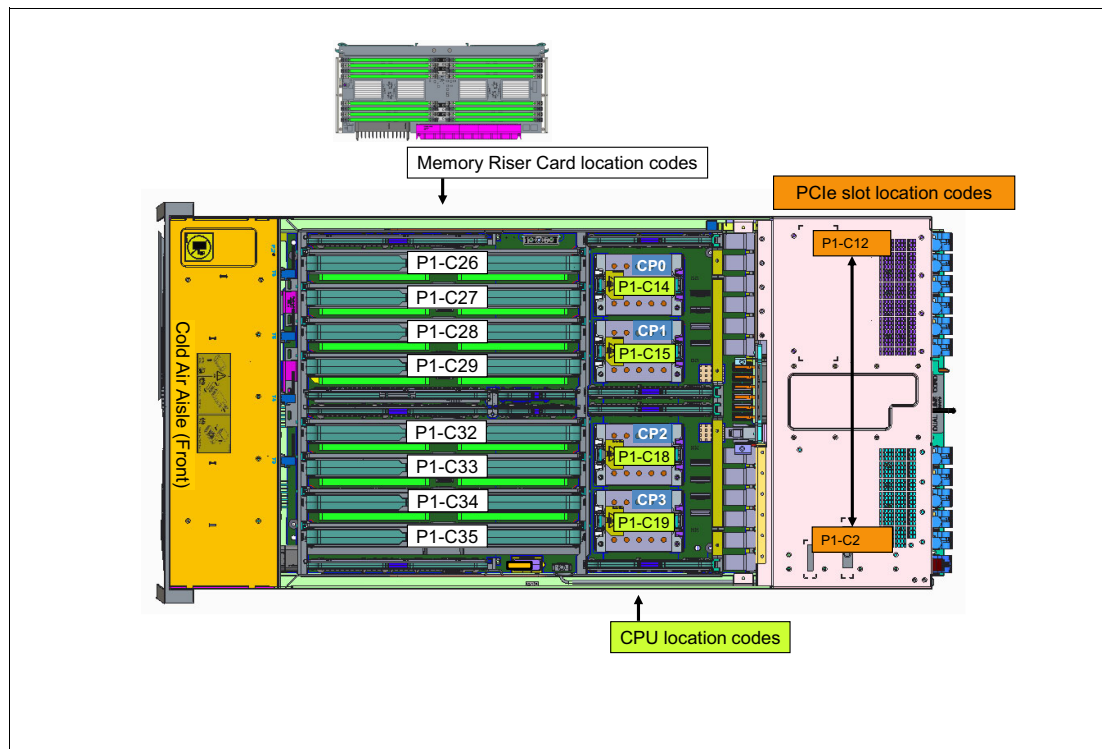


Figure 2-8 Power E950 system board location codes for memory riser cards

¹ A memory expansion of 125% requires the corresponding AME factor to be at 2.25. The AME factor can be set to a much higher value, up to the configuration maximum of 10. The best AME factor for any specific workload must be determined through systematic exploration and testing.

One processor module supports two memory riser cards. A minimum of one memory riser card must be installed; the second riser card is optional. Depending on the number of configured processor modules, the memory riser card placement sequence is determined by Table 2-9 for a 2-processor module and by Table 2-10 for a 4-processor module systems.

Table 2-9 Memory riser card placement sequence for two-processor module configurations

	Processor module CP0		Processor module CP1	
	MC0	MC1	MC0	MC1
Memory controller	MC0	MC1	MC0	MC1
Location code	P1-C26	P1-C27	P1-C28	P1-C29
Placement sequence	1 ^a	3	2 ^a	4

a. A riser card must be installed.

Table 2-10 Memory riser card placement sequence for four-processor module configurations

	Processor module CP0		Processor module CP1		Processor module CP2		Processor module CP3	
	MC0	MC1	MC0	MC1	MC0	MC1	MC0	MC1
Memory controller	MC0	MC1	MC0	MC1	MC0	MC1	MC0	MC1
Location code	P1-C26	P1-C27	P1-C28	P1-C29	P1-C32	P1-C33	P1-C34	P1-C35
Placement sequence	1 ^a	5	2 ^a	6	3 ^a	7	4 ^a	8

a. A riser card must be installed.

The memory buffer is a L4 cache and is built on eDRAM technology (same as the L3 cache), which has a lower latency than an SRAM. Each memory riser card holds four memory buffer chips. One memory buffer chip has 16 MB of L4 cache so that one memory riser card holds 64 MB of L4 cache. A fully populated four-processor module Power E950 server uses eight riser cards with a total of 512 MB of L4 cache capacity. The L4 cache performs several functions that have direct impact on performance and bring a series of benefits for the Power E950 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.

Each memory buffer provides four DDR4 ports that are referenced by labels A, B, C, and D, and every port connects to one IS DIMM slot in turn. The 16 IS DIMM slots are subdivided into two groups of eight slots. The first group is addressed by the A- and B-ports and is composed of IS DIMM slots with location codes C2, C4, C5, C7, C10, C12, C13, and C15. The second group is addressed by the C- and D-ports and is composed of IS DIMM slots with location codes C1, C3, C6, C8, C9, C11, C14, and C16.

Figure 2-9 shows the connection layout between the memory buffer chip ports and the related IS DIMM slots. The first group of IS DIMM slots is indicated by white and the second group of IS DIMM slots is indicated by black.

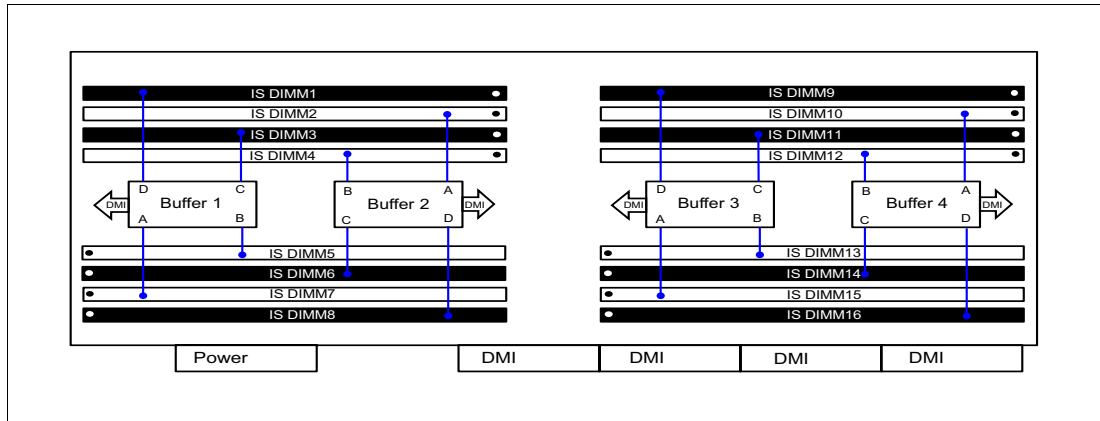


Figure 2-9 Memory buffer chip port connections to IS DIMM connector slots

Figure 2-10 shows the IS DIMM locations codes on the memory riser cards of a Power E950 system. The full locations code for an IS DIMM always starts by the label P1 followed by the location code of the memory riser card. In Figure 2-10, the memory riser card designation in the location codes is Cx, where x represents a variable to be replaced with any one of the related memory riser card location labels C25, C27, C28, C29, C32, C33, C34, or C35.

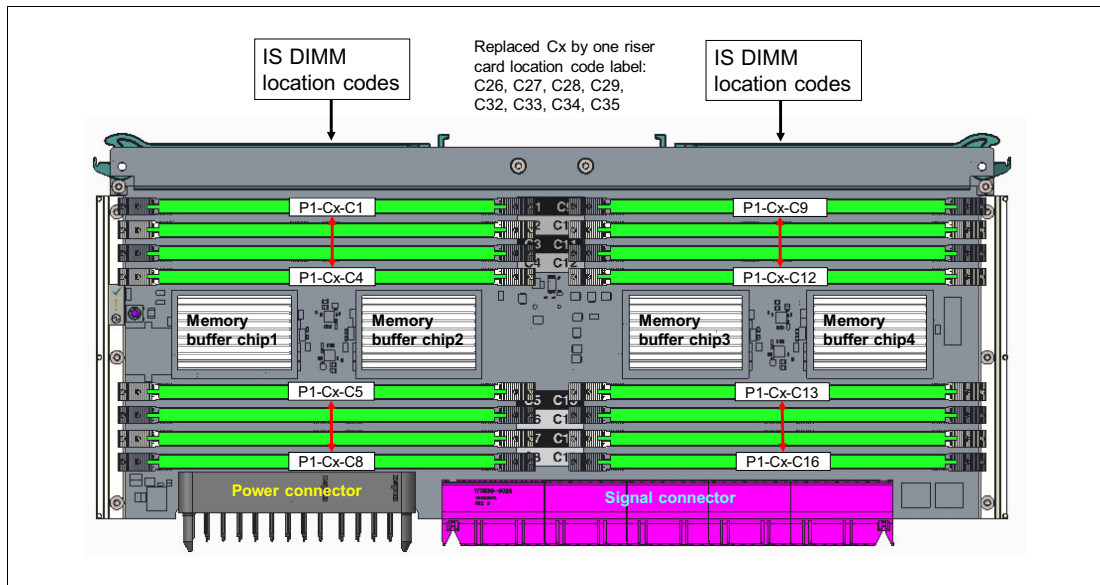


Figure 2-10 IS DIMM location codes on memory riser cards

The IS DIMM connector colors on a physical riser card are white for the first group and black for the second group of IS DIMM slots to assist IS DIMM installations. The first eight IS DIMMs must be installed in the white-colored IS DIMM connectors and the second set of eight IS DIMMs are confined to the remaining, black-colored connectors.

Figure 2-11 shows a physical memory riser card without any of the slots populated with IS DIMMs. The black/white color coding of the IS DIMM connectors is clearly visible.

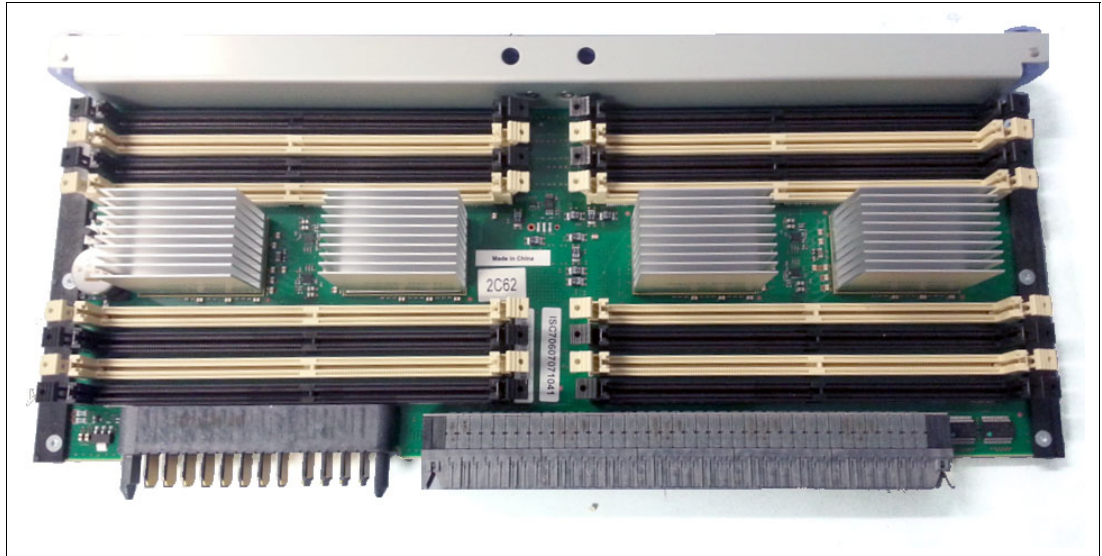


Figure 2-11 Memory riser card without any DIMMs installed

2.2.2 Memory placement rules

For a Power E950 server, each memory feature provides one IS DIMM with the following considerations:

- ▶ A minimum of eight IS DIMMs must be installed per memory riser card, and a minimum of one memory riser card per processor module is required.
- ▶ More memory above the minimum of eight IS DIMMs must also be installed in quantities of eight. So, a memory riser card is populated with either 8 or 16 IS DIMMs.
- ▶ All IS DIMMs on one memory riser card must be the same feature code (FC), which implies that they are all of the same memory capacity.

For the Power E950 server, the following 1600 MHz DDR4 DRAM memory options are available:

- ▶ 8 GB IS DIMM Memory (#EM6A)
- ▶ 16 GB IS DIMM Memory (#EM6B)
- ▶ 32 GB IS DIMM Memory (#EM6C)
- ▶ 64 GB IS DIMM Memory (#EM6D)
- ▶ 128 GB IS DIMM Memory (#EM6E)

Every processor module has one or two memory riser cards that are installed, and different memory features can be used for each of the memory riser cards. Perform the following steps:

1. Start with a minimum of one memory riser card per processor with minimum of eight DIMMs that are installed in the white DIMM connectors. For consistency, install the 'even' riser card first for each processor.
2. The next eight DIMMs are populated in order, that is, CPU0, CPU1, CPU2, and CPU3, while conforming to the rule of the same type and size of DIMMs on each riser card. This processor order is preferred to ensure system configuration consistency; it is not a requirement.

3. After all the even riser cards of the processors are filled, add a second memory riser card to CPU0 with eight IS DIMMs. The second riser card of each processor can have a different IS DIMM capacity from the first riser, but it is preferable that the second riser card has the same type and size of the DIMMs of the first riser card.
4. Fill the second riser card with the next eight IS DIMMS while conforming to the rule of the same type and size of DIMMs on each riser card.
5. Repeat steps 3 and 4 with CPU1, then CPU2, and then CPU3 until all the second riser cards are filled.

The IS DIMMs must be identical in size in at least one memory riser card. However, for optimal performance, the two memory riser cards of each processor module should be the same size by using the same memory FCs. Also, for optimal performance, the amount of memory per processor module should be the same.

The more IS DIMM slots that are filled, the larger the memory bandwidth that is available to the server. If you follow steps 1 on page 62 - 5, the following minimum memory configurations are permitted:

- ▶ The minimum memory that is supported per two POWER9 processors that are installed is 128 GB.
- ▶ The minimum memory that is supported per four POWER9 processors that are installed is 256 GB.

The total memory capacity of a memory riser card is determined by the first eight IS DIMMs FC. Any additional memory capacity that you plan to add after the initial configuration must be the same FC per riser card. Therefore, careful planning for future memory capacity growth must be part of the initial solution design.

2.2.3 Memory activations

Two types of CoD capability are available for processor and memory on the Power E950 server:

Capacity Upgrade on Demand (CUoD)

With CUoD, you can purchase extra permanent processor and memory capacity, and dynamically activate it when you need it.

Elastic Capacity on Demand (Temporary) (Elastic CoD)

With Elastic CoD, you can temporarily activate processors or memory as full-day increments as needed. The charges are based on usage reporting that is collected monthly. The processors and memory can be activated and turned off an unlimited number of times whenever you need extra processing resources.

All memory riser cards and IS DIMMs are capable of CoD. Permanent memory activations are required for at least 50% of the physically installed memory or 128 GB, whichever is larger.

Table 2-11 lists the static memory activation features that are offered for an initial order and any subsequent memory upgrade orders.

Table 2-11 Static memory activation feature codes

Feature code	Feature description	Amount of memory	Type of activation
ELNP	Power IFL Memory Activation	32 GB	Linux
EMAN	128 GB Base Memory activation for MR9 EHC4	128 GB	AIX
EMAP	1 GB Memory activation	1 GB	AIX or Linux
EMAQ	Quantity of 100 1 GB Memory activation	100 GB	AIX or Linux

Elastic CoD for memory is available for memory capacity that is not permanently activated. This CoD type requires a 90 Days Elastic CoD Memory Enablement (#EM9U) feature to be ordered to enable a Power E950 server for Elastic CoD. An Elastic CoD contract must be signed and the participating servers must be registered with IBM before Elastic CoD can be instantiated through #EM9U.

Elastic CoD Memory Days can also be acquired through IBM Marketplace after system installation. For more information about new Elastic CoD features, see the [IBM Digital MarketPlace](#).

For more information about CoD and activation requirements, see 2.3, “Capacity on Demand” on page 68.

2.2.4 Memory throughput

The Power E950 system can be configured with two or four processor modules. Each processor module drives eight memory channels at 9.6 GTps. Any transaction has a 2-byte data read and 1-byte data write simultaneously. Memory bandwidth varies with the workload, but the maximum theoretical memory bandwidth when using IS DIMMs at 1600 Mbps frequency in all 32 slots of a processor module is approximately 230 GBps, as described by the following formula:

$$9.6 \text{ GTps} \times 3 \text{ bytes/channel} \times 8 \text{ channels} = 230.4 \text{ GBps}$$

The total maximum theoretical memory bandwidth of a Power E950 server with two processor modules installed (a 2-socket configuration) is 460.8 GBps. The total maximum theoretical memory bandwidth of a Power E950 server with four processor modules installed (a 4-socket configuration) system is 921.6 GBps.

As data flows from main memory towards the execution units of the POWER9 processor, it passes through the 512 KB L2 and the 64 KB L1 data cache. In many cases, the 10 MB L3 victim cache may also provide the data that is needed for the instruction execution.

Table 2-12 shows the maximum cache bandwidth for a single core as defined by the width of the relevant channels and the related transaction rates on the Power E950 system.

Table 2-12 Power E950 single core architectural maximum cache bandwidth

Cache level of the POWER9 core	Power E950 cache bandwidth ^a			
	3.6 - 3.8 GHz core (#EPWR) [GBps]	3.4 - 3.8 GHz core (#EPWS) [GBps]	3.2 - 3.8 GHz core (#EPWY) [GBps]	3.15 - 3.8 GHz core (#EPWT) [GBps]
L1 64 KB data cache	346 - 365	326 - 365	307 - 365	302 - 365
L2 512 KB cache	346 - 365	326 - 365	307 - 365	302 - 365
L3 10 MB cache	230 - 243	218 - 243	205 - 243	202 - 243

a. Values are rounded to the nearest integer.

The bandwidth figures for the caches are calculated as follows:

- ▶ L1 data cache: In one clock cycle, four 16-byte load operations and two 16-byte store operations can be accomplished. The value varies depending on the core frequency, which is computed as follows:
 - Core running at 3.15 GHz: $(4 \times 16 \text{ B} + 2 \times 16 \text{ B}) \times 3.15 \text{ GHz} = 302.40 \text{ GBps}$
 - Core running at 3.20 GHz: $(4 \times 16 \text{ B} + 2 \times 16 \text{ B}) \times 3.20 \text{ GHz} = 307.20 \text{ GBps}$
 - Core running at 3.40 GHz: $(4 \times 16 \text{ B} + 2 \times 16 \text{ B}) \times 3.40 \text{ GHz} = 326.40 \text{ GBps}$
 - Core running at 3.60 GHz: $(4 \times 16 \text{ B} + 2 \times 16 \text{ B}) \times 3.60 \text{ GHz} = 345.60 \text{ GBps}$
 - Core running at 3.80 GHz: $(4 \times 16 \text{ B} + 2 \times 16 \text{ B}) \times 3.80 \text{ GHz} = 364.80 \text{ GBps}$
- ▶ L2 cache: In one clock cycle, one 64 byte read operation to the core and two 16-byte store operations from the core can be accomplished. The value varies depending on the core frequency, which is computed as follows:
 - Core running at 3.15 GHz: $(1 \times 64 \text{ B} + 2 \times 16 \text{ B}) \times 3.15 \text{ GHz} = 302.40 \text{ GBps}$
 - Core running at 3.20 GHz: $(1 \times 64 \text{ B} + 2 \times 16 \text{ B}) \times 3.20 \text{ GHz} = 307.20 \text{ GBps}$
 - Core running at 3.40 GHz: $(1 \times 64 \text{ B} + 2 \times 16 \text{ B}) \times 3.40 \text{ GHz} = 326.40 \text{ GBps}$
 - Core running at 3.60 GHz: $(1 \times 64 \text{ B} + 2 \times 16 \text{ B}) \times 3.60 \text{ GHz} = 345.60 \text{ GBps}$
 - Core running at 3.80 GHz: $(1 \times 64 \text{ B} + 2 \times 16 \text{ B}) \times 3.80 \text{ GHz} = 364.80 \text{ GBps}$
- ▶ L3 cache: With two clock cycles, one 64 byte read operation to the L2 cache and one 64-byte store operation from the L2 cache can be accomplished. The value varies depending on the core frequency, which is computed as follows:
 - Core running at 3.15 GHz: $(1 \times 64 \text{ B} + 1 \times 64 \text{ B}) \times 3.15 \text{ GHz} / 2 = 201.60 \text{ GBps}$
 - Core running at 3.20 GHz: $(1 \times 64 \text{ B} + 1 \times 64 \text{ B}) \times 3.20 \text{ GHz} / 2 = 204.80 \text{ GBps}$
 - Core running at 3.40 GHz: $(1 \times 64 \text{ B} + 1 \times 64 \text{ B}) \times 3.40 \text{ GHz} / 2 = 217.60 \text{ GBps}$
 - Core running at 3.60 GHz: $(1 \times 64 \text{ B} + 1 \times 64 \text{ B}) \times 3.60 \text{ GHz} / 2 = 230.40 \text{ GBps}$
 - Core running at 3.80 GHz: $(1 \times 64 \text{ B} + 1 \times 64 \text{ B}) \times 3.80 \text{ GHz} / 2 = 243.20 \text{ GBps}$

For a 2-socket Power E950 server that is populated with all its memory IS DIMMs, the overall bandwidths as defined by the width of the relevant channels and the related transaction rates are shown in Table 2-13.

Table 2-13 Power E950 2-processor module maximum cache and memory bandwidth

Memory architecture entity	Power E950 cache and system memory bandwidth for 2-processor module systems ^a			
	16 cores (#EPWR) @ 3.6 - 3.8 GHz [GBps]	20 cores (#EPWS) @ 3.4 - 3.8 GHz [GBps]	22 cores (#EPWY) @ 3.2 - 3.8 GHz [GBps]	24 cores (#EPWT) @ 3.15 - 3.8 GHz [GBps]
L1 64 KB data cache	5,530 - 5,837	6,528 - 7,296	6,758 8,026	7,258 8,755
L2 512 KB cache	5,530 - 5,837	6,528 - 7,296	6,758 8,026	7,258 8,755
L3 10 MB cache	3,686 - 3,891	4,352 - 4,864	4,506 5,350	4,838 5,837
System memory	461	461	461	461

a. Values are rounded to the nearest integer.

For a four-processor module Power E950 system, the accumulated bandwidth values are shown in Table 2-14.

Table 2-14 Power E950 4-processor module maximum cache and memory bandwidth

Memory architecture entity	Power E950 cache and system memory bandwidth for 4-processor module systems ^a			
	32 cores (#EPWR) @ 3.6 - 3.8 GHz [GBps]	40 cores (#EPWS) @ 3.4 - 3.8 GHz [GBps]	44 cores (#EPWY) @ 3.2 - 3.8 GHz [GBps]	48 cores (#EPWT) @ 3.15 - 3.8 GHz [GBps]
L1 64 KB data cache	11,059 - 11,674	13,056 - 14,592	13,517 - 16,051	14,515 - 17,510
L2 512 KB cache	11,059 - 11,674	13,056 - 14,592	13,517 - 16,051	14,515 - 17,510
L3 10 MB cache	7,373 - 7,782	8,704 - 9,728	9,011 - 10,701	9,677 - 11,674
System memory	922	922	922	922

a. Values are rounded to the nearest integer.

2.2.5 Active Memory Mirroring

The Power E950 system can mirror the Power Hypervisor code across multiple memory IS DIMMs. If an IS DIMM that contains the hypervisor code develops an uncorrectable error, its mirrored partner enables the system to continue to operate uninterrupted.

Active Memory Mirroring (AMM) is included by default with all Power E950 systems at an initial order configuration, but you may remove it from the initial order configuration if you want.

The hypervisor code logical memory blocks are mirrored on distinct IS DIMMs to enable more usable memory. There is no specific IS DIMM that hosts the Hypervisor memory blocks, so the mirroring is done at the logical memory block level, not at the IS DIMM level. To enable the AMM feature, the server must have 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), which are responsible for tracking the state of the memory pages that are assigned to partitions
- ▶ Translation control entities (TCEs), which are responsible for providing I/O buffers for the partition's communications
- ▶ Memory that is used by the hypervisor to maintain partition configuration, I/O states, virtual I/O information, and partition state

It is possible to check whether the AMM option is enabled and change its status by using the classical GUI of the HMC opening the CEC Properties window and clicking the **Advanced** tab. If you are using the enhanced GUI of the HMC, the relevant information and controls are in the Memory Mirroring section of the General Settings window of the selected Power E950 system (Figure 2-12).

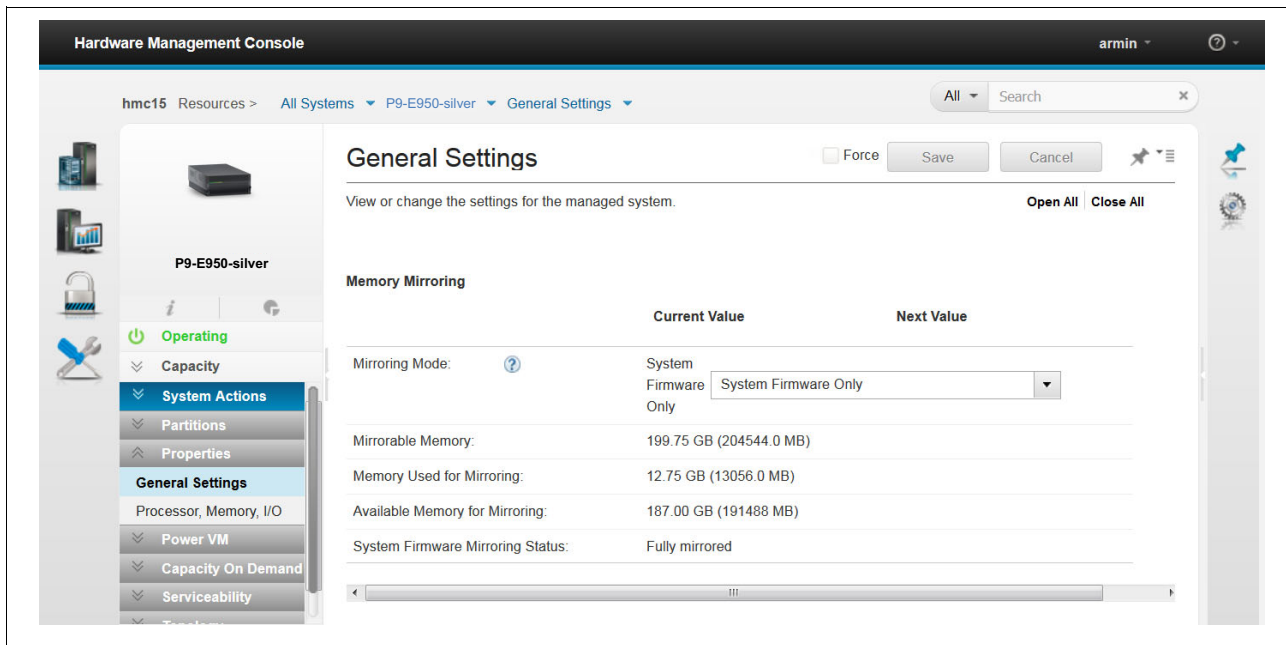


Figure 2-12 Memory Mirroring section in the General Settings window on the HMC enhanced GUI

After a failure on one of the IS DIMMs containing hypervisor data occurs, all the server operations remain active and the Flexible Service Processor (FSP) isolates the failing IS DIMMs. Systems stay in the partially mirrored state until the failing IS DIMM 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 that is used to hold the contents of platform dumps

Partition data: AMM does *not* mirror partition data. It mirrors only the hypervisor code and its components to protect this data against a DIMM failure.

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

2.2.6 Memory error correction and recovery

The memory has error detection, and its correction circuitry is designed so that the failure of any one specific memory module within an error correction code (ECC) can be corrected without any other fault.

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

Other memory protection features include retry capabilities for certain faults that are 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 to be reported to the hypervisor, which supports operating system deallocation of a page that is associated with a hard single-cell fault.

For more information about memory reliability, availability, and serviceability (RAS), see Chapter 4, "Reliability, availability, serviceability, and manageability" on page 123.

2.2.7 Special Uncorrectable Error handling

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

2.3 Capacity on Demand

Several types of CoD offerings are optionally available on the Power E950 server to help meet changing resource requirements in an on-demand environment by using resources that are installed on the system but are not activated. Activation codes are published at [Power Systems Capacity on Demand](#).

The following convention is used in the Order type column in all tables in this section:

Initial	Only available when ordered as part of a new system.
MES	Only available as a Miscellaneous Equipment Specification (MES) upgrade.
Both	Available with a new system or as part of an upgrade.
Supported	Unavailable as a new purchase, but supported when migrated from another system or as part of a model conversion.

2.3.1 Capacity on Demand: New features

Here are some of the new features for CoD:

- ▶ A minimum number of processor cores must be permanently activated. The minimum number of activated cores is one processor's worth. For example, for an 8-core Power E950 server, a minimum of 8 cores must be activated. The activated cores are spread across the processors by the system hypervisor depending on system configurations. More activations are optional and can be purchased later.
- ▶ Permanent CoD memory activations are required for at least 50% of the physically installed memory or 128 GB of activations, whichever is larger.
- ▶ CUoD, Elastic CoD, Utility CoD, and Trial CoD are all available with the Power E950 server.
- ▶ A Power E950 server cannot participate in a Power Enterprise Pool.

2.3.2 Capacity Upgrade on Demand

The Power E950 system includes a number of active processor cores and memory units. They can also include inactive processor cores and memory units. Active processor cores or memory units are processor cores or memory units that are already available for use on your server when it comes from the manufacturer. Inactive processor cores or memory units are processor cores or memory units that are included with your server, but not available for use until you activate them. Inactive processor cores and memory units can be permanently activated by purchasing an activation feature that is called CUoD and entering the provided activation code on the HMC for the server.

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

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

By using CUoD, a company can 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 can be added along with the project, adjusting the hardware platform to meet the project needs so that a company can reduce the initial investment in hardware and acquire only software licenses that are needed on each project phase, reducing the total cost of ownership (TCO) and total cost of acquisition (TCA) of the solution.

Figure 2-13 shows a comparison between two scenarios: a fully activated system versus a system with CUoD resources being activated along with the project timeline.

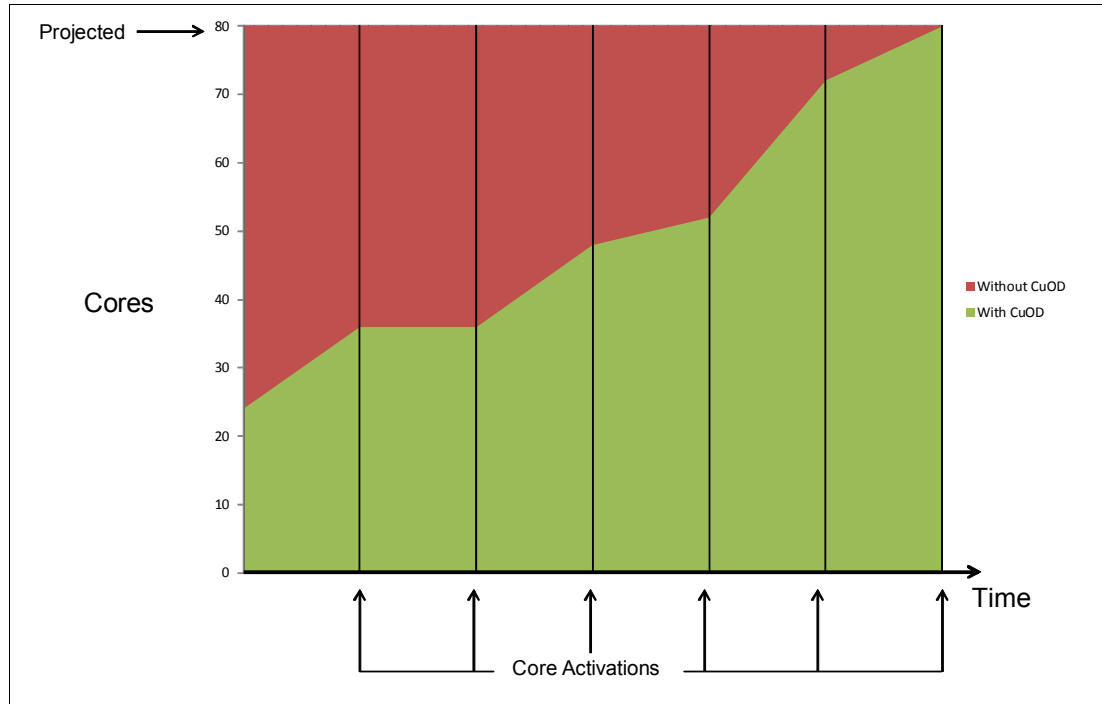


Figure 2-13 Active cores scenarios comparison during a project lifecycle

2.3.3 Processor activations

Table 2-15 lists the processor activation features that are available for the Power E950 server.

Table 2-15 Processor activation features

Feature code	1 core activation feature	Full socket activation for Linux feature
EPWR 8-core 3.6 GHz	EPWV	ELBG - 8 core activations per feature
EPWS 10-core 3.4 GHz	EPWW	ELBP - 10 core activations per feature
EPWT 12-core 3.15 GHz	EPWX	ELBH - 12 core activations per feature
EPWY 11-core 3.15 GHz	EPN3	ELBR - 11 core activations per feature

Table 2-16 lists the memory activation features that are available for the Power E950 server.

Table 2-16 Memory activation features

Feature code	Description	Maximum	Order type ^a
EMAP	1 GB Memory Activation	16384	Both
EMAQ	Quantity of 100 1 GB Memory Activations	163	Both
EMBE	Bundle of 512 1 GB Memory Activations for EMAP/EMAW	4	Both
ELNP	Power IFL Memory Activation 32 GB	12	Both

a. For more information about order types, see 2.3, “Capacity on Demand” on page 68.

2.3.4 Elastic Capacity on Demand (Temporary)

Change of name: Elastic CoD was previously called On/Off CoD. Some websites or documents still refer to Elastic CoD as *On/Off CoD*.

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. When you order an Elastic CoD feature, you receive an enablement code that a system operator uses to make requests for more processor and memory capacity in increments of one processor day or 1 GB memory day. The system monitors the amount and duration of the activations. Both prepaid and post-pay options are available.

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

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

The Power E950 server supports the 90-day temporary Elastic CoD processor and memory enablement features. These features enable the system to activate processor days and GB days equal to the number of inactive resources multiplied by 90 days. Thus, if all the resources are activated by using Elastic CoD, a new enablement code must be ordered every 90 days. If only half of the inactive resources are activated by using Elastic CoD, a new enablement code must be ordered every 180 days.

Before using temporary capacity on your server, you must enable your server. Order an enablement feature (MES only). (The required contracts must be in place.) The 90-day enablement FC for the Power E950 processors is #EP9T. For memory, the enablement FC is #EM9V.

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.

► **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 that are requested, whether the capacity is assigned to partitions or remains in the shared processor pool (SPP).

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 activity occurs). This data is used to determine the proper amount to bill at the end of each billing period (calendar quarter). Failure to report billing data for use of temporary processor or memory capacity during a billing quarter can result in default billing equivalent to 90 processor days of temporary capacity.

For more information about registration, enablement, and usage of Elastic CoD, see [Power Systems Capacity on Demand](#).

Table 2-17 lists the Elastic CoD features that are available for the Power E950 server.

Table 2-17 Elastic CoD features

Feature code	Description	Max	Order ^a type
EP9T	90 Days Elastic CoD Processor Core Enablement	1	MES
MMCB	ECOD GB Memory Day - AIX/Linux	9999	MES
MMCY	ECOD Processor day - AIX/Linux	9999	MES

a. For more information about order types, see 2.3, “Capacity on Demand” on page 68.

2.3.5 Utility Capacity on Demand

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

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

If more 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 that is needed above the base (permanent) level.

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

For more information regarding registration, enablement, and use of Utility CoD, see [IBM Support Planning](#).

2.3.6 Trial Capacity on Demand

A *standard request* for Trial CoD requires you to complete a form that includes contact information and vital product data (VPD) from your Power E950 system with inactive CoD resources.

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

An *exception request* for Trial CoD requires you to complete a form that includes contact information and a VPD from your Power E950 system with inactive CoD resources. An exception request activates all the inactive processors or all inactive memory (or all inactive processors *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, see [Power Systems Capacity on Demand: Trial Capacity on Demand](#).

2.3.7 Software licensing and CoD

For software licensing considerations for the various CoD offerings, see the most recent revision of the [Power Systems Capacity on Demand User's Guide](#).

2.3.8 Solution Edition for Healthcare

The Power E950 Solution Edition for Healthcare product provides a cost-effective 48-core/512 GB processor and memory activation feature package for eligible healthcare industry clients running approved independent software vendor (ISV) applications, for example, Epic. The Power E950 Solution Edition for Healthcare product's minimum requirement is a server with 4 typical 3.15 - 3.8 GHz processor modules, 48 cores (all active), and 512 GB memory (all active). More hardware components can be added as wanted by following normal supported configuration rules.

Table 2-18 lists the processor and memory features that are required for the Power E950 Solution Edition for Healthcare product.

Table 2-18 Solution Edition for Healthcare required features

Feature code	Description	Number required
EMAN	128 GB Base Memory activation for MR9 EHC4	4
ELAN	Base Processor Activation (12) for Healthcare Solution (EHC4)	4
EHC4	Solution Edition for Healthcare typical 3.15 to 3.8 GHz, 12-core Processor Module	4

For eligibility rules and registration of the Power E950 Solution Edition for Healthcare through the sales channel, see [IBM Power Solution Editions](#).

The Power E950 Solution Edition for Healthcare product is available to purchase only in the United States.

2.4 System buses

This section provides more information that is related to the internal buses.

2.4.1 PCIe Gen4

The internal I/O subsystem on the Power E950 server is connected to the PCIe controllers (PECs) on a POWER9 processor in the system. Each POWER9 processor module has three PCIe host bridges (PHBs). Two of the PHBs (PHB0 and PHB2) connect directly to two PCIe Gen4 x16 slots. The third PHB on each POWER9 processor is used for other I/O connections:

- ▶ Four Internal Non-Volatile Memory Express (NVMe) solid-state drive (SSDs)
- ▶ Four USB ports
- ▶ Two PCIe Gen4 x8 slots
- ▶ One PCIe Gen3 x8 slot that is reserved for the initial local area network (LAN) adapter

Bandwidths for the connections are shown in Table 2-19.

Table 2-19 Internal I/O connection speeds

Connection	Locations	Type	Speed	Use
PCIe adapter slot	C2 - C5, C7, C8, C10, and C11	PCIe Gen4 x16	64 GBps	General use
PCIe adapter slot	C6	PCIe Gen3 x8	16 GBps	Reserved for initial LAN adapter
PCIe adapter slot	C9 and C12	PCIe Gen4 x8	32 GBps	Serial Attached SCSI (SAS) adapters for internal drives or general use

Connection	Locations	Type	Speed	Use
NVMe slot	NVMe1 - NVMe4	PCIe Gen3 x4	8 GBps	NVMe SSDs
USB controller	Integrated	PCIe Gen2 x1	1 GBps	USB DVD or general use

A diagram showing the connections is shown in Figure 2-14.

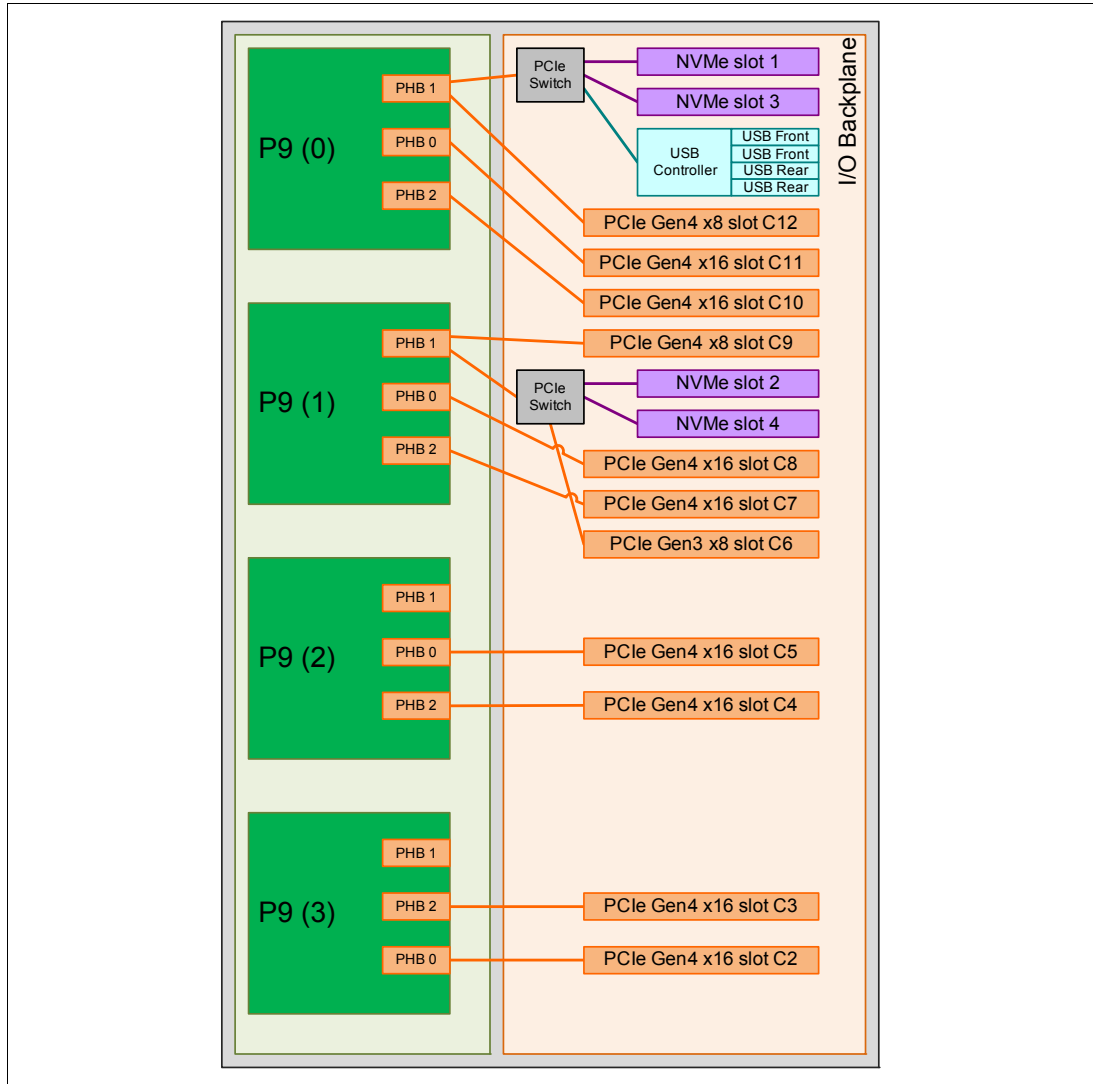


Figure 2-14 Power E950 internal I/O

The Power E950 2-socket configuration supports seven PCIe slots, four NVMe slots, and four USB ports with the minimum two processors sockets populated. Two more PCIe slots become available with each additional populated socket to a maximum of 11 slots.

The Power E950 server supports the connection of up to two PCIe Gen3 I/O Expansion Drawers. Each expansion drawer can provide up to 12 more slots, and connects to the server with two #EJ08 adapters that use a PCIe Gen4 x16 slot for a net addition of 10 PCIe slots.

All of the PCIe connectors are physically x16. However, slot C6, C9, and C12 are logically x8.

For a list of adapters and their supported slots, see 2.5, “PCIe adapters” on page 76.

For more information about PCIe Expansion Drawers, see 2.7.1, “PCIe Gen3 I/O Expansion Drawer” on page 88.

2.5 PCIe adapters

This section covers the type and functions of the PCIe adapters that are supported by the Power E950 system.

The following convention is used in the Order type column in all tables in this section:

Initial	Only available when ordered as part of a new system
MES	Only available as an MES upgrade
Both	Available with a new system or as part of an upgrade
Supported	Unavailable as a new purchase, but supported when migrated from another system or as part of a model conversion

2.5.1 New PCIe adapter features

The following features are new for the PCIe adapters:

- ▶ The Power E950 server supports PCIe Gen4 adapters in internal slots.
- ▶ PCIe Gen3 adapters are supported both internally and in the PCIe Gen3 I/O Expansion Drawer.

2.5.2 PCIe

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

The PCIe interfaces supported on these system nodes are PCIe Gen4, capable of 32 Gbps simplex (64 Gbps duplex) on a single x16 interface. PCIe Gen4 slots also support previous generations (Gen3, Gen2, and Gen1) adapters, which operate at lower speeds, if you place x1, x4, x8, and x16 speed adapters in the same connector size slots first before mixing adapter speeds with connector slot size.

The Power E950 server also supports expansion beyond the slots that are available in the system nodes by attaching one or more PCIe Gen3 I/O Expansion Drawers (#EMX0).

All slots in the Power E950 server and the PCIe Gen3 I/O Expansion Drawer are full-height adapters. Low-profile adapters are not supported in the Power E950 server.

Before adding or rearranging adapters, use the [IBM System Planning Tool](#) to validate the new adapter configuration.

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, see [IBM Prerequisites](#).

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.

2.5.3 LAN adapters

Table 2-20 lists the available LAN adapters that are supported in the Power E950 server.

Table 2-20 Available LAN adapters

Feature code	CCIN	Description	OS support	Order type ^a
EN0W	2CC4	PCIe2 2-port 10/1 GbE BaseT RJ45 Adapter	AIX and Linux	Both
EN0U	2CC3	PCIe2 4-port (10 Gb+1 GbE) Copper SFP+RJ45 Adapter	AIX and Linux	Both
EN0S	2CC3	PCIe2 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter	AIX and Linux	Both
5899	576F	PCIe2 4-port 1 GbE Adapter	AIX and Linux	Both
EC2S	58FA	PCIe3 2-Port 10 Gb NIC&ROCE SR/Cu Adapter	AIX and Linux	Both
EC2U	58FB	PCIe3 2-Port 25/10 Gb NIC&ROCE SR/Cu Adapter	AIX and Linux	Both
EN0K	2CC1	PCIe3 4-port (10 Gb FCoE & 1 GbE) SFP+Copper&RJ45	AIX and Linux	Both
EN0H	2B93	PCIe3 4-port (10 Gb FCoE & 1 GbE) SR&RJ45	AIX and Linux	Both
EN17	2CE4	PCIe3 4-port 10 GbE SFP+ Copper Adapter	AIX and Linux	Both
EN15	2CE3	PCIe3 4-port 10 GbE SR Adapter	AIX and Linux	Both
EC66	2CF3	PCIe4 2-port 100 Gb ROCE EN adapter	AIX and Linux	Both

a. For more information about order types, see 2.5, "PCIe adapters" on page 76.

2.5.4 Graphics adapters

There are currently no graphics adapters that are supported for the Power E950 server.

2.5.5 SAS adapters

Table 2-21 lists the SAS adapters that are available for the Power E950 server.

Table 2-21 Available SAS adapters

Feature code	CCIN	Description	OS support	Order type ^a
EJ0J	57B4	PCIe3 RAID SAS Adapter Quad-port 6 Gb x8	AIX and Linux	Supported
EJ0K	57B4	PCIe3 RAID SAS Adapter Quad-port 6 Gb x8 for MR9	AIX and Linux	Both
EJ10	57B4	PCIe3 SAS Tape/DVD Adapter Quad-port 6 Gb x8	AIX and Linux	Both
EJ14	57B1	PCIe3 12 GB Cache RAID PLUS SAS Adapter Quad-port 6 Gb x8	AIX and Linux	Both

a. For more information about order types, see 2.5, "PCIe adapters" on page 76.

2.5.6 Fibre Channel adapters

Table 2-22 lists the Fibre Channel adapters that are available for the Power E950 server.

Table 2-22 Available Fibre Channel adapters

Feature code	CCIN	Description	OS support	Order type ^a
5729	5729	PCIe2 8 Gb 4-port Fibre Channel Adapter	AIX and Linux	Both
5735	577D	8 Gb PCI Express Dual Port Fibre Channel Adapter	AIX and Linux	Both
EN0A	577F	PCIe3 16 Gb 2-port Fibre Channel Adapter	AIX and Linux	Both
EN12	N/A	PCIe2 8 Gb 4-port Fibre Channel Adapter	AIX and Linux	Both
EN1A	578F	PCIe3 32 Gb 2-port Fibre Channel Adapter	AIX and Linux	Both
EN1C	578E	PCIe3 16 Gb 4-port Fibre Channel Adapter	AIX and Linux	Both

a. For more information about order types, see 2.5, "PCIe adapters" on page 76.

2.5.7 USB ports

The Power E950 server has four USB ports. Two ports are at the front of the system and two are at the rear. The front ports are USB 3.0 and capable of supplying 1.5 A per port. The rear ports are USB 3.0 and capable of supplying 0.9 A per port.

Figure 2-15 shows the location of the USB ports on the Power E950 system.

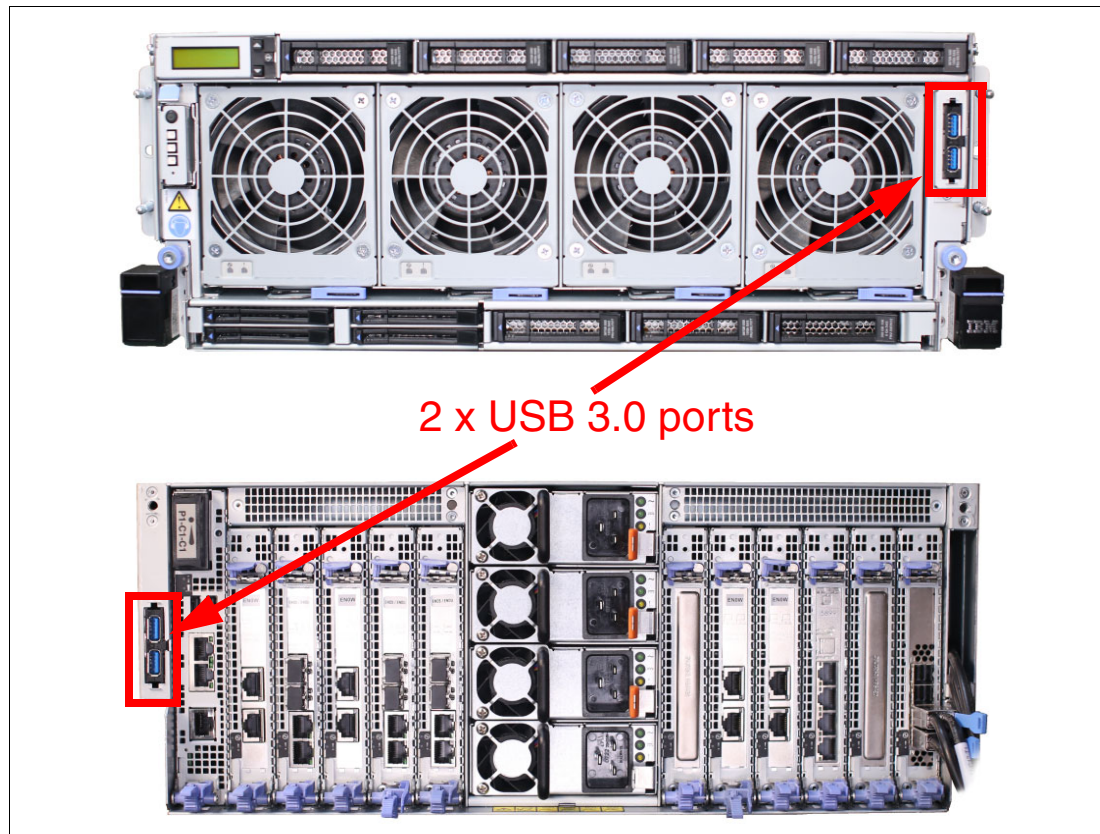


Figure 2-15 The Power E950 server USB ports

All USB ports on the Power E950 server can function with any USB device that is supported by the client operating system to which the adapter is assigned.

There are currently no USB PCIe adapters that are supported in the Power E950 server.

2.5.8 InfiniBand host channel adapters

There are currently no InfiniBand PCIe adapters that are supported in the Power E950 server.

2.5.9 Cryptographic Coprocessor

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

The IBM PCIe Cryptographic Coprocessor adapter highlights the following features:

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

For the most recent firmware and software updates, see IBM CryptoCards [IBM CryptoCards](#).

Table 2-23 lists the cryptographic adapter that is available for the server.

Table 2-23 Available cryptographic adapters

Feature code	CCIN	Description	OS support	Order type ^a
EJ33	4767	PCIe3 Crypto Coprocessor BSC-Gen3 4767	AIX and Linux	Both

a. For more information about order types, see 2.5, “PCIe adapters” on page 76.

2.5.10 CAPI adapters

There are currently no CAPI adapters that are supported in the Power E950 server.

2.6 Internal storage

The Power E950 server supports up to four internal NVMe U.2 (2.5 inch 7 mm small form factor (SFF)) SSDs. Each NVMe drive contains its own controller and connect to the system by using PCIe Gen3 protocols. Optionally, the Power E950 server supports up to eight SFF (2.5 inch) SAS bays for hard disk drives (HDDs) or SSDs. The locations of the internal disks and NVMe SSDs are shown in Figure 2-16.

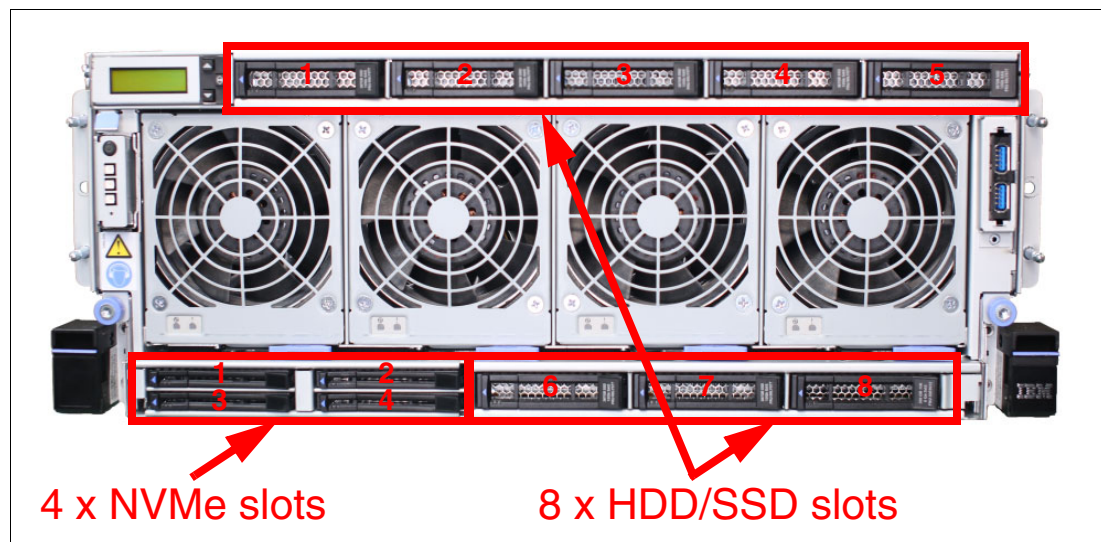


Figure 2-16 Power E950 server with disk and NVMe locations highlighted

2.6.1 Backplane features

The Power E950 server supports up to four internal NVMe SSD drives and up to eight SAS drives. One backplane is required in each Power E950 server. The backplanes that are available on the Power E950 server are shown in Table 2-24.

Table 2-24 Available backplanes

Feature code	CCIN	Description	Drive support	PCIe SAS adapters required
EJOB	2D37	Storage Backplane with Zero DASD	4 x NVMe 0 x SAS	0
EJBB	2D37	Storage Backplane Base DASD 8 SAS 2.5" HDD/SDD bays	4 x NVMe 8 x SAS	One in slot C12
EJSB	2D37	Storage Backplane Split DASD 8 SAS 2.5" HDD/SDD bays	4 x NVMe 4 + 4 x SAS	One in slot C12 One in slot C9

All backplanes are available with an initial order or as an MES upgrade.

Note: To control the internal SAS drives, #EJ0K must be used and must be in slot C12 or slots C12 and C9, depending on the direct access storage device (DASD) backplane feature that is installed.

The NVMe drives are not by the SAS adapters. If #EJBB is selected, a SAS adapter feature (#EJ0K) must be installed in slot C12, which drives all eight SAS drives. If #EJSB is selected, two SAS adapters (#EJ0K) must be installed, one in slot C9 and a second in slot C12. Each adapter drives four SAS drives.

The connections between adapters and disks for #EJSB are shown in Figure 2-17.

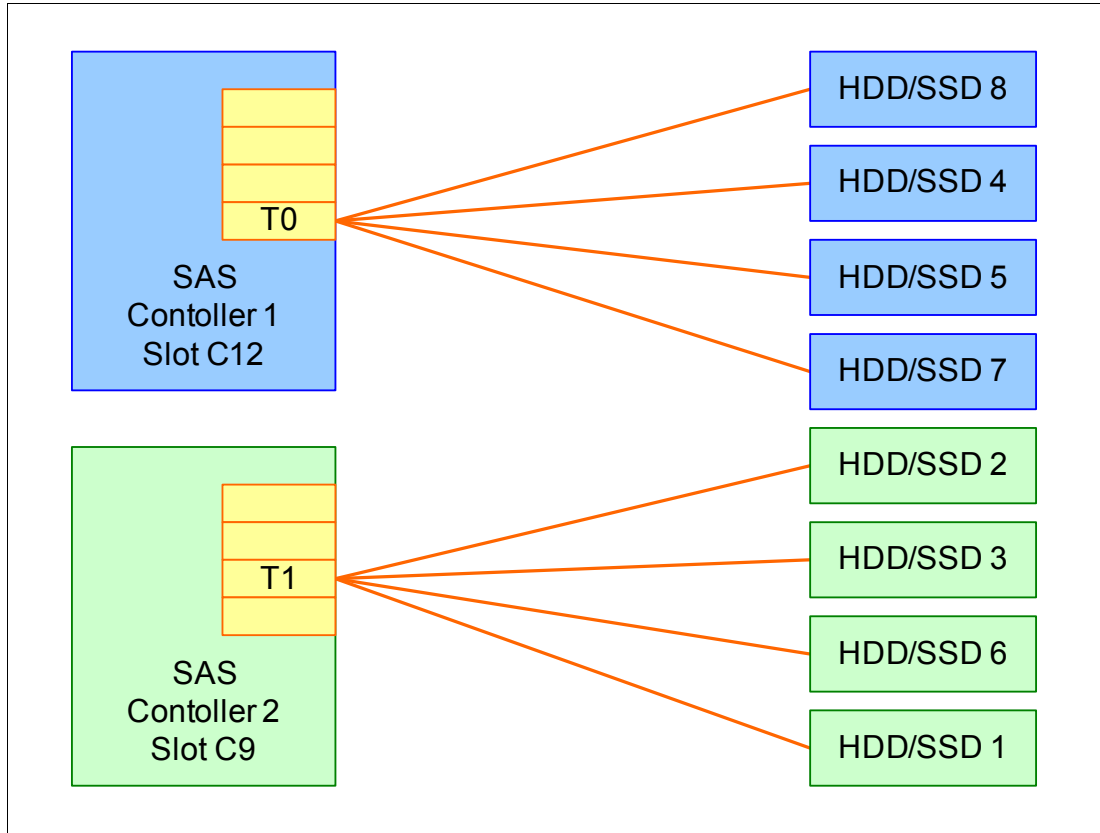


Figure 2-17 Internal SAS connections for #EJSB

Physical HDD/SSD locations are shown in Figure 2-18.

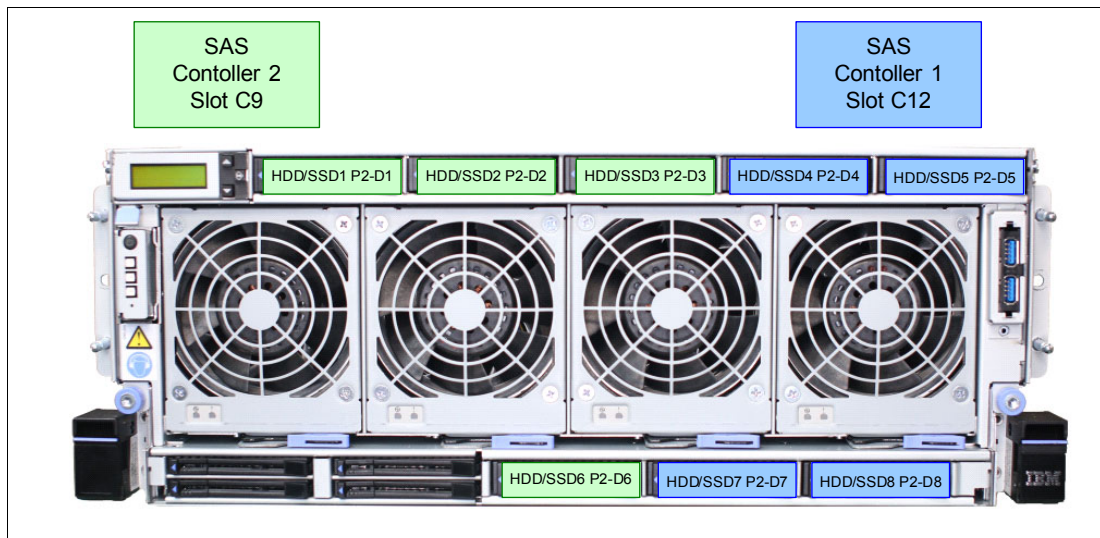


Figure 2-18 The Power E950 server with the HDD/SSD locations highlighted

2.6.2 Internal NVMe SSD drives

The internal NVMe SSD drives are intended for boot purposes only and not as general-purpose drives.

Table 2-25 shows the available internal NVMe SSD drives.

Table 2-25 Available internal NVMe SSD features

Feature code	CCIN	Description	OS support	Order type ^a
EC5J	59B4	Mainstream 800 GB SSD NVMe U.2 module	AIX and Linux	Both
EC5K	59B5	Mainstream 1.6 TB GB SSD NVMe U.2 module	AIX and Linux	Both
EC5L	59B6	Mainstream 3.2 TB GB SSD NVMe U.2 module	AIX and Linux	Both

a. For more information about order types, see 2.5, “PCIe adapters” on page 76.

2.6.3 Supported RAID functions

The SAS adapters that drive disks in the #EJBB and #EJSB backplanes offer RAID 0, 5, 6, and 10. When more features are configured, the server supports hardware RAID 0, 5, 6, 10, 5T2, 6T2, and 10T2, depending on the storage adapter feature:

- ▶ 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 and provides T10 Data Integrity Fields.
- ▶ 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 specific 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 and provides T10 Data Integrity Fields.
- ▶ 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 more 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 sustains data parity during the rebuild process and provides T10 Data Integrity Fields.
- ▶ RAID 10 is a striped set of mirrored arrays.
It is a combination of RAID 0 and RAID 1. RAID 10 uses mirrored pairs to redundantly store data. The array must contain an even number of disks. Two is the minimum number of disks that is needed to create a RAID 10 array. The data is striped across the mirrored pairs.

RAID 10 can tolerate multiple disk failures. If one disk in each mirrored pair fails, the array still functions by operating in Degraded mode. You can continue to use the array normally because for each Failed disk, the data is stored redundantly on its mirrored pair. However, if both members of a mirrored pair fail, the array is placed in the Failed state and is not accessible.

When a RAID 10 disk array is created, the controller automatically attempts to select the disks for each mirrored pair from a different controller connector (a different cable to a different device enclosure). For example, if four disks that are selected for the disk array are on one of the controller connectors and another four disks selected are on another of the controller's connectors, the controller automatically attempts to create each mirrored pair from one disk on each controller connector. In the event of a controller port, cable, or enclosure failure, each mirrored pair continues to operate in a Degraded mode. Such redundancy requires careful planning when you are determining where to place devices.

RAID 5T2, RAID 6T2, and RAID 10T2 are RAID levels with IBM Easy Tier® enabled. They require that both types of disks exist on the system under the same controller (HDDs and SSDs), and that both types are configured under the same RAID type.

2.6.4 Easy Tier

Both HDDs and SSDs may be in the same array if they are using the Easy Tier function. If the Easy Tier function is not being used, then HDDs and SSDs cannot be mixed on a single array.

When the SDDs 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). This is called *Easy Tier*.

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*.

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 2-19 shows an Easy Tier array.

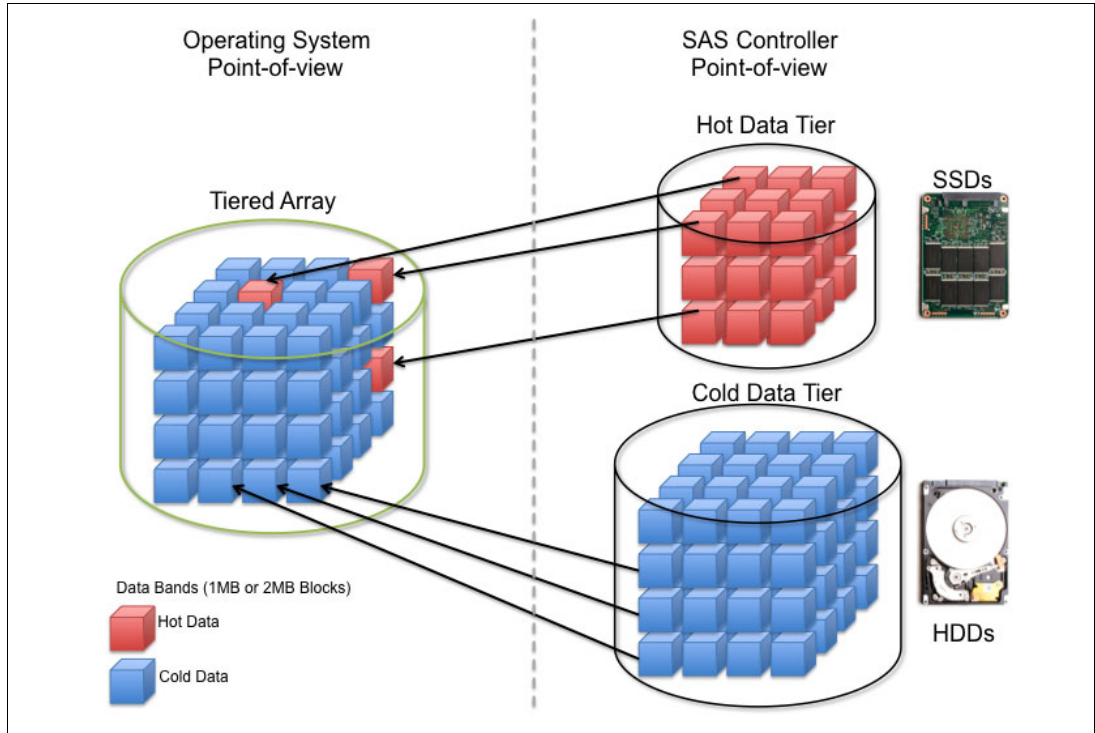


Figure 2-19 Easy Tier array

The Easy Tier configuration is accomplished through a standard operating system SAS adapter configuration utility. Figure 2-20 and Figure 2-21 on page 86 show two examples of tiered array creation for AIX.

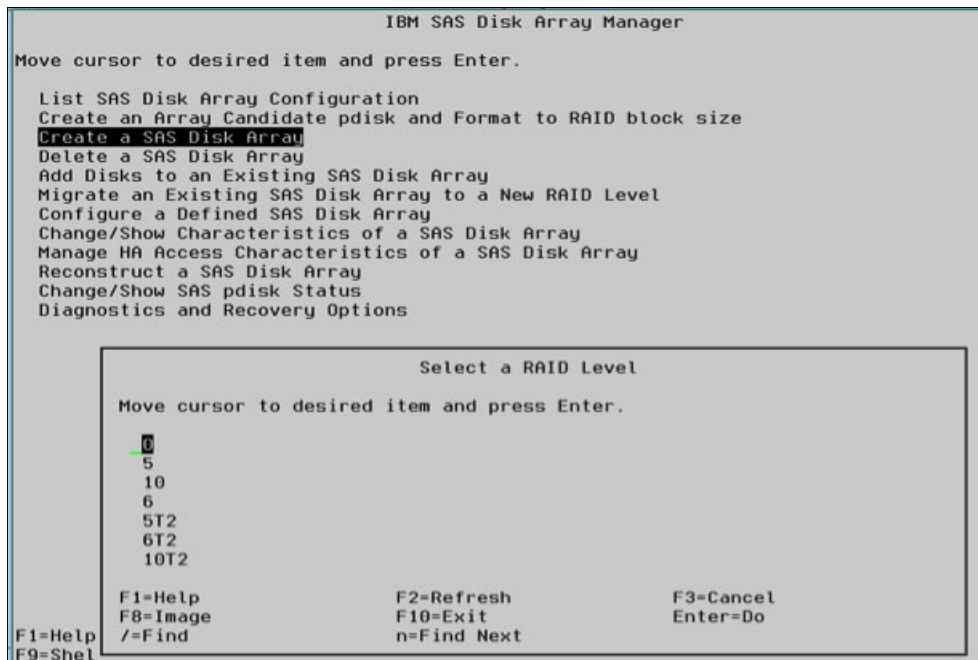


Figure 2-20 Array type selection screen on AIX RAID Manager

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	← RAID 5T2
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	← RAID 6T2
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 (0/0)	666.6GB	← RAID 10T2
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 2-21 Tiered arrays (RAID 5T2, RAID 6T2, and RAID 10T2) example on AIX RAID Manager

To support Easy Tier, make sure that the server is running at least the following minimum levels:

- ▶ Virtual I/O Server (VIOS) 2.2.3.3 with interim fix IV56366 or later
- ▶ AIX 7.1 TL3 SP3 or later
- ▶ AIX 6.1 TL9 SP3 or later
- ▶ Red Hat Enterprise Linux 6.5 or later
- ▶ SUSE Linux Enterprise Server 11 SP3 or later

2.6.5 Media drawers

IBM multimedia drawers, such as the 7226-1U3 or 7214-1U2, or tape units, such as the TS2240, TS2340, TS3100, TS3200, and TS3310, can be connected by using external SAS ports.

2.6.6 External DVD drives

There is a trend to use good quality USB flash drives rather than DVD drives. Being mechanical, DVD drives are less reliable than solid-state technology.

If you feel that you do need a DVD drive, IBM offers a stand-alone external USB unit (#EUA5), which is shown in Figure 2-22.



Figure 2-22 #EUA5: Stand-alone USB DVD drive with cable

Note: If you use an external/stand-alone USB drive that does not have its own power supply, you should use a USB socket at the front of the system to ensure that enough current is available.

2.6.7 RDX removable disk drives

Power Systems servers support RDX removable disk drives, which are commonly used for quick backups. The available RDX unit for the Power E950 server is an external/stand-alone RDX unit (#EUA4). #EUA4 is available for purchase only in the United States.

Various disk drives are available, as shown in Table 2-26.

Table 2-26 RDX disk drives

Feature code	Part number	Description
1107	46C5379	500 GB Removable Disk Drive
EU01	46C2335	1 TB Removable Disk Drive
EU2T	46C2975	2 TB Removable Disk Drive

2.7 External I/O subsystems

This section describes the PCIe I/O Expansion Drawer (#EMX0) that can be attached to the Power E950 system.

2.7.1 PCIe Gen3 I/O Expansion Drawer

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

The PCIe Gen3 I/O Expansion Drawer is a 4U high, PCIe Gen3-based, and 19-inch rack mountable I/O drawer. It offers two PCIe FanOut Modules (#EMXG), each of them providing six full-length, full-height PCIe slots, which are labeled C1 - C6. Slots C1 and C4 are x16 slots, and C2, C3, C5, and C6 are x8 slots. Slots C1 and C4 of the 6-slot fan-out module in a PCIe Gen3 I/O drawer are Single Root I/O Virtualization (SR-IOV) enabled.

An #EMX0 drawer can be configured with one or two #EMXG fan-out modules. Adding a second fan-out module is not a hot-plug operation and requires scheduled downtime.

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 x16 PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ08) adapter and two CXP cables are required to connect a system node to a PCIe fan-out module in the I/O expansion drawer.

Concurrent repair and addition or removal of PCIe adapters 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 that is used with the previous generation server's #5802/5803/5877/5873 12X attached I/O drawers.

As stated before, the maximum number of PCIe Gen3 I/O drawers depends on the number of installed processor modules. A Power E950 system configuration with two processor modules supports two PCIe Gen3 I/O drawers, and a configuration with four processor modules supports four PCIe Gen3 I/O drawers. This delivers an additional maximum of 24 PCIe Gen3 slot capacity for the two processor module configuration, and an additional maximum of 48 PCIe Gen3 slots for the four processor module configuration.

Note: At the initial availability date August 17, 2018, the number of supported PCIe I/O Expansion Drawers (#EMX0) is limited to a maximum of two.

Note: When the operating system is Linux, PowerVM is required for support of the I/O Expansion Drawers.

Each fan-out module must be attached through one optical cable adapter (#EJ08) that is installed in one x16 PCIe Gen4 slot of the Power E950 server. Therefore, a two processor module Power E950 configuration with two I/O drawers that are attached provides a maximum of 31 adapter slots per server, and a four processor module Power E950 configuration with four I/O drawers that are attached provides a maximum of 59 free adapter slots per server.

Table 2-27 lists the maximum PCIe Gen3 I/O drawer configurations for the Power E950 server that are supported by the system's design.

Table 2-27 Maximum PCIe Gen3 I/O drawer configurations that are supported by design.

Power E950 configuration	Maximum number of attached PCIe Gen3 I/O drawers	Maximum number of PCIe slots on the server
Power E950 server with two processor modules	2	31
Power E950 server with four processor modules	4	59

Figure 2-23 shows the back view of the PCIe Gen3 I/O Expansion Drawer.

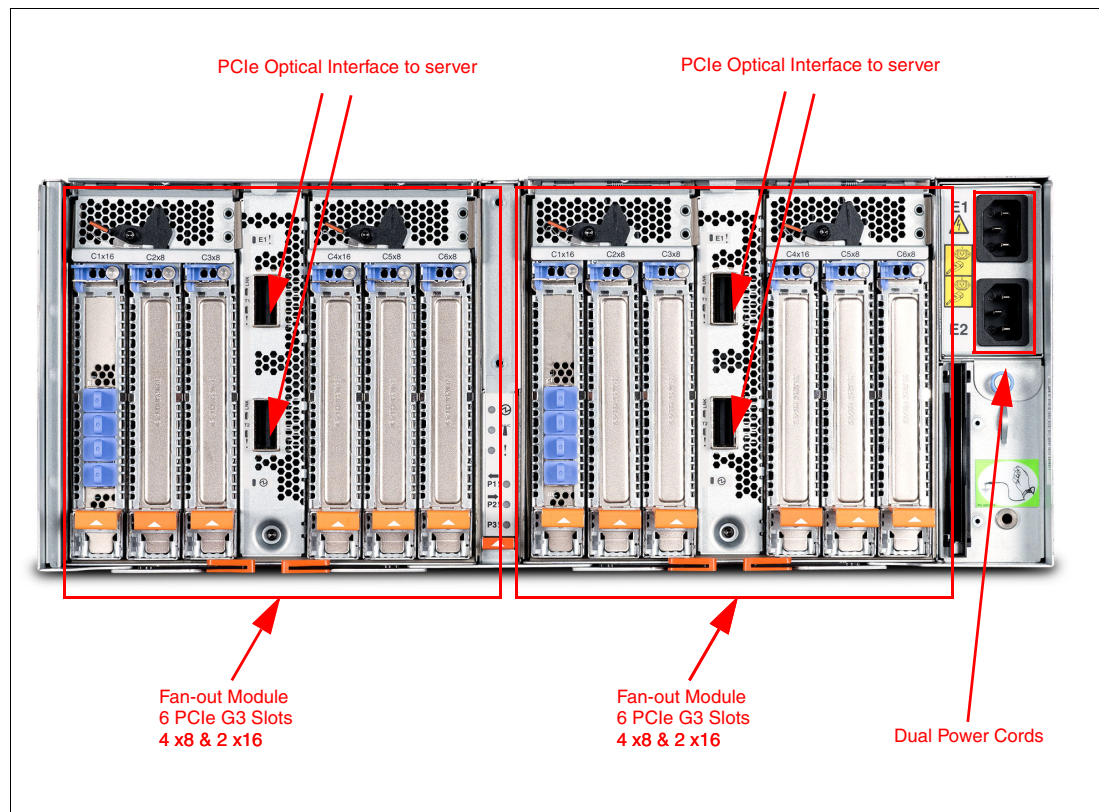


Figure 2-23 Rear view of the PCIe Gen3 I/O Expansion Drawer

PCIe Gen1, Gen2, and Gen3 full-high adapters are supported. The set of full-high PCIe adapters that are supported is found in the Sales Manual, and are identified by feature number. For details and rules that are associated with the specific adapters that are supported and their supported placement in x8 or x16 slots, see [IBM Knowledge Center](#).

2.7.2 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 Copper CXP Cable Pair for PCIe3 Expansion Drawer (#ECCS)
- ▶ 3.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC7)
- ▶ 10.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC8)

Recommendation: Place any attached PCIe Gen3 I/O Expansion Drawer in the same rack as the POWER9 server for ease of service. The drawers may be installed in separate racks if the application or other rack content requires it. As a preferred practice, use 3 m cables for PCIe drawers in the same rack as the system unit and 10 m cables for drawers in a different rack.

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 adapter. Likewise, the bottom two ports must be cabled together:

1. Connect an Active Optical Cable (AOC) 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 FanOut 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 FanOut module in your expansion drawer.
5. Repeat steps 1 - 4 for the other PCIe3 6-slot FanOut module in the expansion drawer, if required.

Figure 2-24 shows connector locations for the PCIe Gen3 I/O Expansion Drawer.

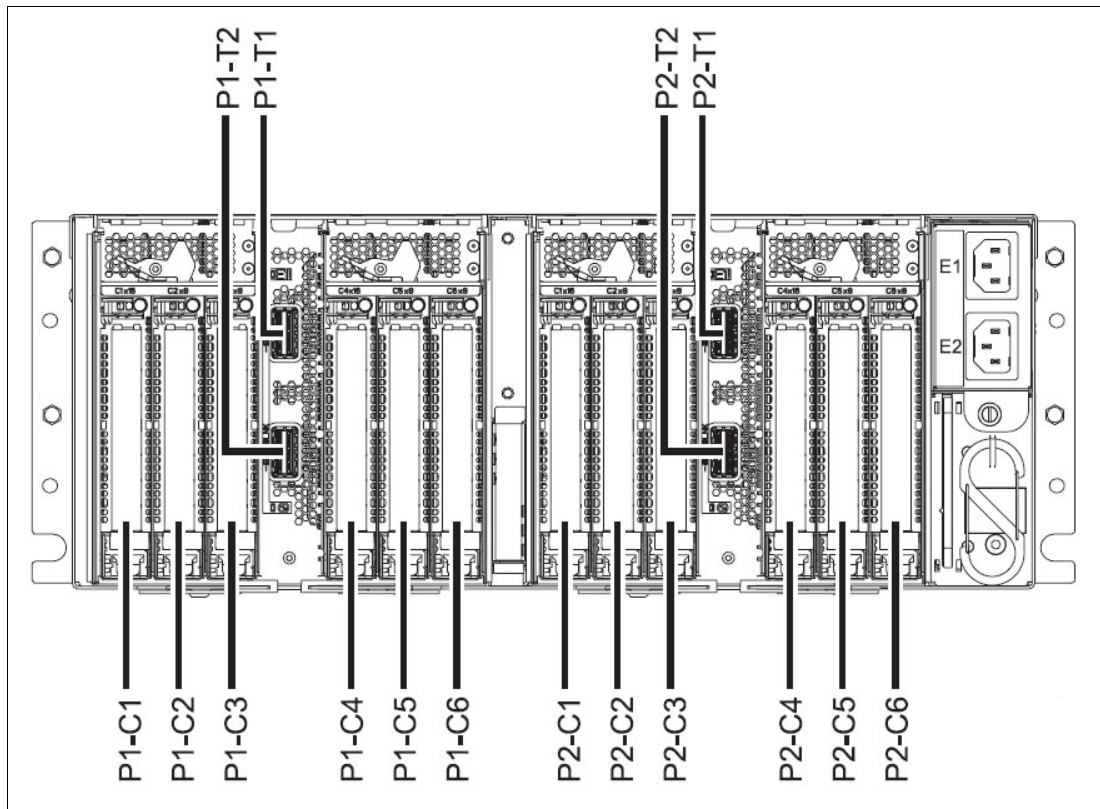


Figure 2-24 Connector locations for the PCIe Gen3 I/O Expansion Drawer

Figure 2-25 shows typical optical cable connections.

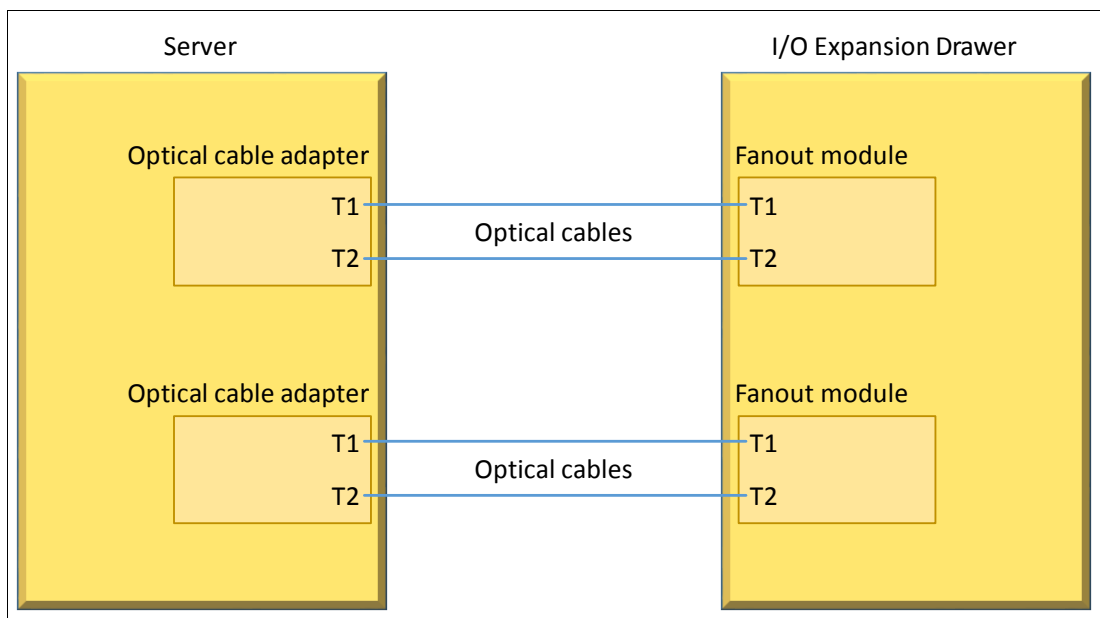


Figure 2-25 Typical optical cable connection

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 E950 system node. However, it is a preferred practice to use the PCIe adapter slot priority information while selecting slots for installing a PCIe3 Optical Cable Adapter (#EJ08).

Table 2-28 shows the PCIe adapter slot priorities in the Power E950 server for #EJ08.

Table 2-28 PCIe adapter slot priorities for PCIe3 Optical Cable Adapter (#EJ08)

Feature code	Description	Slot priorities	
		Two processor modules	Four processor modules
EJ08	PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer	11, 8, 10, and 7 ^a	11, 8, 5, 3, 10, 7, 4, and 2 ^a

a. For information about how the slot numbers that are listed relate to physical location codes, see Figure 2-24 on page 91.

The following figures show configurations that are supported. For simplification, we have not shown every possible combination of the I/O expansion drawer fan-out modules to server attachments.

Figure 2-26 shows an example of a Power E950 server with two processor modules and a maximum of two PCIe Gen3 I/O Expansion Drawers. This configuration is fully supported at the initial availability date of the Power E950 server on 17 August 2018.

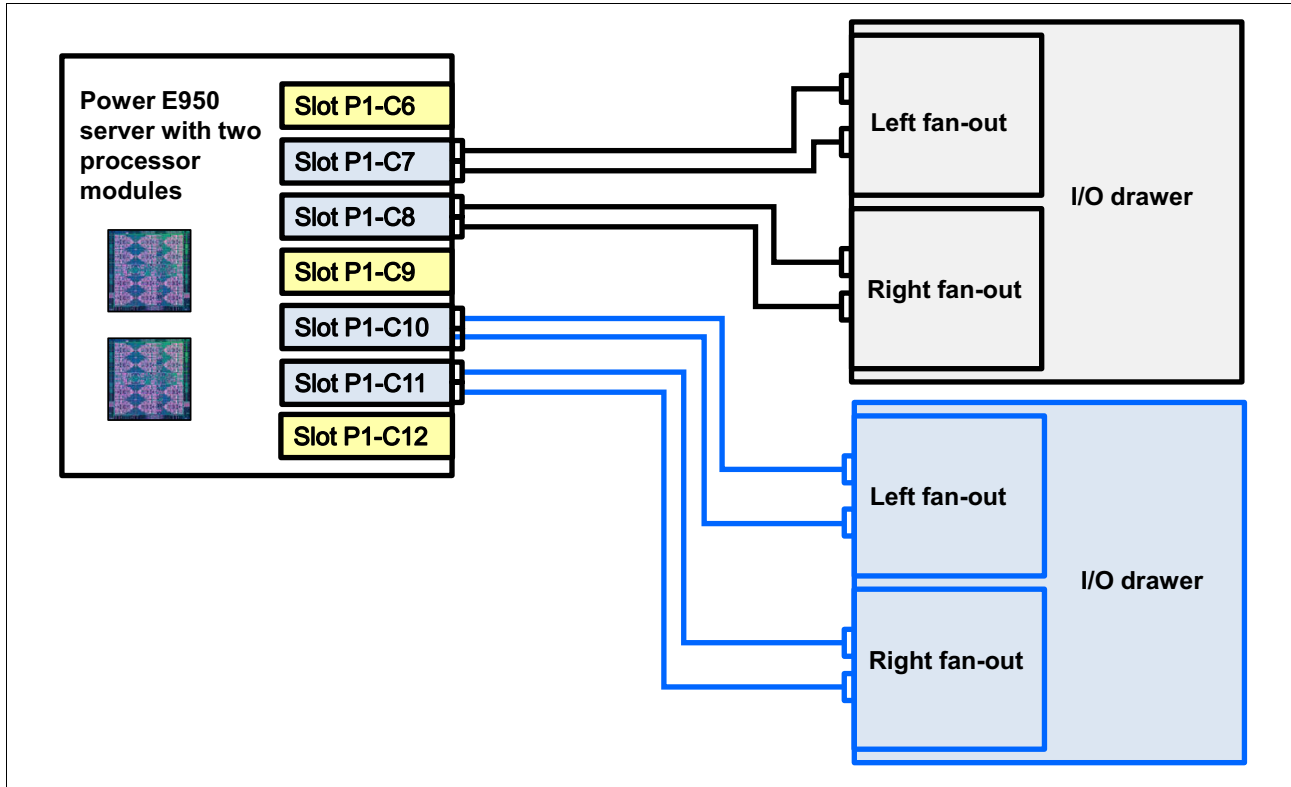


Figure 2-26 Example of a Power E950 server with two processor modules and a maximum of two I/O drawers

Figure 2-27 shows an example of a Power E950 server with four processor modules and a maximum of four PCIe Gen3 I/O Expansion Drawers. At the initial availability date of 17 August 2018, the number of supported PCIe I/O Expansion Drawers (#EMX0) is limited to a maximum of two, but the configuration that is shown in Figure 2-27 has been included for future reference.

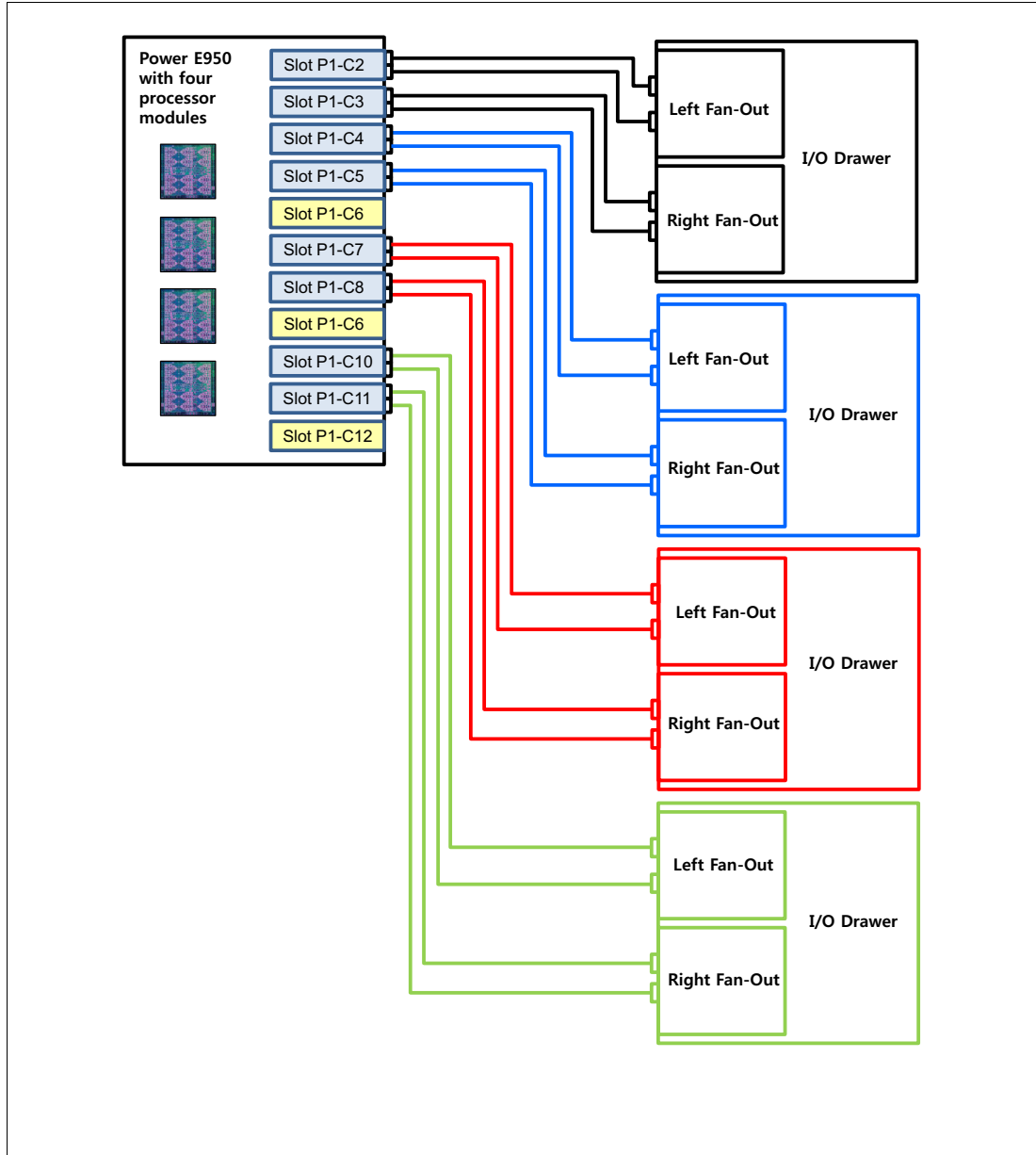


Figure 2-27 Example of a Power E950 server and a maximum of four PCI Gen3 I/O expansion drawers

2.7.3 PCIe Gen3 I/O Expansion Drawer system power control network cabling

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

2.8 External disk subsystems

This section describes the following external disk subsystems that can be attached to the Power E950 server:

- ▶ EXP24S SFF Gen2-bay Drawer for High-density Storage (#5887)
- ▶ EXP24SX and EXP12SX SAS Storage Enclosures (#ESLS)
- ▶ EXP12SX SAS Storage Enclosures (#ESLL)
- ▶ IBM System Storage

Note: The EXP24S is support only (at a time after the initial general availability (GA)) and cannot be ordered together with the Power E950 server.

2.8.1 EXP24S SFF Gen2-bay Drawer

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

Note: The EXP24S is support only (at a time after the initial GA) and cannot be ordered together with the Power E950 server.

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

With AIX and 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 2-28 shows the front of the unit and the groups of disks on each mode.

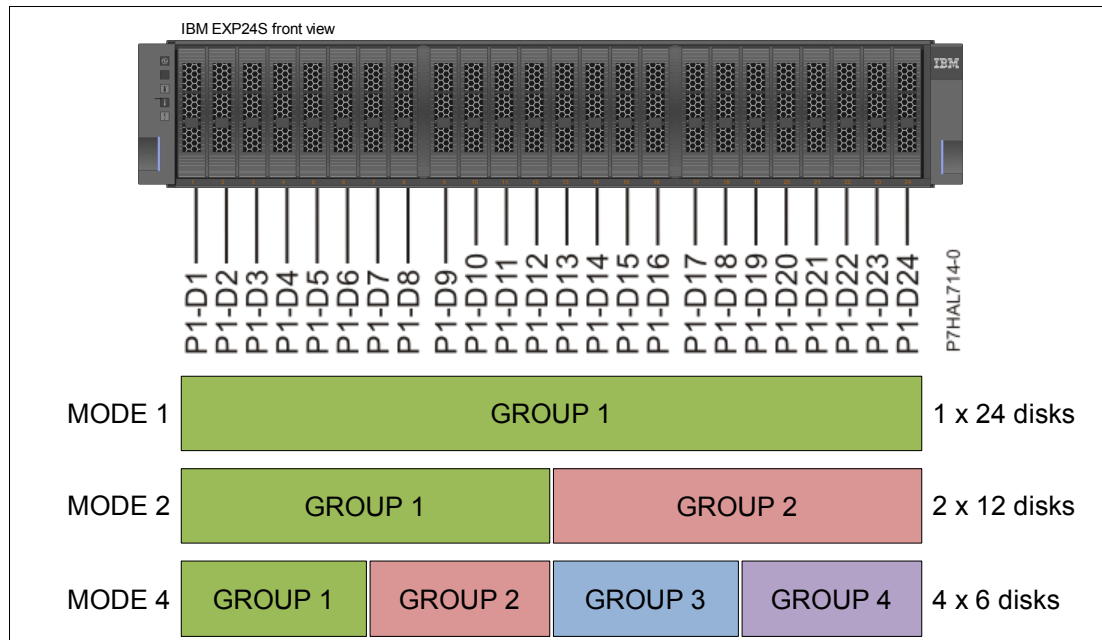


Figure 2-28 EXP24S front view with location codes and disk groups depending on its mode of operation

Mode setting is done by IBM Manufacturing. 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 (ESM) modules (B).

Figure 2-29 shows the mode stickers.

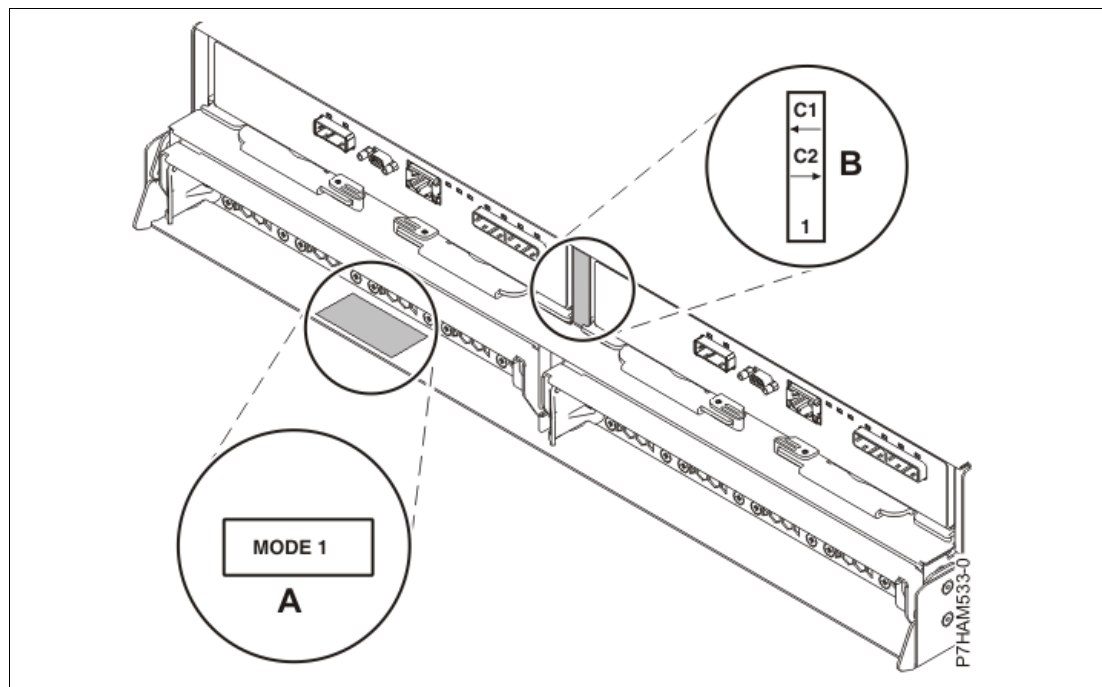


Figure 2-29 Mode sticker locations at the rear of the #5887 disk drive enclosure

The EXP24S SAS ports are attached to a SAS PCIe adapter or pair of adapters by using SAS YO or X cables. Cable length varies depending on the FC, and the correct length should be calculated while considering routing for proper airflow and ease of handling. Figure 2-30 shows both types of SAS cables.

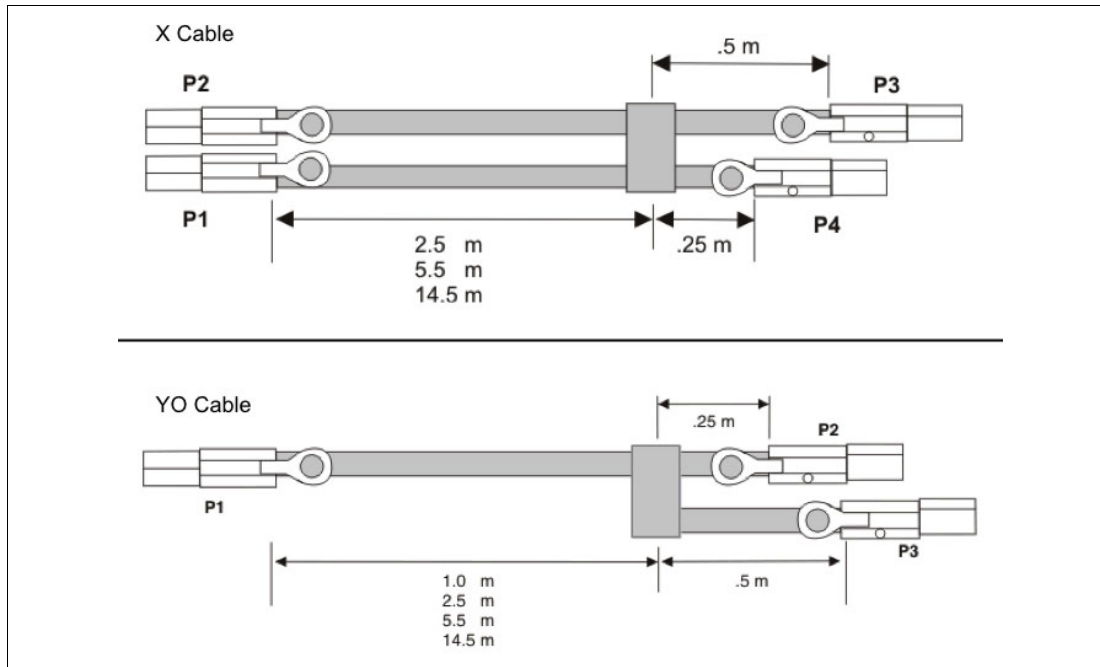


Figure 2-30 Diagram of SAS cables types X and YO

The following SAS adapters support the EXP24S:

- ▶ PCIe3 RAID SAS Adapter Quad-port 6 Gb x8 (#EJ0J)
- ▶ PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6 Gb x8 (#EJ0L) at a time after GA
- ▶ PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6 Gb x8 for MR9 (#EJ0K)
- ▶ PCIe3 12 GB Cache RAID Plus SAS Adapter Quad-port 6 Gb x8 (#EJ14)

The EXP24S drawer can support up to 24 SAS SFF Gen-2 disks.

Table 2-29 lists the available disk options.

Table 2-29 Available disks for the EXP24S

Feature code	Description	OS support
ES0G	775 GB SFF-2 SSD for AIX/Linux	AIX and Linux
ES0Q	387 GB SFF-2 4 K SSD for AIX/Linux	AIX and Linux
ES0S	775 GB SFF-2 4 K SSD for AIX/Linux	AIX and Linux
ES19	387 GB SFF-2 SSD for AIX/Linux	AIX and Linux
ES2C	387 GB SFF-2 SSD for AIX/Linux	AIX and Linux
ES62	3.86 - 4.0 TB 7200 RPM 4 K SAS LFF-1 Nearline Disk Drive (AIX/Linux)	AIX and Linux

Feature code	Description	OS support
ES64	7.72 - 8.0 TB 7200 RPM 4 K SAS LFF-1 Nearline Disk Drive (AIX/Linux)	AIX and Linux
ES78	387 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	AIX and Linux
ES7E	775 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux	AIX and Linux
ES80	1.9 TB Read Intensive SAS 4 K SFF-2 SSD for AIX/Linux	AIX and Linux
ES85	387 GB SFF-2 SSD 4 K eMLC4 for AIX/Linux	AIX and Linux
ES8C	775 GB SFF-2 SSD 4 K eMLC4 for AIX/Linux	AIX and Linux
ES8F	1.55 TB SFF-2 SSD 4 K eMLC4 for AIX/Linux	AIX and Linux
1752	900 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX and Linux
1953	300 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX and Linux
1964	600 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX and Linux
ESD3	1.2 TB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	AIX and Linux
ESDP	600 GB 15 K RPM SAS SFF-2 Disk Drive - 5xx Block (AIX/Linux)	AIX and Linux
ESEV	600 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX and Linux
ESEZ	300 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	AIX and Linux
ESF3	1.2 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX and Linux
ESFP	600 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	AIX and Linux
ESFT	1.8 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096	AIX and Linux

There are six SAS connectors at 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. The T1 connectors are not used. While configuring the drawer, special configuration FCs indicate for the mode of operation in which the disks and ports will be split:

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

Figure 2-31 shows the rear connectors of the EXP24S drawer, how they relate to the modes of operation, and disk grouping.

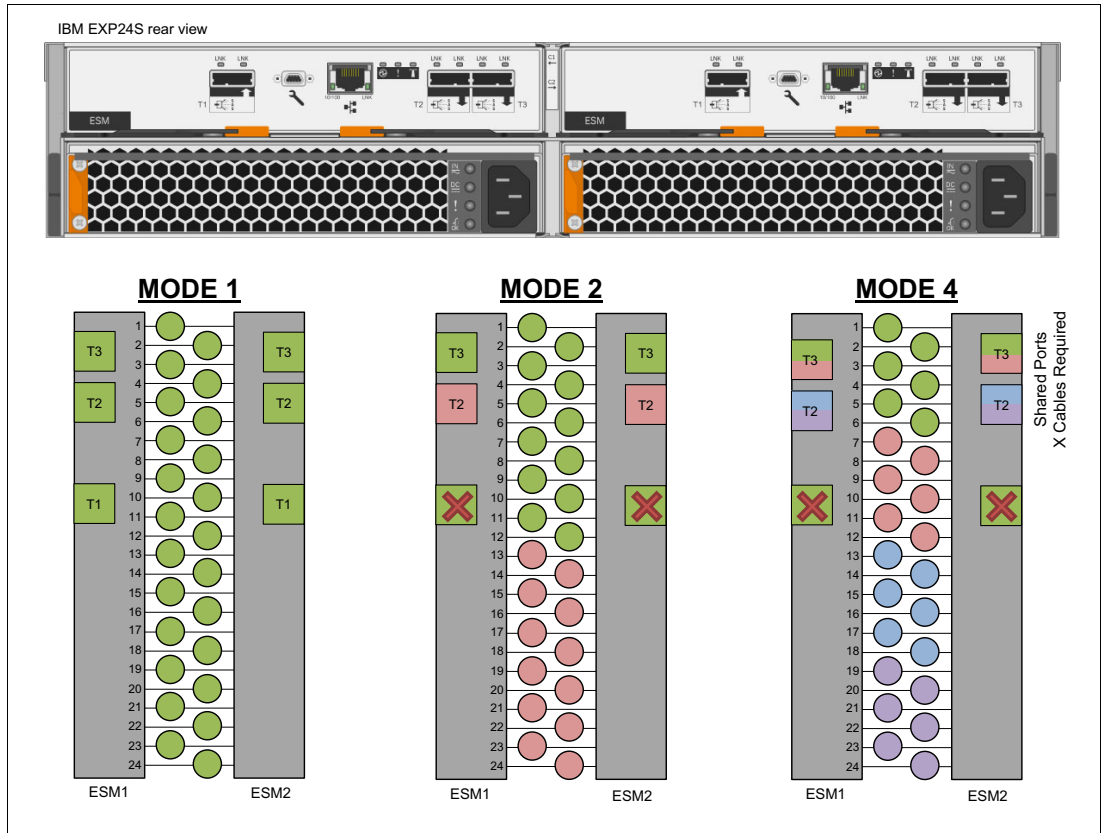


Figure 2-31 Rear view of EXP24S with the three modes of operation and the disks that are assigned to each port

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.

EXP24S common usage scenarios

The EXP24S drawer is versatile in the ways that it can be attached to Power Systems servers. This section describes the most common usage scenarios for EXP24S and VIOSes by using standard PCIe SAS adapters (#EJ0J).

Note: Not all possible scenarios are included. For more information about supported scenarios, see “Planning for serial-attached SCSI cables” in [IBM Knowledge Center](#).

Scenario 1: Basic non-redundant connection

This scenario assumes a single VIOS with a single PCIe SAS adapter (#EJ0J) and an EXP24S expansion drawer set on mode 1, allowing up to 24 disks to be attached to the server.

Figure 2-32 shows the connection diagram and components of the solution.

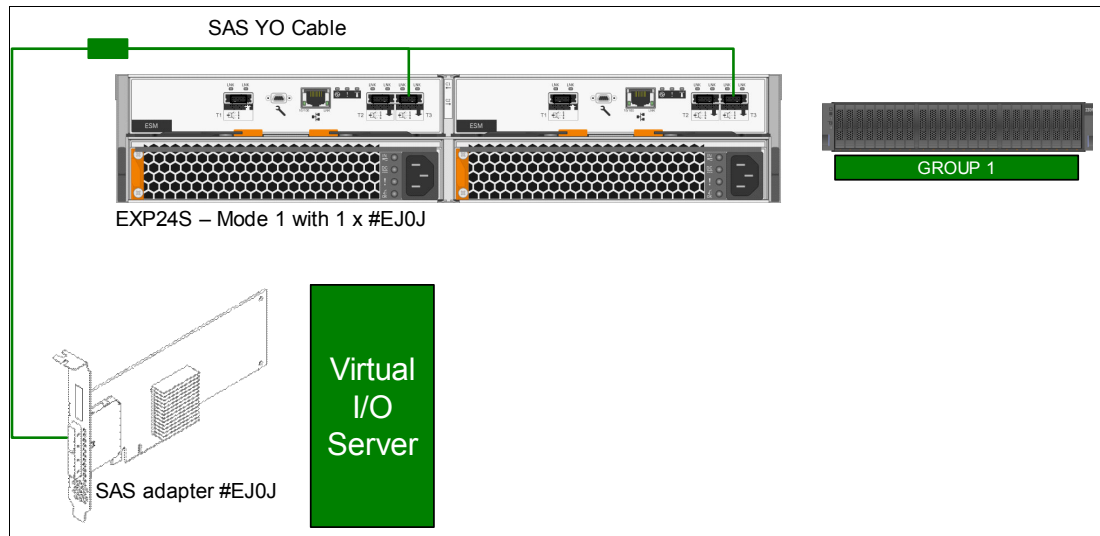


Figure 2-32 Scenario 1: Basic non-redundant connection

For this scenario, the following items are the required components:

- ▶ One EXP24S drawer (#5887) with mode 1
- ▶ One PCIe SAS adapter (#EJ0J)
- ▶ One SAS YO cable 3 Gbps with proper length (provided by manufacturing)

Scenario 2: Basic redundant connection

This scenario assumes a single VIOS with two PCIe SAS adapters (#EJ0J) and an EXP24S expansion drawer set on mode 1, allowing up to 24 disks to be attached to the server.

Figure 2-33 shows the connection diagram and components of the solution.

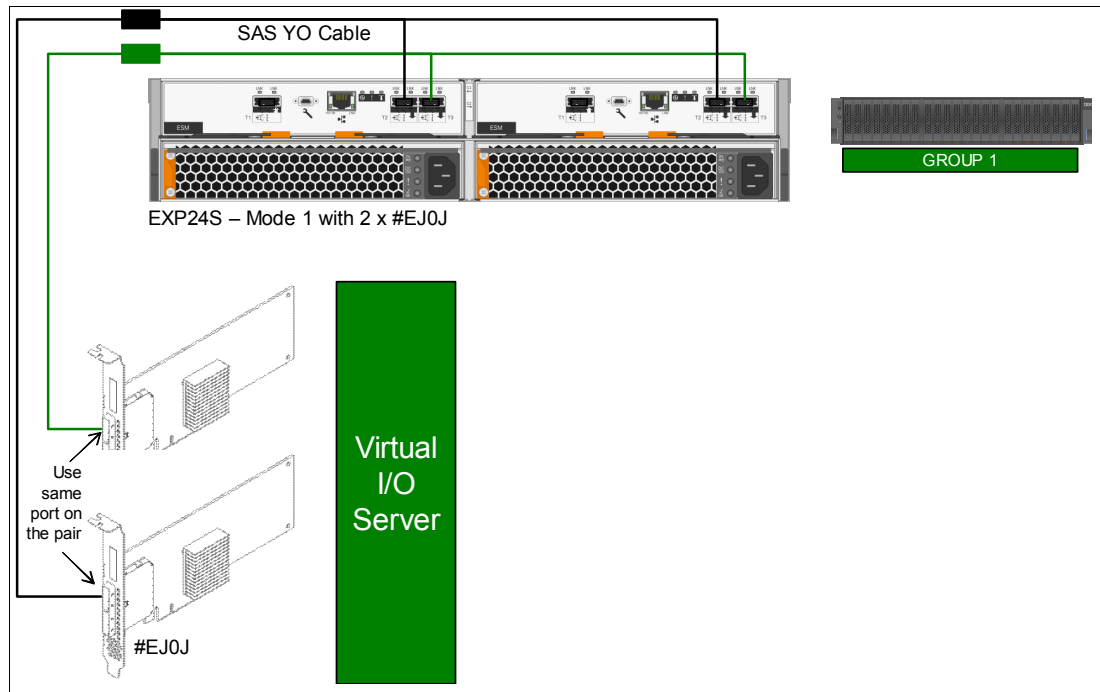


Figure 2-33 Scenario 2: Basic redundant connection

For this scenario, the following items are the required components:

- ▶ One EXP24S drawer (#5887) with mode 1
- ▶ Two PCIe SAS adapter (#EJ0J)
- ▶ Two SAS YO cables 3 Gbps with proper length (provided by manufacturing)

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

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

This scenario assumes dual VIOs with two PCIe SAS adapters (#EJ0J) each and an EXP24S expansion drawer set on mode 2, allowing up to 12 disks to be attached to each VIO.

Figure 2-34 shows the connection diagram and components of the solution.

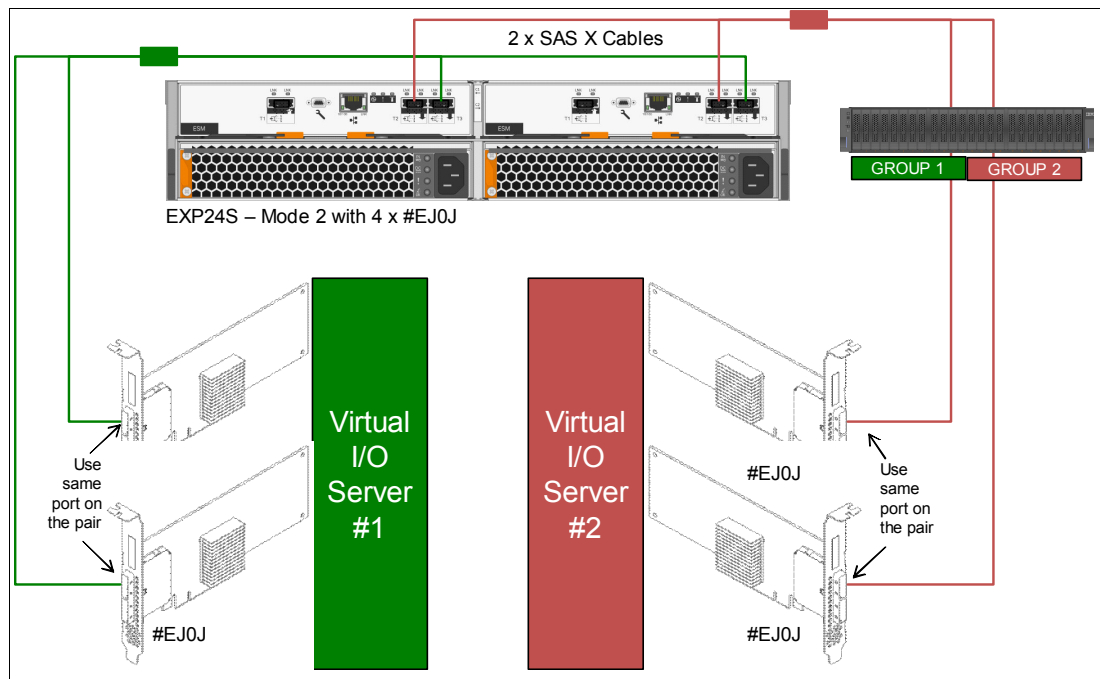


Figure 2-34 Dual Virtual I/O Servers sharing a single EXP24S

For this scenario, the following items are the required components:

- ▶ One EXP24S drawer (#5887) with mode 2
- ▶ Four PCIe SAS adapter (#EJ0J)
- ▶ Two SAS X cables 3 Gbps with proper length (provided by manufacturing)

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

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

This scenario assumes dual VIOs with two PCIe SAS adapters (#EJ0J) each and two EXP24S expansion drawers set on mode 2, allowing up to 24 disks to be attached to each VIOS (two per drawer). If compared to “Scenario 3: Dual Virtual I/O Servers sharing a single EXP24S drawer” on page 100, 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 correctly mirrored.

Figure 2-35 shows the connection diagram and components of the solution.

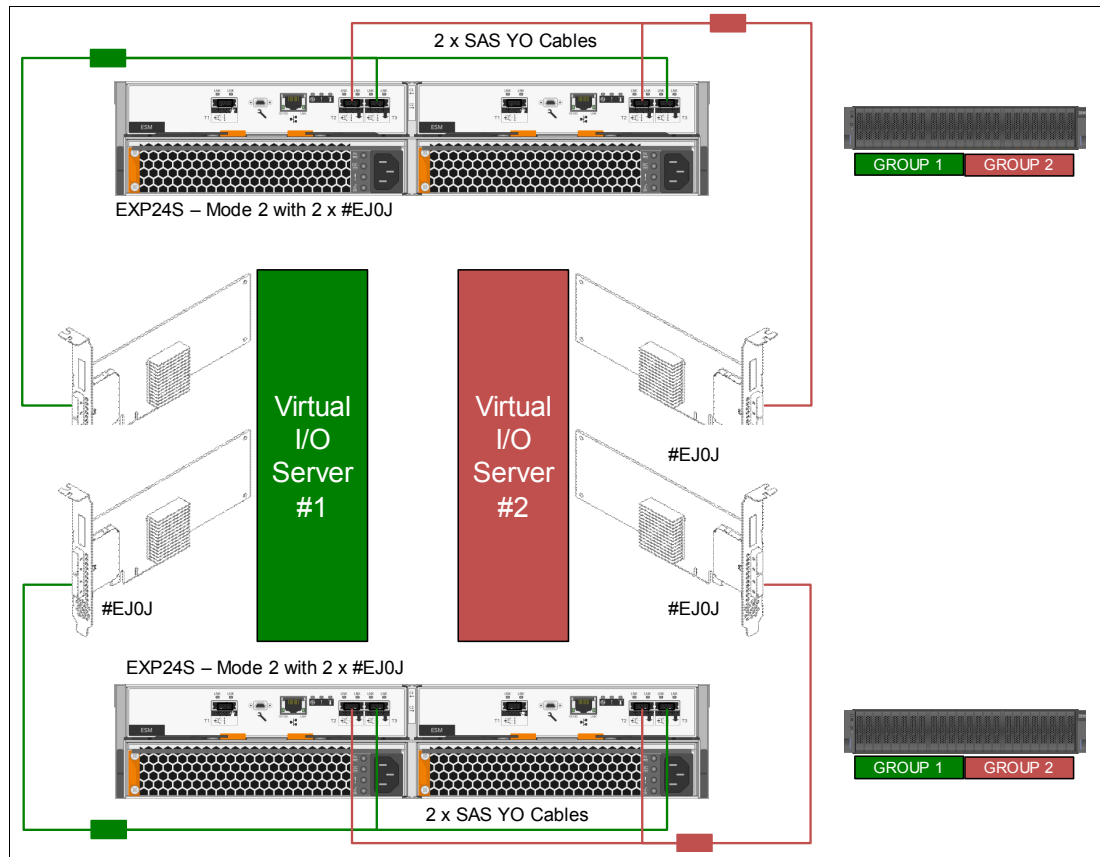


Figure 2-35 Dual Virtual I/O Servers sharing two EXP24S

For this scenario, the following items are the required components:

- ▶ Two EXP24S drawers (#5887) with mode 2
- ▶ Four PCIe SAS adapter (#EJ0J)
- ▶ Four SAS YO cables 3 Gbps with proper length (provided by manufacturing)

There is no SSD support in this scenario.

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

This scenario assumes four VIOs with two PCIe SAS adapters (#EJ0J) each and two EXP24S expansion drawers set on mode 4, allowing up to 12 disks to be attached to each VIO (six per drawer). This scenario has the benefit of allowing disks from different EXP24S drawers to be mirrored, allowing for hot maintenance of all of the EXP24S drawers if all data is properly mirrored.

Figure 2-36 shows the connection diagram and components of the solution.

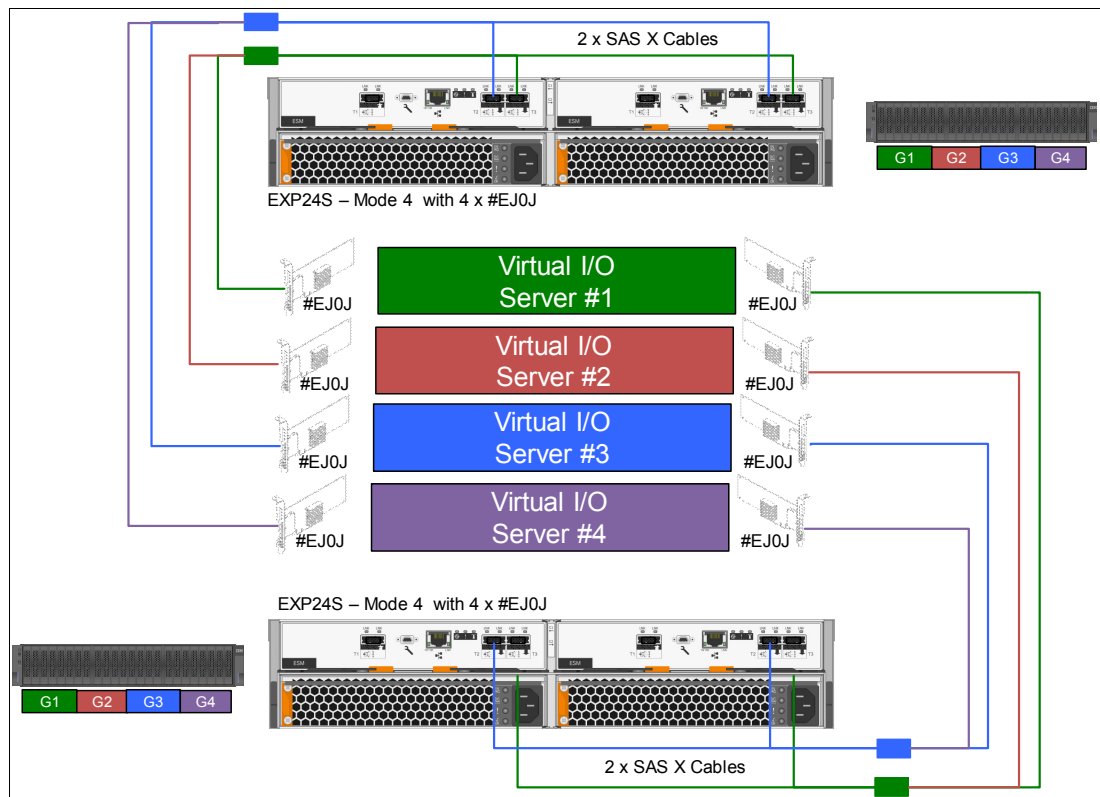


Figure 2-36 Four Virtual I/O Servers sharing two EXP24S drawers

For this scenario, here are the required FCs:

- ▶ Two EXP24S drawers (#5887) with mode 4
- ▶ Eight PCIe SAS adapter (#EJ0J)
- ▶ Four SAS X cables 3 Gbps with proper length (provided by manufacturing)

There is no SSD support in this scenario.

Other scenarios

For more information about direct connection to logical partitions (LPARs), different adapters, and cables, see “5887 disk drive enclosure” in [IBM Knowledge Center](#).

2.8.2 EXP24SX and EXP12SX SAS Storage Enclosure

The EXP24SX is a storage expansion enclosure with 24 2.5-inch SFF SAS bays. It supports HDDs or SSDs. The EXP12SX is a storage expansion enclosure with twelve 3.5-inch large form factor (LFF) SAS bays. The EXP12SX supports HDD only.

The following PCIe3 SAS adapters support the EXP24SX and EXP12SX enclosures:

- ▶ PCIe3 RAID SAS Adapter Quad-port 6 Gb x8 (#EJ0J)
- ▶ PCIe3 RAID SAS Adapter Quad-port 6 Gb x8 for MR9 (#EJ0K)
- ▶ PCIe3 12 GB Cache RAID Plus SAS Adapter Quad-port 6 Gb x8 (#EJ14)

Protecting the drives is highly advised, but not required for other operating systems. All POWER operating system environments that use SAS adapters with write cache require the cache to be protected by using pairs of adapters.

The EXP24SX and EXP12SX enclosures have many high-reliability design points:

- ▶ SAS bays that support hot-swap.
- ▶ Redundant and hot-plug power and fan assemblies.
- ▶ Dual power cords.
- ▶ Redundant and hot-plug ESMs.
- ▶ Redundant data paths to all drives.
- ▶ LED indicators on drives, bays, ESMs, and power supplies that support problem identification.
- ▶ Through the SAS adapters/controllers, drives that can be protected with RAID and mirroring and hot-spare capability.

Notes:

- ▶ For the EXP24SX, A maximum of twenty-four 2.5-inch SSDs or 2.5-inch HDDs are supported in the #ESLS 24 SAS bays. You cannot mix HDDs and SSDs in the same mode-1 drawer. You can mix HDDs and SSDs in a mode-2 or mode-4 drawer, but you cannot mix them within a logical split of the drawer. For example, in a mode-2 drawer with two sets of 12 bays, one set can hold SSDs and one set can hold HDDs, but you cannot mix SSDs and HDDs in the same set of 12 bays.
- ▶ You can mix the EXP24S, EXP24SX, and EXP12SX drawers on the same server and on the same PCIe3 adapters.
- ▶ The EXP12SX does not support SSD.

The cables that are used to connect an #ESLL or #ESLS storage enclosure to a server are different from the cables that are used with the #5887 disk drive enclosure. Attachment between the SAS controller and the storage enclosure SAS ports is through the appropriate SAS YO12 or X12 cables. The PCIe Gen3 SAS adapters support 6 Gb throughput. The EXP12SX and EXP24SX drawers support up to 12 Gb throughput if future SAS adapters support that capability.

The cable options are:

- ▶ 3.0M SAS X12 Cable (Two Adapter to Enclosure) (#ECDJ)
- ▶ 4.5M SAS X12 AOC (Two Adapter to Enclosure) (#ECDK)
- ▶ 10M SAS X12 AOC (Two Adapter to Enclosure) (#ECDL)
- ▶ 1.5M SAS YO12 Cable (Adapter to Enclosure) (#ECDT)
- ▶ 3.0M SAS YO12 Cable (Adapter to Enclosure) (#ECDU)
- ▶ 4.5M SAS YO12 AOC (Adapter to Enclosure) (#ECDV)
- ▶ 10M SAS YO12 AOC (Adapter to Enclosure) (#ECDW)

There are six SAS connectors at the rear of the EXP24SX and EXP12SX enclosures to which SAS adapters or controllers are attached. They are labeled T1, T2, and T3; there are two T1, two T2, and two T3 connectors. The T1 connectors are not used.

- ▶ 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 a dual adapters configuration.
- ▶ In mode 2 or mode 4, four ports are used, two T2 and two T3 connectors, to access all SAS bays.

Figure 2-37 shows connector locations for the EXP24SX and EXP12SX storage enclosures.

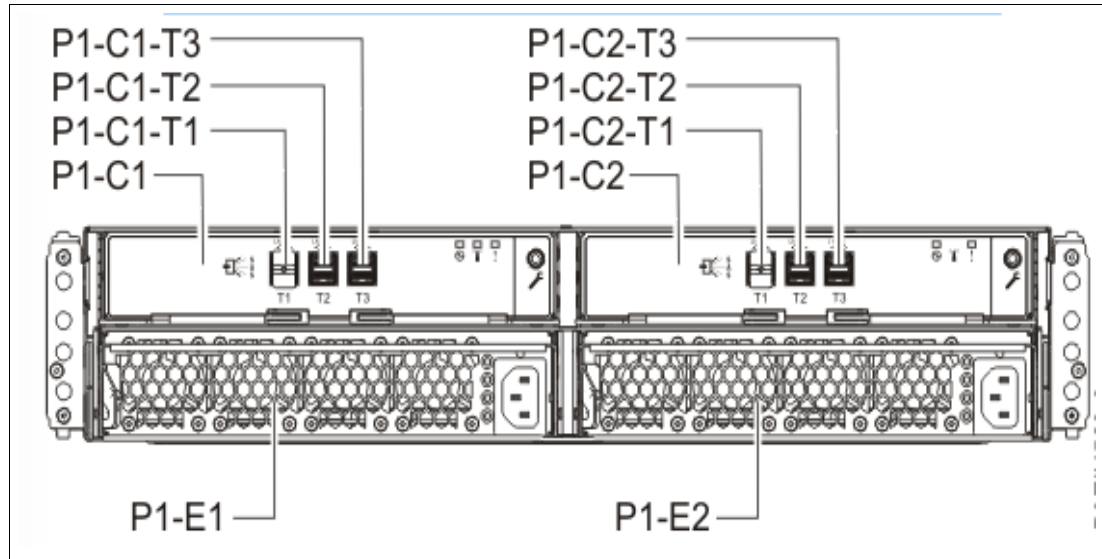


Figure 2-37 Connector locations for the EXP24SX and EXP12SX storage enclosures

Mode setting is done by IBM Manufacturing. If you must change the mode after installation, ask your IBM System Services Representative (SSR) for help, and direct them to [Mode Change on Power EXP24SX and EXP12SX SAS Storage Enclosures \(Features #ESLL, #ESLS, #ELLL, #ELLS\)](#).

For more information about SAS cabling and cabling configurations, see “Connecting an #ESLL or #ESLS storage enclosure to your system” in [IBM Knowledge Center](#).

2.8.3 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 [Hybrid Storage Solutions](#).

The following section highlights a few of the offerings.

IBM Storwize Family

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

IBM FlashSystem Family

The IBM FlashSystem® family delivers extreme performance to derive measurable economic value across the data architecture (servers, software, applications, and storage). IBM offers a comprehensive flash portfolio with the IBM FlashSystem family. For more information, see [IBM FlashSystem](#).

IBM XIV Storage System

The IBM XIV® Storage System hardware is part of the IBM Spectrum Accelerate™ family and 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 Storage Systems extend ease of use with integrated management for large and multi-site XIV deployments, reducing operational complexity and enhancing capacity planning. For more information, see [IBM XIV Storage System](#).

IBM System Storage DS8000

The IBM System Storage DS8000® storage subsystem 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 [IBM Knowledge Center](#).

2.9 Operating system support

The Power E950 server supports the following operating systems:

- ▶ AIX
- ▶ Linux

In addition, the VIOS can be installed in special partitions that provide support to other partitions running AIX or Linux operating systems for using features such as virtualized I/O devices, PowerVM LPM, or PowerVM Active Memory Sharing.

For more information about the software that is available on Power Systems, see [IBM Power Systems Software](#).

2.9.1 AIX operating system

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

IBM periodically releases maintenance packages (service packs or technology levels) for the AIX operating system. Information about these packages, downloading, and obtaining the CD-ROM can be found at [Fix Central](#).

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

The Service Update Management Assistant (SUMA), which can help you automate the task of checking and downloading operating system downloads, is part of the base operating system. For more information about the `suma` command, see [IBM Knowledge Center](#).

Table 2-30 shows the minimum supported AIX levels when using any I/O configuration.

Table 2-30 Supported minimum AIX levels for any I/O

Version	Technology level	Service pack	Planned availability
7.2	3		14 September 2018
7.2	2	2	7 August 2018
7.2	1	5	January 2019
7.1	5	2	7 August 2018
7.1	4		January 2019
6.1 ^a	9	12	7 August 2018

a. AIX 6.1 service extension is required.

Table 2-31 shows the minimum supported AIX levels when using virtual I/O only.

Table 2-31 Supported minimum AIX levels for virtual I/O only

Version	Technology level	Service pack	Planned availability
7.2	3		14 September 2018
7.2	2	1	7 August 2018
7.2	1	1	7 August 2018
7.1	5	1	7 August 2018
7.1	4	2	7 August 2018
6.1 ^a	9	7	7 August 2018

a. AIX 6.1 service extension is required.

For compatibility information for hardware features and the corresponding AIX Technology Levels, see [IBM Prerequisites](#).

2.9.2 Linux operating system

Linux is an open source, cross-platform operating system that runs on numerous platforms, such as embedded systems and mainframe computers. It provides an UNIX like implementation across many computer architectures.

Here are the supported versions of Linux on the Power E950 server:

- ▶ Red Hat Enterprise Linux 7.5 for Power LE (p8compat) or later
- ▶ Red Hat Enterprise Linux for SAP with Red Hat Enterprise Linux 7 for Power LE version 7.5 or later
- ▶ SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 11 Service Pack 4
- ▶ SUSE Linux Enterprise Server 12 Service Pack 3 or later
- ▶ SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 12 Service Pack or later

Service and productivity tools

Service and productivity tools are available in a YUM repository that you can use to download, and then install all recommended packages for your Red Hat, SUSE Linux, or Fedora distribution. You can find the repository at [Service and productivity tools](#).

Learn about developing on the IBM Power Architecture®, find packages, get access to cloud resources, and discover tools and technologies by going to the [Linux on IBM Power Systems Developer Portal](#).

The IBM Advance Toolchain for Linux on Power is a set of open source compilers, runtime libraries, and development tools that you can use to take leading-edge advantage of POWER hardware features on Linux. For more information, see [Advance toolchain for Linux on Power](#).

For more information about SUSE Linux Enterprise Server, see [SUSE Linux Enterprise Server](#).

For more information about Red Hat Enterprise Linux, see [Red Hat Enterprise Linux](#).

2.9.3 Virtual I/O Server

The minimum required level of VIOS for the Power E950 server is VIOS 2.2.6.23 or later.

IBM regularly updates the VIOS code. For more information about the latest updates, see [Fix Central](#).



Virtualization

Virtualization is a key factor for productive and efficient use of IBM Power Systems servers. In this chapter, you find a brief description of virtualization technologies that are available for POWER9 processor-based systems. The following IBM Redbooks publications provide more information about the virtualization features:

- ▶ *IBM PowerVM Best Practices*, SG24-8062
- ▶ *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- ▶ *IBM PowerVM Virtualization Active Memory Sharing*, REDP-4470
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- ▶ *IBM Power Systems SR-IOV: Technical Overview and Introduction*, REDP-5065

3.1 IBM POWER Hypervisor

Power Systems servers that are 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 the total cost of ownership (TCO) and better use 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 that you can make business-driven policies to deliver resources that are based on time, cost, and service-level requirements.

Combined with features in the POWER9 processors, the IBM POWER Hypervisor™ delivers functions that enable other system technologies, including logical partitioning (LPAR) technology, virtualized processors, IEEE VLAN-compatible virtual switch, virtual SCSI adapters, virtual Fibre Channel adapters, and virtual consoles. The POWER Hypervisor is a basic component of the system's firmware and offers the following functions:

- ▶ Provides an abstraction between the physical hardware resources and the LPARs that use them.
- ▶ Enforces partition integrity by providing a security layer between LPARs.
- ▶ Controls the dispatch of virtual processors to physical processors.
- ▶ Saves and restores all processor state information during a logical processor context switch.
- ▶ Controls hardware I/O interrupt management facilities for LPARs.
- ▶ Provides virtual local area network (LAN) channels between LPARs 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 or whether it is connected to the managed console. It requires memory to support the resource assignment of the LPARs on the server. The amount of memory that is required by the POWER Hypervisor firmware varies according to several factors:

- ▶ Memory usage for hardware page tables (HPTs)
- ▶ Memory usage to support I/O devices
- ▶ Memory usage for virtualization

Memory usage for hardware page tables

Each partition on the system has its own HPT that contributes to hypervisor memory usage. The HPT is used by the operating system to translate from effective addresses to physical real addresses in the hardware. This translation from effective to real addresses allows multiple operating systems to run simultaneously in their own logical address space. Whenever a virtual processor for a partition is dispatched on a physical processor, the hypervisor indicates to the hardware the location of the partition HPT that should be used when translating addresses.

The amount of memory for the HPT is based on the maximum memory size of the partition and the HPT ratio. The default HPT ratio is 1/128th (for AIX, Virtual I/O Server (VIOS), and Linux partitions) of the maximum memory size of the partition. AIX, VIOS, and Linux use larger page sizes (16 and 64 KB) instead of using 4 KB pages. Using larger page sizes reduces the overall number of pages that must be tracked, so the overall size of the HPT can be reduced. As an example, for an AIX partition with a maximum memory size of 256 GB, the HPT would be 2 GB.

When defining a partition, the maximum memory size that is specified should be based on the amount of memory that can be dynamically added to the dynamic partition (DLPAR) without having to change the configuration and restart the partition.

In addition to setting the maximum memory size, the HPT ratio can also be configured. The **hpt_ratio** parameter for the **chsyscfg** Hardware Management Console (HMC) command can be issued to define the HPT ratio that is used for a partition profile. The valid values are 1:32, 1:64, 1:128, 1:256, or 1:512. Specifying a smaller absolute ratio (1/512 is the smallest value) decreases the overall memory that is assigned to the HPT. Testing is required when changing the HPT ratio because a smaller HPT might incur more CPU consumption because the operating system might need to reload the entries in the HPT more frequently. Most customers choose to use the IBM provided default values for the HPT ratios.

Memory usage for I/O devices

In support of I/O operations, the hypervisor maintains structures that are called the translation control entities (TCEs), which provide an information path between I/O devices and partitions. The TCEs provide the address of the I/O buffer, indications of read versus write requests, and other I/O-related attributes. There are many TCEs in use per I/O device, so multiple requests can be active simultaneously to the same physical device. To provide better affinity, the TCEs are spread across multiple processor chips or drawers to improve performance while accessing the TCEs.

For physical I/O devices, the base amount of space for the TCEs is defined by the hypervisor based on the number of I/O devices that are supported. A system that supports high-speed adapters can also be configured to allocate more memory to improve I/O performance. Linux is the only operating system that uses these additional TCEs so that the memory can be freed for use by partitions if the system is using only AIX.

Memory usage for virtualization features

Virtualization requires more memory to be allocated by the POWER Hypervisor for hardware statesave areas and various virtualization technologies. For example, on POWER9 processor-based systems, each processor core supports up to eight simultaneous multithreading (SMT) threads of execution, and each thread contains over 80 different registers.

The POWER Hypervisor must set aside save areas for the register contents for the maximum number of virtual processors that is configured. The greater the number of physical hardware devices, the greater the number of virtual devices, the greater the amount of virtualization, and the more hypervisor memory is required. For efficient memory consumption, wanted and maximum values for various attributes (processors, memory, and virtual adapters) should be based on business needs, and not set to values that are significantly higher than actual requirements.

Predicting memory that is used by the POWER Hypervisor

The IBM System Planning Tool (SPT) is a resource that can be used to estimate the amount of hypervisor memory that is required for a specific server configuration. After the SPT executable file is downloaded and installed, you can define a configuration by selecting the correct hardware platform, selecting the installed processors and memory, and defining partitions and partition attributes. SPT can estimate the amount of memory that will be assigned to the hypervisor, which assists you when you change an existing configuration or deploy new servers.

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

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

Virtual SCSI

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

Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions either fast and secure communication on the same server without any need for physical interconnection or 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 (SEA).

Virtual Fibre Channel

A virtual Fibre Channel adapter is a virtual adapter that provides client LPARs with a Fibre Channel connection to a storage area network through the VIOS partition. The VIOS partition provides the connection between the virtual Fibre Channel adapters on the VIOS partition and the physical Fibre Channel adapters on the managed system.

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.

3.1.1 POWER processor modes

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

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

- ▶ POWER7 compatibility mode

This is the mode for POWER7+ and POWER7 processors, implementing Version 2.06 of the IBM Power Instruction Set Architecture (ISA). For more information, see [IBM Knowledge Center](#).

- ▶ POWER8 compatibility mode

This is the native mode for POWER8 processors implementing Version 2.07 of the IBM Power ISA. For more information, see [IBM Knowledge Center](#).

- ▶ POWER9 compatibility mode

This is the native mode for POWER9 processors implementing Version 3.0 of the IBM Power ISA. For more information, see [IBM Knowledge Center](#).

Figure 3-1 shows the available processor modes on a POWER9 processor-based system.

General	Processors	Memory	I/O	Virtual Adapters	Power Controlling	Settings
Detailed below are the current processing settings for this partition profile.						
Processing mode						
<input type="radio"/> Dedicated <input checked="" type="radio"/> Shared						
Processing units						
Total managed system processing units : 16.00						
Minimum shared processing units : <input type="text" value="0.1"/>						
Desired shared processing units : <input type="text" value="4.0"/>						
Maximum shared processing units : <input type="text" value="8.0"/>						
Shared processor pool: <input type="text" value="DefaultPool (0)"/>						
Virtual processors						
Minimum processing units required for each virtual processor : 0.10						
Newer operating system levels support : 0.05						
Minimum virtual processors : <input type="text" value="1.0"/>						
Desired virtual processors : <input type="text" value="8.0"/>						
Maximum virtual processors : <input type="text" value="16.0"/>						
Sharing mode						
<input checked="" type="checkbox"/> Uncapped Weight : <input type="text" value="128"/>						
Processor compatibility mode: <input type="text" value="default"/>						
<input type="button" value="OK"/>	<input type="button" value="Cancel"/>	<input type="button" value="Help"/>	<input type="text" value="default"/> POWER7 POWER8 POWER9_base			

Figure 3-1 POWER9 processor modes

Processor compatibility mode is important when Live Partition Mobility (LPM) migration is planned between different generation of servers. An LPAR that potentially might be migrated to a machine that is managed by a processor from another generation must be activated in a specific compatibility mode.

Table 3-1 shows an example where the processor mode must be selected when a migration from POWER9 to POWER8 is planned.

Table 3-1 Processor compatibility modes for a POWER9 to POWER8 migration

Source environment POWER9 server		Destination environment POWER8 server			
		Active migration		Inactive migration	
Wanted processor compatibility mode	Current processor compatibility mode	Wanted processor compatibility mode	Current processor compatibility mode	Wanted processor compatibility mode	Current processor compatibility mode
POWER9	POWER9	Fails because the wanted processor mode is not supported on the destination.		Fails because the wanted processor mode is not supported on the destination.	
POWER9	POWER8	Fails because the wanted processor mode is not supported on the destination.		Fails because the wanted processor mode is not supported on the destination.	
Default	POWER9	Fails because the wanted processor mode is not supported on destination.		Default	POWER8
POWER8	POWER8	POWER8	POWER8	POWER8	POWER8
Default	POWER8	Default	POWER8	Default	POWER8
POWER7	POWER7	POWER7	POWER7	POWER7	POWER7

3.2 Active Memory Expansion

Active Memory Expansion (AME) is an optional feature code (FC) for the Power E950 server.

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

AME 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 1000% 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 AME feature is not supported by IBM i and the Linux operating systems.

3.3 Single Root I/O Virtualization

Single Root I/O Virtualization (SR-IOV) is an extension to the Peripheral Component Interconnect 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 a VIOS.

POWER9 provides the following SR-IOV enhancements:

- ▶ Faster ports: 10 Gb, 25 Gb, 40 Gb, and 100 Gb
- ▶ More virtual functions (VFs) per port: Sixty VFs per port (120 VFs per adapter) for 100-Gb adapters
- ▶ vNIC and vNIC failover support for Linux

Here are the hardware requirements to enable SR-IOV:

- ▶ PCIe2 4-port (10-Gb Fibre Channel over Ethernet (FCoE) and 1 GbE) SR & RJ45 Adapter (#EN0H)
- ▶ PCIe2 4-port (10-Gb FCoE and 1 GbE) SFP+Copper and RJ4 Adapter (#EN0K)
- ▶ PCIe3 4-port 10 GbE SR Adapter (#EN15)

For more information, see *IBM Power Systems SR-IOV: Technical Overview and Introduction*, REDP-5065.

3.4 PowerVM

The PowerVM platform is the family of technologies, capabilities, and offerings that delivers industry-leading virtualization on Power Systems servers. It is the umbrella branding term for Power Systems virtualization (logical partitioning, IBM Micro-Partitioning®, POWER Hypervisor, VIOS, LPM, and more). As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and software.

Note: PowerVM Enterprise Edition License Entitlement is now included with each Power E950 server. PowerVM Enterprise Edition is available as a hardware feature (#EPVV) and supports up to 20 partitions per core, VIOS, and multiple shared processor pools (MSPPs). It also offers LPM, Active Memory Sharing, and IBM PowerVP™ performance monitoring.

Logical partitions

LPARs and virtualization increase the usage of system resources and add a level of configuration possibilities.

Logical partitioning is the ability to make a server that is run as though it were two or more independent servers. When you logically partition a server, you divide the resources on the server into subsets called LPARs. You can install software on an LPAR, and the LPAR runs as an independent logical server with the resources that you allocated to the LPAR. LPAR is the equivalent of a virtual machine (VM).

You can assign processors, memory, and input/output devices to LPARs. You can run AIX and Linux, and VIOS in LPARs. VIOS provides virtual I/O resources to other LPARs with general-purpose operating systems.

LPARs share a few system attributes, such as the system serial number, system model, and processor FCs. All other system attributes can vary from one LPAR to another.

Micro-Partitioning

When you use the Micro-Partitioning technology, you can allocate fractions of processors to an LPAR. An LPAR that uses fractions of processors is also known as a *shared processor partition* or *micropartition*. Micropartitions run over a set of processors that is called a *shared processor pool* (SPP), and virtual processors are used to let the operating system manage the fractions of processing power that are assigned to the LPAR. 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.

On the POWER9 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 0.05 of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under the control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC.

The Power E950 server supports up to 48 cores in a single system. Here are the maximum numbers:

- ▶ 48 dedicated partitions
- ▶ 960 micropartitions (maximum of 20 micropartitions per physical active core)

The maximum amounts are supported by the hardware, but the practical limits depend on application workload demands.

Processing mode

When you create an LPAR, you can assign entire processors for dedicated use, or you can assign partial processing units from an SPP. This setting defines the processing mode of the LPAR.

Dedicated mode

In dedicated mode, physical processors are assigned as a whole to partitions. The SMT feature in the POWER9 processor core allows the core to run instructions from two, four, or eight independent software threads simultaneously.

Shared dedicated mode

On POWER9 processor-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 an SPP. 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, LPARs 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). 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. All partitions receive a 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 SMT is turned on or off.

3.4.1 Multiple shared processor pools

MSPPs are supported on POWER9 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 used from the physical SPP.

Micropartitions are created and then identified as members of either the default processor pool or a user-defined SPP. 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 an SPP is assured of 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 an SPP do not use their capacity entitlement, the unused capacity is ceded and other uncapped micropartitions within the same SPP are allocated the additional capacity according to their uncapped weighting. In this way, the entitled pool capacity of an SPP is distributed to the set of micropartitions within that SPP.

All Power Systems servers that support the MSPPs capability have a minimum of one (the default) SPP and up to a maximum of 64 SPPs.

3.4.2 Virtual I/O Server

The VIOS is part of PowerVM. It is specific appliance that allows the sharing of physical resources between LPARs to allow more efficient usage (for example, consolidation). In this case, the VIOS owns the physical resources (SCSI, Fibre Channel, network adapters, or 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 3-2 shows an overview of a VIOS configuration.

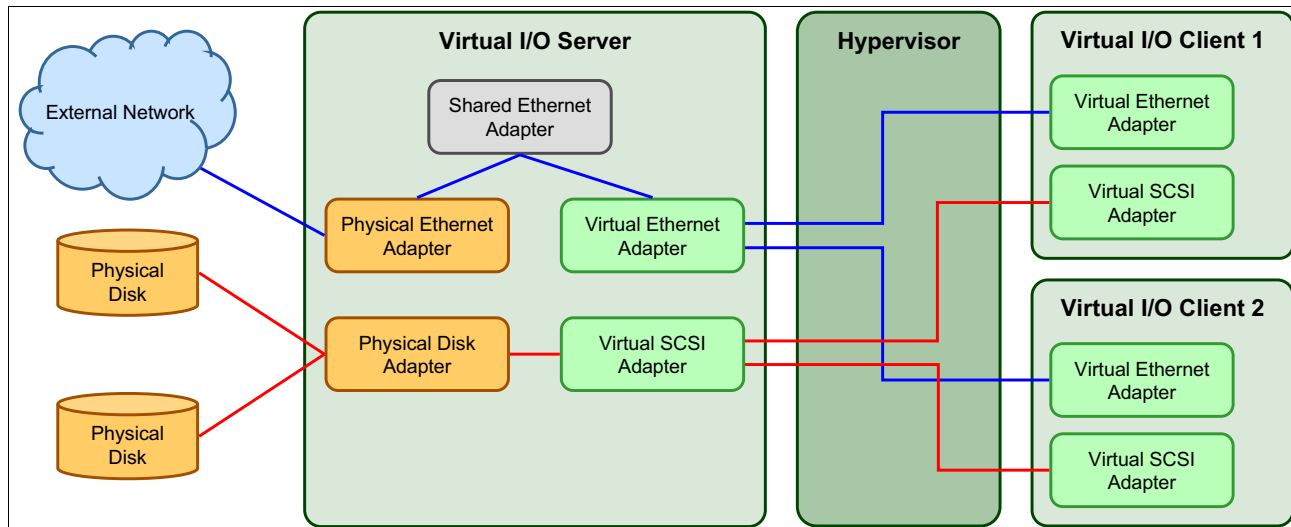


Figure 3-2 Architectural view of the VIOS

It is a preferred practice to run two VIOSes per physical server.

Shared Ethernet Adapter

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

By using the SEA, several client partitions can share one physical adapter, and you can connect internal and external VLANs by using a physical adapter. The SEA 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.

Virtual SCSI

Virtual SCSI is used to view a virtualized implementation of the SCSI protocol. Virtual SCSI is based on a client/server relationship. The VIOS LPAR owns the physical resources and acts as a server or, in SCSI terms, a target device. The client LPARs access the virtual SCSI backing storage devices that are provided by the VIOS as clients.

The virtual I/O adapters (a virtual SCSI server adapter and a virtual SCSI client adapter) are configured by using a managed console or through the Integrated Virtualization Manager (IVM) on smaller systems. The virtual SCSI server (target) adapter is responsible for running any SCSI commands that it receives. It is owned by the VIOS partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN-attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the VIOS.

N_Port ID Virtualization

N_Port ID Virtualization (NPIV) is a technology that allows multiple LPARs 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 has one or more virtual Fibre Channel adapters, each with their own 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 *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

3.4.3 Live Partition Mobility

LPM allows you to move a running LPAR from one system to another without disruption. Inactive partition mobility allows you to move a powered-off LPAR from one system to another one.

LPM provides systems management flexibility and improves system availability by:

- ▶ Avoiding planned outages for hardware upgrade or firmware maintenance.
- ▶ Avoiding unplanned downtime. With preventive failure management, if a server indicates a potential failure, you can move its LPARs to another server before the failure occurs.

For more information and requirements for LPM, see *IBM PowerVM Live Partition Mobility*, SG24-7460.

3.4.4 Active Memory Sharing

Active Memory Sharing provides system memory virtualization capabilities, allowing multiple partitions to share a common pool of physical memory.

The physical memory of a Power Systems server can be assigned to multiple partitions in either dedicated or shared mode. A 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.

3.4.5 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 kinds of applications. On heavily loaded systems, this behavior might lead to a shortage of the available memory resources, forcing paging by the Active Memory Sharing partition operating systems, the Active Memory Deduplication pool, or both, which might decrease overall system performance.

Active Memory Deduplication allows the POWER Hypervisor to map dynamically identical partition memory pages to a single physical memory page within a shared memory pool. This way enables a better usage of the Active Memory Sharing shared memory pool, increasing the system's overall performance by avoiding paging. Deduplication can cause the hardware to incur fewer cache misses, which also leads to improved performance.

Active Memory Deduplication depends on the Active Memory Sharing feature being available, and it uses 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 suggest to the POWER Hypervisor that some pages (such as frequently referenced read-only code pages) are good for deduplication.

3.4.6 Remote Restart

Remote Restart is a high availability option for partitions. If there is an error that causes a server outage, a partition that is configured for Remote Restart can be restarted on a different physical server. At times, it might take longer to start the server, in which case the Remote Restart function can be used for faster reprovisioning of the partition. Typically, this can be done faster than restarting the server that stopped and then restarting the partitions. The Remote Restart function relies on technology similar to LPM where a partition is configured with storage on a SAN that is shared (accessible) by the server that will host the partition.

HMC V9R2 provides the following enhancements to the Remote Restart feature.

- ▶ Remote restart a partition with reduced or minimum CPU/memory on the target system.
- ▶ Remote restart by choosing a different virtual switch on the target system.
- ▶ Remote restart the partition without turning on the partition on the target system.
- ▶ Remote restart the partition for test purposes when the source-managed system is in the Operating or Standby state.
- ▶ Remote restart through the REST API.



Reliability, availability, serviceability, and manageability

This chapter provides information about reliability, availability, and serviceability (RAS) design and features of the IBM Power E950 system.

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 Power E950 specific RAS enhancements

The Power E950 system defines the most current implementation of a POWER processor system based on a symmetric multiprocessor (SMP) architecture with four processor modules in a single central electronic complex.

The Power E950 server inherits many of the basic RAS design elements of the Power E850 and Power E850C servers. The Power E950 RAS characteristics are substantially enhanced in comparison to its POWER8 processor-based predecessor systems.

This section provides reference lists of strategic RAS features that are implemented in Power E950 servers. More detailed explanations for individual features are given in subsequent sections of this chapter and in [Power Processor-Based Systems RAS](#).

The following list highlights the most important RAS improvements for the Power E950 system:

- ▶ Peripheral Component Interconnect Express (PCIe) adapter blind-swap cassette (BSC) implementation
- ▶ Easy access to I/O adapters in BSCs from the rear of the system
- ▶ Concurrently maintainable Serial Attached SCSI (SAS) storage controller
- ▶ Strategic voltage regulator modules (VRMs) implemented as independently serviceable field replaceable units (FRUs) with predictive callout support
- ▶ Single cooling domain with reduced number of fans
- ▶ Concurrently maintainable fans
- ▶ Operator panel that is composed of concurrently maintainable independent base and LCD units
- ▶ Time-of-day battery concurrently maintainable and accessible from rear of the system

The following RAS features are specific to the enterprise class servers and are shared by the Power E950 and Power E980 systems¹:

- ▶ Core-contained checkstops
- ▶ Extended L2/L3/L4 cache line delete
- ▶ IBM memory buffer support and spare DRAM module capacity with 4 bit wide (x4) dual inline memory modules (DIMMs)
- ▶ Memory row repair
- ▶ Active Memory Mirroring (AMM) for Hypervisor
- ▶ Internal Non-Volatile Memory Express (NVMe) drive boot support
- ▶ Voltage regulators with N+1 phase redundancy
- ▶ Redundant/spare voltage phases on voltage converters for levels feeding processors, Power E980 memory CDIMMs, or Power E950 memory riser cards
- ▶ PCIe3 optical cable adapter with new routing for clock logic within the card and extra recovery procedures for faults during initial program load (IPL)
- ▶ New concurrently maintainable operator panel design that is composed of a separate base and LED unit and a USB connection
- ▶ Time-of-day battery concurrent maintenance

¹ Some RAS features that re listed in the previous list are included again for completeness.

The following list shows the POWER9 processor base RAS features that are shared among all POWER9 processor-based systems. It also shows important infrastructure-related RAS features that pertain to the Power E950 system but are shared with the POWER9 scale-out servers Power System S914, Power System S922, Power System S924, Power System H922, and Power System H924.

- ▶ Traditional POWER9 processor RAS features include First-Failure Data Capture (FFDC), processor instruction retry, L2/L3 cache error correction code (ECC) protection with cache line-delete, and a power/cooling monitor function integrated into an on chip controller (OCC)
- ▶ OCC error handling with power safe mode
- ▶ New POWER9 cyclic redundancy check (CRC) including retry capability and spare data lane support for the processor fabric bus
- ▶ Memory ECC with Chipkill handling
- ▶ Memory scrubbing
- ▶ Memory preserving IPL
- ▶ Dynamic memory relocation
- ▶ Enhanced error handling (EEH) for all adapters
- ▶ I/O adapter concurrent maintenance with PowerVM virtualization or operating system software-based redundancy support
- ▶ Hot swap direct access storage devices (DASDs)
- ▶ At least n+1 redundancy and concurrent maintenance support for power supplies and fans of each system node
- ▶ Power cord redundancy
- ▶ Redundant vital product data (VPD)
- ▶ Emergency power-off (EPOW) reporting
- ▶ Concurrent firmware updates

Table 4-1 compares the RAS features of IBM POWER9 scale-out and POWER9 enterprise class systems.

Table 4-1 POWER9 server RAS highlight comparison

Feature	POWER9 1- and 2-socket systems ^a	Power E950 server	IBM Power System E980 server
Base POWER9 Processor RAS features: <ul style="list-style-type: none"> ▶ FFDC ▶ Processor Instruction Retry ▶ L2/L3 Cache ECC protection with cache line-delete ▶ Power/cooling monitor function that is integrated into processors' OCC ▶ CRC checked processor fabric bus retry with spare data lane 	Yes ^b	Yes	Yes
POWER9 Enterprise RAS features: <ul style="list-style-type: none"> ▶ Extended L2/L3 cache line delete ▶ Core-contained checkstops 	No	Yes	Yes

Feature	POWER9 1- and 2-socket systems ^a	Power E950 server	IBM Power System E980 server
POWER9 Multi-node Enterprise RAS: <ul style="list-style-type: none"> ▶ Across node ½ bandwidth capability ▶ Asynchronous clocking across nodes 	N/A	N/A	Yes
PCIe hot-plug with processor-integrated PCIe controller (PEC)	Yes	Yes	Yes
Memory DIMM support with ECC checking supporting x4 Chipkill	Yes	Yes	Yes
IBM memory buffer support and Spare DRAM module capability with x4 DIMMS	No	Yes	Yes
x8 DIMM with Chipkill correction for marked faulty DRAM	N/A	N/A	Yes
Custom DIMM support with extra spare DRAMs	No	No	Yes
AMM for the Hypervisor	No	Yes - Feature	Yes - Base
Redundant/spare voltage phases on voltage converters for levels feeding processor and memory DIMMs or risers	No	Redundant	Both redundant and spare
Redundant global processor clocks with concurrent failover	No	No	Yes
Redundant service processor with concurrent failover	No	No	Yes
Multi-node support	No	No	Yes

a. Power S914, Power S922, Power S924, Power H922, and Power H924.

b. Some features require PowerVM.

4.2 Reliability

Highly reliable systems are built with highly reliable components. On IBM POWER processor-based systems, this basic principle is expanded upon by using a clear design for the reliability architecture and methodology. A concentrated, systematic, and architecture-based approach improves the 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 POWER9 processor-based 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 POWER9 chip supports many cores per processor module, and the PCIe Gen4 I/O controller function is integrated into the processor module, which generates a PCIe bus directly from the processor module. Additionally, the POWER9 chip also provides compression and encryption functional units and the integrated circuit logic to attach external accelerators and devices through the Coherent Accelerator Processor Interface (CAPI), OpenCAPI, and NVLink protocols.

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 7 years. By selecting higher reliability grade components, the frequency of all failures is lowered, and the failure of parts 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).

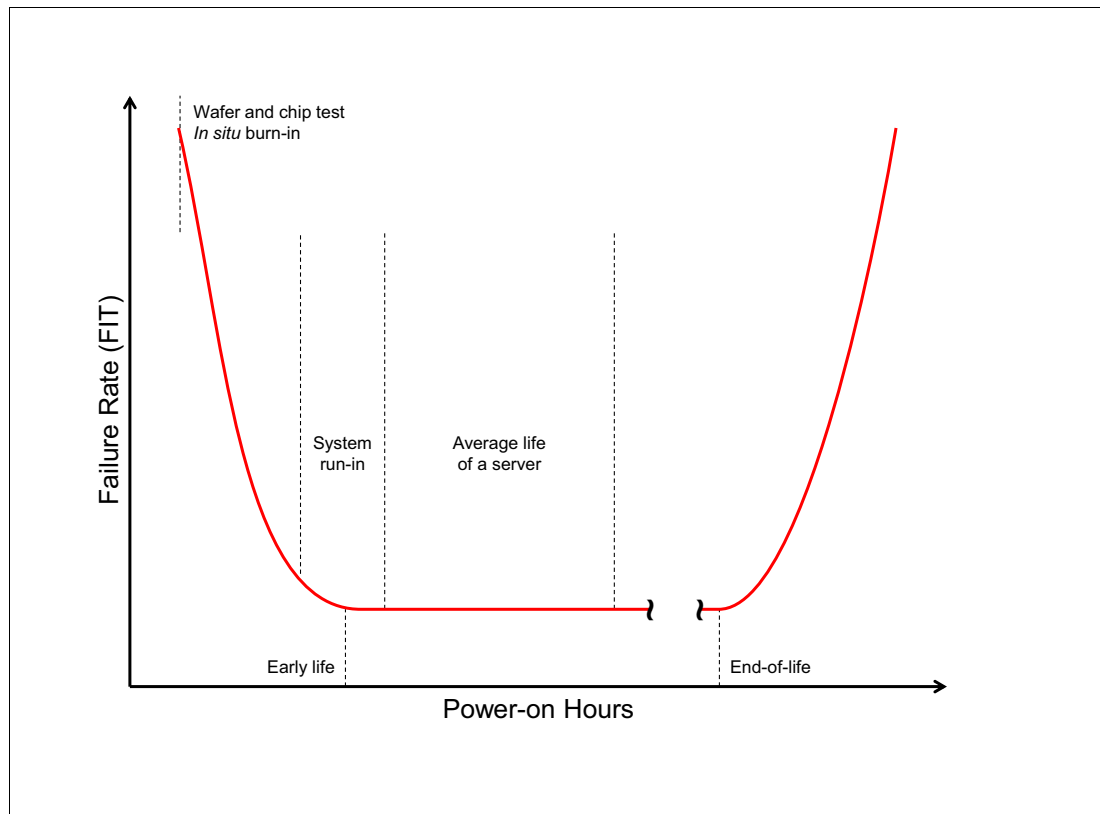


Figure 4-1 Failure rate bathtub curve

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 POWER9 processor chips, are positioned on the system board so that they receive clear air flow during operation. POWER9 systems use a premium fan with an extended life to further reduce overall system failure rate and provide adequate cooling for the critical system components.

4.3 Processor RAS details

The more reliable a system or subsystem is, the more available it should be. Nevertheless, considerable effort is made to design systems that can detect faults that 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 error detection and fault isolation (ED/FI).

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 that is built into each system. The dedicated service processor then works with the hardware to determine the course of action to be taken for each fault.

4.3.1 Correctable error introduction

Intermittent or soft errors are typically tolerated within the hardware design by using ECC 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, which 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 POWER9 processor-based systems, 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, you avoid having to replace any hardware for a solid correctable fault. 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. The more advanced RAS design of POWER processor-based systems means that in some cases the system might be able to stop the application by using the hardware that failed. This RAS design requires that uncorrectable errors are detected by the hardware and reported to software layers, and the software layers are responsible for determining how to minimize the impact of faults.

One extra advantage of the special ECC that is used in data error detection is that the hardware can distinguish between an initial ECC error that is related to a specific component of the data path and one that was passed along from earlier data transfer stages. This advantage allows the correct component, the one originating the fault, to be reported as the component to be replaced.

The advanced RAS features that are built into POWER9 processor-based systems handle certain “uncorrectable” errors in ways that minimize the impact of the faults, even keeping an entire system running after experiencing a failure.

Depending on the fault, a recovery might use the virtualization capabilities of PowerVM so 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 error handling

Layer 2 (L2) and Layer 3 (L3) caches and directories can correct single-bit errors and detect double-bit errors by using ECC (SEC/DED ECC). When a persistent correctable error occurs in these caches, the system can purge the data in the cache (writing to another level of the hierarchy) and delete it. Beyond soft error correction, the intent of the POWER9 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 restart, or individual Core Power on Reset by using the Power-On Reset Engine (PORE)).

Soft errors that are detected in the level 1 cache are also correctable by a try again operation that is handled by the hardware. But instead of using error correcting code, intermittent L1 cache errors can be corrected by using data from elsewhere in the cache hierarchy. A portion of an L1 cache can be disabled (set delete) to avoid outages due to persistent hard errors. If too many errors are observed across multiple sets, the core that uses the L1 cache can be predictively deallocated.

Separate from the system caches and the description above are cache directories that provide indexing to the caches. These also have single-bit error correction, but uncorrectable directory errors typically result in system checkstops.

Beyond soft error correction, the intent of the POWER9 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.

Beyond the L1, L2, and L3 functional units, single-bit correcting ECC is used in multiple areas of the processor as the standard means of protecting data against single-bit errors. This includes a number of the internal buses where data is passed between units.

4.3.4 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 stops 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 stops as a result of the operation. Software layers are not required to handle such faults.

Only when data is loaded to be processed by a processor core or sent out to an I/O adapter is any further action needed. In such cases, if data is used as owned by a partition, then the partition operating system might be responsible for stopping itself or just the program by using the marked page. If data is owned by the hypervisor, then the hypervisor might choose to stop, resulting in a system-wide outage.

However, the exposure to such events is minimized because cache-lines can be deleted, which eliminates the repetition of an uncorrectable fault that is in a particular cache-line.

4.3.5 Cyclic redundancy check and lane repair for processor fabric buses

ECC is used internally in various data paths as data is transmitted between processor units. However, externally to the processor, high-speed data buses can be susceptible to occasional multiple bit errors due to electrical noise, timing drift, and various other factors.

Previous POWER processor implementations such as POWER8 processors use CRC to detect multiple bit errors on the memory bus. The data can be corrected by the memory controller's ability to retry a faulty operation. If the memory bus experiences multiple CRC errors that must be corrected by retry, the memory controller can be dynamically retrained to reestablish optimal bus performance. If retraining a bus does not correct a persistent soft error, the fault might be because of a faulty bit line on the bus itself. The memory bus contains a dynamic spare bit line (dynamic memory channel repair) function that allows the memory controller to identify a persistently faulty bit line and to substitute a spare.

With the introduction of the POWER9 processor, the CRC code for error detection, the retry capability on error conditions, and the ability to substitute a faulty data lane are extended to the processor fabric bus interfaces for the Power E950 on the system board processor interconnect (X-bus).

4.3.6 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 POWER9 processor-based 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 are detected. This complements the try again abilities of the memory buffer module that is used in the Power E950 and Power E980 servers.

4.3.7 Predictive processor deallocation

Because of the amount of self-healing that is incorporated in POWER9 processor-based systems and as the extensive error recovery features that are implemented, it is rare that an entire processor core must be predictively deallocated due to a persistent recoverable error.

If such cases do occur, PowerVM can start a process for deallocating the failing processor dynamically at run time. This process interacts with the operating system that holds access to the processor in question, and requires that control over the processor be ceded by the operating system.

4.3.8 Core-contained checkstops and PowerVM handled errors

Core hardware faults that cannot be contained by processor instruction retry and the other previously described features that are defined in the hierarchy might be handled through PowerVM by a technique called core-contained checkstops.

The core-contained checkstop technology allows PowerVM to be signaled when such faults occur and stop the code that is being used by the failing processor core. This feature allows the outage that is associated with the fault to be contained to the logical partition (LPAR) by using the core that was being used when the uncorrectable fault occurred.

The core-contained checkstop feature is beneficial for scale-up IBM Power Systems servers such as the Power E950 server, which typically host many LPARs. However, a core-contained checkstop signaling that a fault occurred on a core running a hypervisor instruction typically results in hypervisor termination and a full system outage.

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.

PowerVM can handle certain other hardware faults without stopping applications, such as an error in specific data structures (faults in translation tables or lookaside buffers).

4.3.9 PCIe controller and enhanced error handling

Each processor has three elements that are called PCIe hubs that generate the various PCIe Gen4 buses that are used in the system. The hub can “freeze” operations when certain faults occur, and in certain cases can retry and recover from the fault condition. This hub freeze behavior prevents faulty data from being written out through the I/O hub system and prevents reliance on faulty data within the processor complex when certain errors are detected.

Along with this hub freeze behavior is what is termed as Enhanced Error Handling for I/O (EEH for I/O). This capability signals device drivers when various PCIe bus-related faults occur. Device drivers may attempt to restart the adapter after such faults (EEH recovery.)

A clock error in the PCIe clocking can be signaled and recovered by using EEH in any system that incorporates redundant PCIe clocks with dynamic failover enabled.

4.3.10 Memory channel checkstops and hypervisor memory mirroring

The memory controller that communicates between the processor and the memory buffer has its own set of methods for containing errors or retry operations.

Some severe faults require that memory under a portion of the controller becomes inaccessible to prevent reliance on incorrect data. There are cases where the fault can be limited to just one memory channel. In these cases, the memory controller asserts what is known as a *channel checkstop*. In systems without hypervisor memory mirroring, a channel checkstop usually results in a system outage. However, with hypervisor memory mirroring, the hypervisor continues to operate despite the memory channel checkstop.

4.3.11 Persistent guarding of failed elements

Not all processor core or processor module faults can be corrected by using the techniques that are described in this chapter. Therefore, a provision is made for faults that require 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 restart with the failed component that is unconfigured from the system.

The auto-restart (restart) option, when enabled, can restart the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced failure.

The auto-restart (restart) option must be enabled from the Advanced System Management Interface (ASMI).

4.4 Memory RAS details

One POWER9 processor-based module of the Power E950 enterprise class system provides two integrated memory controllers to facilitate access to the system's main memory. One memory controller drives four differential memory interface (DMI) channels, and every DMI channel connects to one dedicated memory buffer chip. Each memory buffer chip provides four DDR4 memory ports and one 16 MB L4 cache. One DDR4 port connects to one DIMM slot that is populated with one DDR4 industry-standard DIMM (IS DIMM).

Four memory buffer chips are mounted with their 16 associated IS DIMM slots on one memory riser card. Every processor module of a Power E950 server uses either one or two memory riser cards. The DDR4 technology-based IS DIMMs are available with 8 GB, 16 GB, 32 GB, 64 GB, and 128 GB capacity.

Figure 4-2 shows the Power E950 memory subsystem design for one POWER9 processor-based module with two memory controllers and eight DMI channels connecting to two memory riser cards.

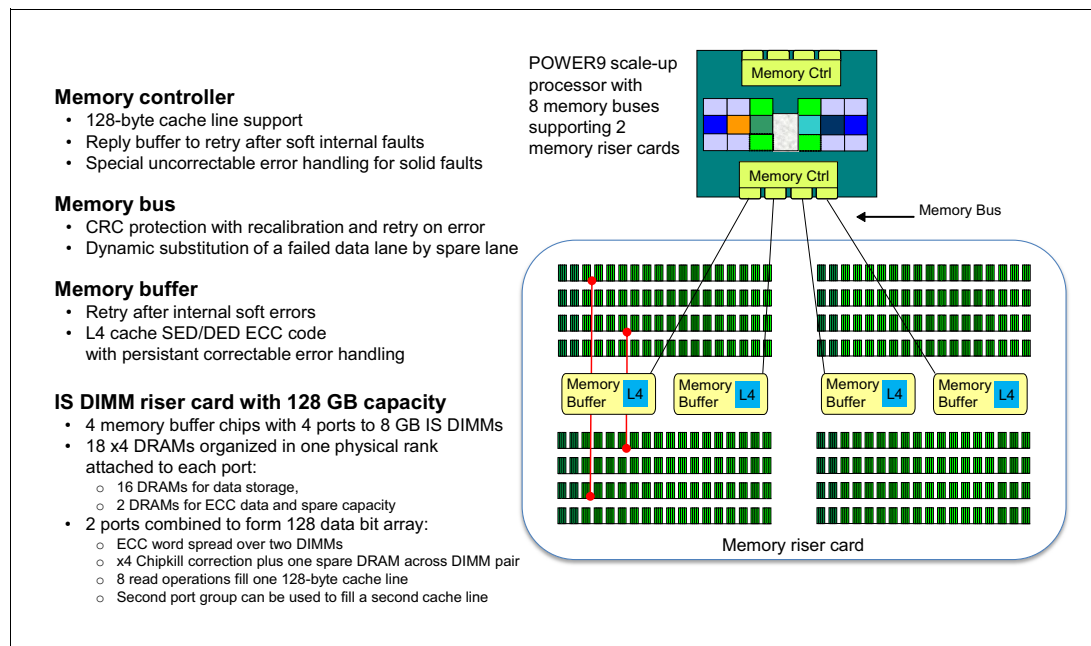


Figure 4-2 Power E950 memory protection features

The memory buffer chip is manufactured in 22-nm lithography and incorporates similar technologies that are used by POWER9 processor-based functional units to avoid soft errors. The integrated L4 cache is based on embedded DRAM (eDRAM) technology with soft error hardening and persistent error-handling features. The memory buffer implements a try again for many internally detected faults. This function complements a replay buffer in the memory controller in the processor, which also handles internally detected soft errors.

The bus between a processor memory controller and a memory buffer chip uses CRC error detection that is coupled with the ability to retry a memory access operation in case a soft error occurs. The bus features dynamic recalibration capabilities and a spare data lane that can be substituted for a failing bus lane through the recalibration process.

Each memory buffer on a memory riser card uses four ports to connect to IS DIMMs. One port gives access to one IS DIMM. The IS DIMMs are available in 8 GB, 16 GB, 32 GB, 64 GB, and 128 GB capacities. The 8 GB IS DIMMs are based on 18 DRAM modules of 4 Gb density that is organized in one physical rank. Sixteen of the x4 DRAM modules store data and two DRAM modules are available for error correction data and spare capacity. The 16 GB IS DIMM uses DRAM modules with 8 Gb density, but share the other structural parameters with the 8 GB IS DIMM. The 32 GB, 64 GB, and 128 GB IS DIMMs are using 36 DRAM modules of 8 Gb density that are organized in two physical ranks with 18 DRAM modules per rank. The 64 GB IS DIMMs are composed of two-high (2H) 3D-stacked (3DS) DRAM modules, and the 128 GB IS DIMMs are composed of four-high (4H) 3D-stacked (3DS) DRAM modules.

Two ranks on different IS DIMMs are combined into a 128-bit ESS word. The ECC that is deployed can correct the result of an entire DRAM module that is faulty per IS DIMM rank pair. This process is also known as *Chipkill correction*. Then, it can correct at least one other bit within the ECC word.

One dedicated spare DRAM module per IS DIMM rank pair is available to take over the work of a failing DRAM module. This substitution avoids the need to replace the IS DIMM for a single Chipkill event.

In addition to the protection that is provided by the ECC and sparing capabilities, the memory subsystem implements memory scrubbing to identify and correct single-bit soft errors. The PowerVM Hypervisor is informed about incidents of single-cell persistent (hard) faults for deallocation of associated pages. However, because of the ECC and sparing capabilities that are used, such memory page deallocation is not relied upon for repair of faulty hardware.

Finally, should an uncorrectable error in data be encountered, the memory that is affected is marked with a Special Uncorrectable Error (SUE) code and handled as described in 4.3.4, “Cache uncorrectable error handling” on page 130.

4.5 PCIe I/O subsystem RAS details

In POWER8 processor-based systems, the external I/O hub and bridge adapters were eliminated in favor of a topology that integrates the PEC and the PCIe host bridge (PHB) logic into the processor module. PCIe buses that are generated directly from a PHB can drive individual I/O slots or a PCIe switch. The integrated PEC supports try again (end-point error recovery) and freezing features. With POWER9 processors, this design was carried forward to support PCIe 4.0 technology.

4.5.1 I/O subsystem availability and enhanced error handling

Multi-path I/O and Virtual I/O Server (VIOS) for I/O adapters and RAID for storage devices must be used to prevent application outages when I/O adapter faults occur.

To permit soft or intermittent faults to be recovered without failover to an alternative device or I/O path, Power Systems hardware supports 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 stop if there is an intermittent hard error or other unrecoverable errors while protecting against reliance on data that cannot be corrected. This action often is done by “freezing” access to the I/O subsystem with the fault. Freezing prevents data from flowing to and from an I/O adapter and causes the hardware or firmware to respond with a defined error signature whenever an attempt is made to access the device. If necessary, a SUE code can be used to mark a section of data as bad when the freeze is first started.

IBM device drivers under AIX are fully EEH-capable. For Linux under PowerVM, EEH support extends to many frequently used devices. There might be various third-party PCI devices that do not provide native EEH support.

4.5.2 PCIe Gen3 I/O Expansion drawer RAS

PCIe Gen3 I/O Expansion Drawers (#EMX0) can be used with Power E950 systems to increase I/O capacity. Figure 4-3 shows the functional components of the #EMX0 drawer.

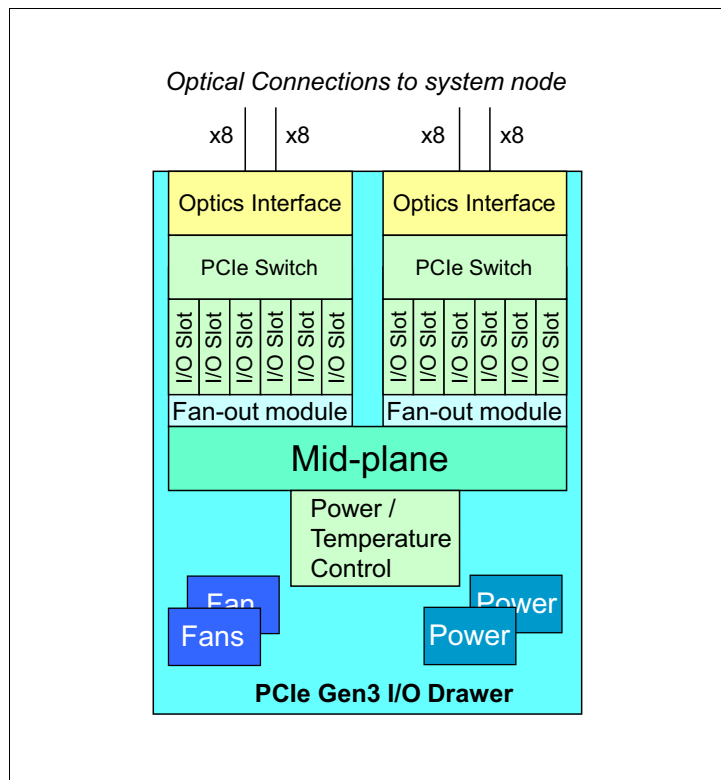


Figure 4-3 PCI3 Gen3 Expansion Drawer structural diagram

These I/O drawers are attached by using a connecting card that is called a PCIe3 Optical Cable Adapter (#EJ07) that plugs in to a PCIe slot of a Power E950 system node. The cable cards for POWER9 processor-based servers are redesigned in certain areas to improve error handling. These improvements include new routing for clock logic within the cable card and extra recovery for faults during IPL.

Each I/O drawer contains up to two PCIe FanOut Modules (#EMXG). An I/O module uses x16 PCIe lanes that are controlled from a processor in a system node. An I/O module that uses a PCIe switch to supply six PCIe slots is supported.

Two Active Optical Cables (AOCs) are used to connect a PCIe3 cable adapter to the equivalent card in the I/O drawer module. Although these cables are not redundant (since the FW830 firmware), the loss of one cable reduces the I/O bandwidth (the number of lanes that is available to the I/O module) by 50%.

Infrastructure RAS features for the I/O drawer include redundant power supplies, fans, and DC outputs of voltage regulators (phases).

The impact of the failure of an I/O drawer component is summarized for the most significant cases in Table 4-2.

Table 4-2 PCIe Gen3 I/O Expansion Drawer RAS feature matrix

Faulty component	Impact of failure	Impact of repair	Prerequisites
I/O adapter in an I/O slot.	Loss of function of the I/O adapter.	I/O adapter can be repaired while the rest of the system continues to operate.	Multipathing I/O adapter redundancy, where implemented, can be used to prevent application outages.
First fault on a data lane (in the optics between the PCIe3 cable adapter in the system and the I/O module).	None: Spare used.	No repair needed: Integrated sparing feature.	None.
A second data lane fault or other failure of one active optics cable.	System continues to run, but the number of active lanes that are available to the I/O module is reduced.	The associated I/O module must be taken down for repair; the rest of the system may remain active.	None.
Other failure of PCIe3 cable adapter in the system or I/O module.	Loss of access to all the I/O of the connected I/O module.	The associated I/O module must be taken down for repair; the rest of the system can remain active.	Systems with a Hardware Management Console (HMC).
One fan.	System continues to run with remaining fan.	Concurrently repairable.	None.
One power supply.	System continues to run with remaining power supply.	Concurrently repairable.	None.
VRM associated with an I/O module.	System continues to run for a phase failure transition to n mode. Other faults impact all the I/O in the module.	The associated I/O module cannot be active during repair; the rest of the system can remain active.	Systems with an HMC.

Faulty component	Impact of failure	Impact of repair	Prerequisites
Chassis Management Card (CMC).	No impact to the running system, but after it is powered off, the I/O drawer cannot be reintegrated until the CMC is repaired.	The I/O drawer must be powered off to repair (loss of use of all I/O in the drawer).	Systems with an HMC.
Midplane.	Depending on the source of the failure, this failure might take down the entire I/O drawer.	The I/O drawer must be powered off to repair (loss of use of all I/O in the drawer).	Systems with an HMC.

4.6 Enterprise systems availability

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

The following advanced RAS features pertain to the Power E950 enterprise class system:

- ▶ **Dynamic Processor Sparing**

Enterprise class systems are Capacity Upgrade on Demand (CUoD)-capable. Processor sparing helps minimize the effect on server performance that is caused by a failed processor. An inactive processor is activated if a failing processor reaches a predetermined error threshold, which helps to maintain performance and improve system availability.

Dynamic processor sparing happens dynamically and automatically when dynamic logical partitioning (DLPAR) is used and the failing processor is detected before failure. Dynamic processor sparing does not require purchasing an activation code. Instead, it requires only that the system have inactive CUoD processor cores available.

- ▶ **Dynamic Memory Sparing**

Enterprise class systems are CUoD-capable. Dynamic memory sparing helps minimize the effect on server performance that is caused by a failed memory feature. Memory sparing occurs when on-demand inactive memory is automatically activated by the system to temporarily replace failed memory until a service action can be performed.

- ▶ **AMM for Hypervisor**

The hypervisor is the core part of the virtualization layer. Although minimal, its operational data must be in memory DIMMs. If there is a failure of DIMM, the hypervisor can become inoperative. The AMM for Hypervisor allows for the memory blocks that are used by the hypervisor to be written in two distinct DIMMs. If an uncorrectable error is encountered during a read, the data is retrieved from the mirrored pair and operations continue normally. AMM for Hypervisor is an optional configurable feature of the Power E950 server.

4.7 Availability effects of a solution architecture

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

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

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

- ▶ Will you need to move your workloads off an entire server during service or planned outages?
- ▶ If you use a clustering solution to move the workloads, how will the failover time affect your services?
- ▶ If you use a server evacuation solution to move the workloads, how long does it take to migrate all the partitions with your current server configuration?

4.7.1 Clustering

A Power Systems server that is running under PowerVM, AIX, and Linux support many clustering solutions. These solutions meet requirements for application availability regarding server outages and data center disaster management, reliable data backups, and so on. These offerings include distributed applications with IBM Db2® PureScale, HA solutions that use clustering technology with IBM PowerHA® SystemMirror®, and disaster management across geographies with PowerHA SystemMirror Enterprise Edition.

For more information, see the following resources:

- ▶ *IBM PowerHA SystemMirror for i: Using Geographic Mirroring (Volume 4 of 4)*
- ▶ *IBM PowerHA SystemMirror for i: Using IBM Storwize (Volume 3 of 4)*
- ▶ *IBM PowerHA SystemMirror for i: Using DS8000 (Volume 2 of 4)*
- ▶ *IBM PowerHA SystemMirror for i: Preparation (Volume 1 of 4)*
- ▶ *IBM PowerHA SystemMirror V7.2.1 for IBM AIX Updates*
- ▶ *IBM PowerHA SystemMirror V7.2 for IBM AIX Updates*
- ▶ *Guide to IBM PowerHA SystemMirror for AIX Version 7.1.3, SG24-8167*
- ▶ *IBM PowerHA SystemMirror for AIX Cookbook, SG24-7739*

4.7.2 Virtual I/O redundancy configurations

Within each server, the partitions can be supported by a single VIOS. However, if a single VIOS is used and that VIOS stops for any reason (hardware or software caused), all of the partitions that use that VIOS stop.

The usage of redundant VIOS servers mitigates this risk. Maintaining the redundancy of adapters within each VIOS (in addition to having redundant VIOSes) avoids most faults that keep a VIOS from running. Therefore, multiple paths to networks and SANs are advised.

A partition that is accessing data from two distinct VIOSes, each one with multiple network and SAN adapters to provide connectivity, is shown in Figure 4-4.

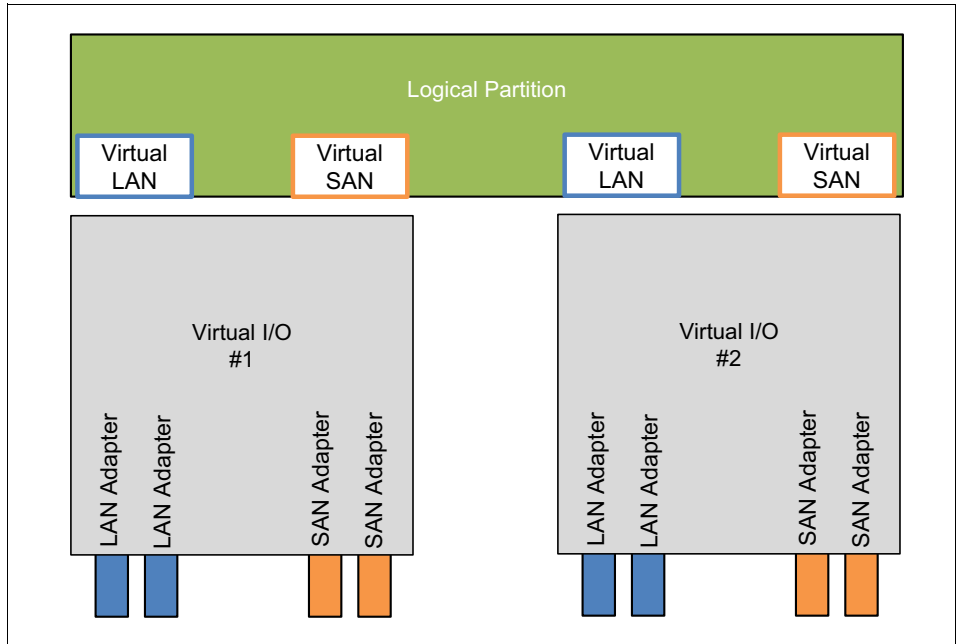


Figure 4-4 Partition with dual redundant Virtual I/O Servers

Because each VIOS can be considered as an AIX based partition, each VIOS also must access a boot image, have paging space, and so on, under a root volume group or rootvg. The rootvg can be accessed through a SAN, the same as the data that partitions use.

Alternatively, a VIOS can use the U.2 NVMe internal storage option available for Power E950 systems. The Power E950 server supports four U.2 NVMe devices, which can be individually assigned to independent partitions. The NVMe devices are accessible from the front of the system and are concurrent maintainable. To use storage that is locally attached (DASD devices or solid-state drives (SSDs)) through SAS adapters offers another option to provide boot devices to VIOS partitions. However they are accessed, the rootvgs should use mirrored or RAID drives with redundant access to the devices for best availability.

4.7.3 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility (LPM) allows you to move a running LPAR (including its operating system and running applications) from one system to another without any shutdown and without disrupting the operation of that LPAR. Inactive partition mobility allows you to move a powered-off LPAR from one system to another.

LPM provides systems management flexibility and improves system availability through the following functions:

- ▶ Avoid planned outages for hardware or firmware maintenance by moving LPARs to another server and then performing the maintenance. LPM 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 LPARs to another server and then performing the upgrade. This approach allows your users to continue their work without disruption.
- ▶ Avoid unplanned downtime. With preventive failure management, you can move a server's LPARs to another server before the failure occurs if a server indicates a potential failure. Partition mobility can help avoid unplanned downtime.
- ▶ Take advantage of server optimization:
 - Consolidation: You can consolidate workloads that run on several small, underused servers onto a single large server.
 - Deconsolidation: You can move workloads from server-to-server to optimize resource use and workload performance within your computing environment. With LPM, you can manage workloads with minimal downtime.

Server Evacuation: This PowerVM function allows you to perform a server evacuation operation. Server Evacuation is used to move all migration-capable LPARs from one system to another if there are no active migrations in progress on the source or the target servers.

With the Server Evacuation feature, multiple migrations can occur based on the concurrency setting of the HMC. Migrations are performed as sets, with the next set of migrations starting when the previous set completes. Any upgrade or maintenance operations can be performed after all the partitions are migrated and the source system is powered off.

You can migrate all the migration-capable AIX, IBM i, and Linux partitions from the source server to the destination server by running the following command from the HMC command line:

```
migr\lpar -o m -m source_server -t target_server --all
```

Hardware and operating system requirements for Live Partition Mobility

LPM is supported by default with enterprise systems. It also is supported by all operating systems that are compatible with POWER9 processor-based technology.

The VIOS partition cannot be migrated.

For more information about LPM and how to implement it, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

4.8 Serviceability

The purpose of serviceability is to repair the system while attempting to minimize or eliminate service cost (within budget objectives) and maintain application availability and high customer satisfaction. Serviceability includes system installation, Miscellaneous Equipment Specification (MES) (system upgrades or fallbacks), and system maintenance or repair. Depending on the system and warranty contract, service might be performed by the customer, an IBM System Services Representative (SSR), or an authorized warranty service provider.

The serviceability features that are delivered in this system provide a highly efficient service environment by incorporating the following attributes:

- ▶ Design for SSR Set Up and Customer Installed Features (CIFs).
- ▶ ED/FI.
- ▶ FFDC.
- ▶ The Guiding Light service indicator architecture is used to control a system of integrated LEDs that lead the individual servicing the machine to the correct part as quickly as possible.
- ▶ Service labels, service cards, and service diagrams are available on the system and delivered through the HMC.
- ▶ Step-by-step service procedures are available through the HMC.

This section provides an overview of how these attributes contribute to efficient service in the progressive steps of error detection, analysis, reporting, notification, and repair found in all POWER processor-based systems.

4.8.1 Detecting errors

The first and most crucial component of a solid serviceability strategy is the ability to detect accurately and effectively errors when they occur.

Although not all errors are a threat to system availability, those errors that go undetected can cause problems because the system has no opportunity to evaluate and act if necessary. POWER processor-based systems employ IBM Z® server-inspired error detection mechanisms, which extend from processor cores and memory to power supplies and hard disk drives (HDDs).

4.8.2 Error checkers, fault isolation registers, and first failure data capture

POWER processor-based systems contain specialized hardware detection circuitry that is used to detect erroneous hardware operations. Error-checking hardware ranges from parity error detection that is coupled with Processor Instruction Retry and bus try again, to ECC correction on caches and system buses.

Within the processor/memory subsystem error checker, error-checker signals are captured and stored in hardware FIRs. The associated logic circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, runtime error diagnostic tests can be deterministic so that for every check station, the unique error domain for that checker is defined and mapped to FRUs that can be repaired when necessary.

Integral to the Power Systems design is the concept of FFDC. FFDC is a technique that involves sufficient error-checking stations and co-ordination of faults so that faults are detected and the root cause of the fault is isolated. FFDC also expects that necessary fault information can be collected at the time of failure without needing to re-create the problem or run an extended tracing or diagnostics program.

For many faults, a good FFDC design means that the root cause is isolated at the time of the failure without intervention by an IBM SSR. For all faults, good FFDC design still makes failure information available to the IBM SSR. This information can be used to confirm the automatic diagnosis. More detailed information can be collected by an IBM SSR for rare cases where the automatic diagnosis is not adequate for fault isolation.

4.8.3 Service processor

In POWER9 processor-based systems, the Flexible Service Processor (FSP) is a microprocessor that is powered separately from the main instruction processing complex.

The service processor performs the following serviceability functions:

- ▶ Several remote power control options
- ▶ Reset and boot features
- ▶ Environmental monitoring

The service processor interfaces with the OCC function, which monitors the server's built-in temperature sensors and sends instructions to the system fans to increase the rotational speed when the ambient temperature is above the normal operating range. By using a designed operating system interface, the service processor notifies the operating system of potential environmentally related problems so that the system administrator can take appropriate corrective actions before a critical failure threshold is reached. The service processor can also post a warning and start an orderly system shutdown in the following circumstances:

- The operating temperature exceeds the critical level (for example, failure of air conditioning or air circulation around the system).
- The system fan speed is out of operational specification (for example, because of multiple fan failures).
- The server input voltages are out of operational specification. The service processor can shut down a system in the following circumstances:
 - The temperature exceeds the critical level or remains above the warning level for too long.
 - Internal component temperatures reach critical levels.
 - Non-redundant fan failures occur.
- ▶ PowerVM Hypervisor (system firmware) and HMC connection surveillance

The service processor monitors the operation of the firmware during the boot process and monitors the hypervisor for termination. The hypervisor monitors the service processor and can perform a reset and reload if it detects the loss of the service processor. If the reset/reload operation does not correct the problem with the service processor, the hypervisor notifies the operating system, which can then take appropriate action, including calling for service. The FSP also monitors the connection to the HMC and can report loss of connectivity to the operating system partitions for system administrator notification.

- ▶ **Uncorrectable error recovery**

When enabled, the auto-restart (restart) option can restart the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced (power) failure.

The auto-restart (restart) option must be enabled from the ASMI.

- ▶ **Concurrent access to the service processors menus of the ASMI**

This access allows nondisruptive abilities to change system default parameters, interrogate service processor progress and error logs, set and reset service indicators (Guiding Light for enterprise servers), and access all service processor functions without powering down the system to the standby state.

The administrator or IBM SSR can access dynamically the menus from any web browser-enabled console that is attached to the Ethernet service network concurrently with normal system operation. Some options, such as changing the hypervisor type, do not take effect until the next restart.

- ▶ **Managing the interfaces for connecting uninterruptible power source systems to the POWER processor-based systems and performing timed power-on (TPO) sequences.**

4.8.4 Diagnosing

General diagnostic objectives are created to detect and identify problems so that they can be resolved quickly. The IBM diagnostic strategy includes the following elements:

- ▶ Provide a common error code format that is equivalent to a System Reference Code (SRC), 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 an IBM back-end system.

By using the extensive network of advanced and complementary error detection logic that is built directly into hardware, firmware, and operating systems, the Power Systems servers can perform considerable self-diagnosis.

Because of the FFDC technology that is designed into IBM servers, re-creating diagnostic tests for failures or requiring user intervention is unnecessary. Solid and intermittent errors are correctly detected and isolated at the time that the failure occurs. Runtime and boot time diagnostic tests fall into this category.

Boot time

When a Power Systems server starts, the service processor initializes the system hardware. Boot time diagnostic testing uses a multitier approach for system validation, starting with managed low-level diagnostic tests that are supplemented with system firmware initialization and configuration of I/O hardware, followed by OS-initiated software test routines.

To minimize boot time, the system determines which of the diagnostic tests are required to be started to ensure correct operation. This determination based on the way that the system was powered off or on the boot-time selection menu.

Host Boot IPL

On POWER9 processor-based systems, the boot process is initialized by the FSP, and one part of the running firmware performs the central electronics complex chip initialization. A component that is external to the POWER9 processor that is called the Power NOR (PNOR) chip stores the Host Boot (HB) firmware and the self-boot engine (SBE) code. During system IPL, the PNOR starts the SBE, which eventually loads the HB base code onto the POWER9 chip.

With this HB initialization, new progress codes are available. An example of an FSP progress code is C1009003. During the HB IPL, progress codes, such as CC009344, appear.

If there is a failure during the HB process, a new HB system memory dump is collected and stored. This type of memory dump includes HB memory and is offloaded to the HMC when it is available.

Processor Runtime Diagnostics

All Power Systems servers can monitor critical system components during run time. They also can take corrective actions when recoverable faults occur. In POWER9 processor-based systems that are virtualized through PowerVM, the IBM hardware error-check architecture with the FFDC implementation reports errors in the central electronics complex to a special service partition that is under the control of the PowerVM Hypervisor. The special service partition runs the Processor Runtime Diagnostics (PRD) code, which ingests and processes the error information and directs the error management. The hypervisor can restart the special service partition and reload the PRD code in case the Runtime Diagnostics service becomes unavailable. On previous POWER7 and POWER8 processor-based systems, the PRD code was run on the FSP.

Extensive diagnostic and fault analysis routines were developed and improved over many generations of POWER processor-based servers. These routines enable quick and accurate predefined responses to actual and potential system problems. The PRD code running in the special service partition 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 an LPAR. 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 operating system can start the Call Home application to report the service event to IBM. For optional HMC-managed servers, the event is reported to the HMC, which can start the Call Home request to IBM. I/O devices can also include specific exercisers that can be started by the diagnostic facilities for problem recreation (if required by service procedures).

4.8.5 Reporting

If a system hardware or environmentally induced failure is detected, Power Systems servers report the error through various mechanisms. The analysis result is stored in system NVRAM. You can use error log analysis (ELA) to display the failure cause and the physical location of the failing hardware.

Using the Call Home infrastructure, the system automatically can send an alert through a phone line to a pager, or call for service if there is a critical system failure. A hardware fault also illuminates the amber system fault LED that is on the system node to alert the user of an internal hardware problem.

On POWER9 processor-based servers, hardware and software failures are recorded in the system log. When a management console is attached, an ELA routine analyzes the error, forwards the event to the Service Focal Point (SFP) application that is running on the management console, and notifies the system administrator that it isolated a likely cause of the system problem. The service processor event log also records unrecoverable checkstop conditions, forwards them to the SFP application, and notifies the system administrator.

After the information is logged in the SFP application, a Call Home service request is started and the pertinent failure data with service parts information and part locations is sent to the IBM service organization if the system is correctly configured. This information also contains the client contact information that is defined in the IBM Electronic Service Agent (ESA) guided setup wizard. In HMC V8R8.1.0, a Serviceable Event Manager is available to block problems from being automatically transferred to IBM. For more information, see “Service Event Manager” on page 160.

Error logging and analysis

When the root cause of an error is identified by a fault isolation component, an error log entry is created with the following types of basic data:

- ▶ An error code that uniquely describes the error event.
- ▶ The location of the failing component.
- ▶ The part number of the component to be replaced, including pertinent data, such as engineering and manufacturing levels.
- ▶ Return codes.
- ▶ Resource identifiers.
- ▶ FFDC data.

Data that contains information about the effect that the repair has on the system is also included. Error log routines in the operating system and FSP can then use this information and decide whether the fault is a Call Home candidate. If the fault requires support intervention, a call is placed with service and support. A notification is sent to the contact that is defined in the ESA-guided setup wizard.

Remote support

The Remote Management and Control (RMC) subsystem is delivered as part of the base operating system, which includes the operating system that runs on the HMC. RMC provides a secure transport mechanism across the local area network (LAN) interface between the operating system and the optional HMC and is used by the operating system diagnostic application for transmitting error information. It performs several other functions, but those functions are not used for the service infrastructure.

Service Focal Point application for partitioned systems

A critical requirement in a logically partitioned environment is to ensure that errors are not lost before being reported for service. Also, an error should be reported only once, regardless of how many LPARs experience the potential effect of the error. The SFP application on the management console or in the Integrated Virtualization Manager (IVM) is responsible for aggregating duplicate error reports, and ensures that all errors are recorded for review and management. The SFP application provides other service-related functions, such as controlling service indicators, setting up Call Home, and providing guided maintenance.

When a local or globally reported service request is made to the operating system, the operating system diagnostic subsystem uses the RMC subsystem to relay error information to the optional HMC. For global events (platform unrecoverable errors, for example), the service processor also forwards error notification of these events to the HMC, which provides a redundant error-reporting path in case the errors are in the RMC subsystem network.

The first occurrence of each failure type is recorded in the Manage Serviceable Events task on the management console. This task then filters and maintains a history of duplicate reports from other LPARs or from the service processor. It then looks at all active service event requests within a predefined timespan, analyzes the failure to ascertain the root cause and, if enabled, starts a Call Home for service. This methodology ensures that all platform errors are reported through at least one functional path, which results in a single notification for a single problem. Similar service functions are provided through the SFP application on the IVM for providing service functions and interfaces on non-HMC partitioned servers.

Extended error data

Extended error data (EED) is data that is collected automatically at the time of a failure or manually later. Although the data that is collected depends on the invocation method, it includes information, such as firmware levels, operating system levels, other FIR values, recoverable error threshold register values, and system status.

The data is formatted and prepared for transmission back to IBM to assist the service support organization with preparing a service action plan for the IBM SSR or for more analysis.

System memory dump handling

In certain circumstances, an error might require a memory dump to be automatically or manually created. In this event, the memory dump can be offloaded to the optional HMC. Specific management console information is included as part of the information that can be sent to IBM Support for analysis. If more information that relates to the memory dump is required, or if viewing the memory dump remotely becomes necessary, the management console memory dump record notifies the IBM Support center regarding on which managements console the memory dump is. If no management console is present, the memory dump might be on the FSP or in the operating system, depending on the type of memory dump that was started and whether the operating system is operational.

4.8.6 Notifying

After a Power Systems server detects, diagnoses, and reports an error to an appropriate aggregation point, it notifies the client and, if necessary, the IBM Support organization. Depending on the assessed severity of the error and support agreement, this client notification might range from a simple notification to having field service personnel automatically dispatched to the client site with the replacement part.

Client Notify

When an event is important enough to report but does not indicate the need for a repair action or to call home to IBM Support, it is classified as *Client Notify*. Clients are notified because these events might be of interest to an administrator. The event might be a symptom of an expected systemic change, such as a network reconfiguration or failover testing of redundant power or cooling systems, including the following examples:

- ▶ Network events, such as the loss of contact over a LAN
- ▶ Environmental events, such as ambient temperature warnings
- ▶ Events that need further examination by the client (although these events do not necessarily require a part replacement or repair action)

Client Notify events are serviceable events because they indicate that something happened that requires client awareness if the client wants to take further action. These events can be reported to IBM at the discretion of the client.

Call Home

Call Home refers to an automatic or manual call from a customer location to an IBM Support structure with error log data, server status, or other service-related information. The Call Home feature starts the service organization so that the appropriate service action can begin. Call Home can be done through HMC or most non-HMC managed systems.

Although configuring a Call Home function is optional, clients are encouraged to implement this feature to obtain service enhancements, such as reduced problem determination and faster and potentially more accurate transmission of error information. The use of the Call Home feature can result in increased system availability. The ESA application can be configured for automated Call Home. For more information, see 4.9.4, “Electronic Services and Electronic Service Agent” on page 158.

Vital product data and inventory management

Power Systems servers store VPD internally, which keeps a record of how much memory is installed, how many processors are installed, the manufacturing level of the parts, and so on. These records provide valuable information that can be used by remote support and IBM SSRs, which enables the IBM SSRs to help keep the firmware and software current on the server.

IBM Service and Support Problem Management database

At the IBM Support center, historical problem data is entered into the IBM Service and Support Problem Management database. All of the information that is related to the error, along with any service actions that are taken by the IBM SSR, is recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

4.8.7 Locating and servicing

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts that require service. POWER processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Packaging for service

The following service enhancements are included in the physical packaging of the systems to facilitate service:

- ▶ Color coding (touch points)
Blue-colored touch points delineate components that cannot be concurrently maintained (they might require that the system is turned off for removal or repair).
- ▶ Tool-less design
Selected IBM systems support tool-less or simple tool designs. These designs require no tools (or require basic tools such as flathead screw drivers) to service the hardware components.
- ▶ Positive retention
Positive retention mechanisms help ensure proper connections between hardware components, such as from cables to connectors, and between two cards that attach to each other. Without positive retention, hardware components risk become loose during shipping or installation, which prevents a good electrical connection. Positive retention mechanisms, such as latches, levers, thumb-screws, pop Nylatches (U-clips), and cables are included to help prevent loose connections and aid in installing (seating) parts correctly. These positive retention items do not require tools.

Guiding Light

High-end systems are usually repaired by IBM Support personnel. The enclosure and system identify LEDs that are on solid, and can be used to follow the path from the system to the enclosure and down to the specific FRU.

Guiding Light uses a series of flashing LEDs, allowing a service provider to quickly and easily identify the location of system components. Guiding Light can also handle multiple error conditions simultaneously, which might be necessary in some complex high-end configurations.

In these situations, Guiding Light waits for the servicer's indication of what failure to attend first and then illuminates the LEDs to the failing component.

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

IBM Knowledge Center

[IBM Knowledge Center](#) provides you with a single information center where you can access product documentation for IBM systems hardware, operating systems, and server software. The latest version of the documentation is accessible on the internet.

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

Service labels

Service providers use these labels to assist with maintenance actions. Service labels are in various formats and positions and are intended to transmit readily available information to the IBM SSR during the repair process.

The following service labels are available:

- ▶ Location diagrams

These diagrams are strategically positioned on the system hardware and relate information about the placement of hardware components. Location diagrams can include location codes, drawings of physical locations, concurrent maintenance status, or other data that is pertinent to a repair. Location diagrams are especially useful when multiple components are installed, such as DIMMs, sockets, processor cards, fans, adapter, LEDs, and power supplies.

- ▶ Remove or replace procedure labels

These labels contain procedures that often are found on a cover of the system or in other locations that are accessible to the IBM SSR. These labels provide systematic procedures (including diagrams) that describe how to remove and replace certain serviceable hardware components.

- ▶ Numbered arrows

These arrows are used to indicate the order of operation and serviceability direction of components. Various serviceable parts, such as latches, levers, and touch points, must be pulled or pushed in a certain direction and order so that the mechanisms can engage or disengage. Arrows often improve the ease of serviceability.

Operator panel

The operator panel on a POWER processor-based system is an LCD display (two rows by 16 elements) that is used to present boot progress codes, which indicate advancement through the system power-on and initialization processes. The operator panel also is used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, which enable an IBM SSR or client to change various boot-time options, and for other limited service functions. The operator panel of the Power E950 server is composed of a base unit and a separate LCD unit that are individually concurrent maintainable.

Concurrent maintenance

The IBM POWER9 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 concurrently maintainable when properly configured. Concurrent maintenance is facilitated by the redundant design for the power supplies, fans, and physical storage.

In addition to these components, the operator panel can be replaced concurrently by using the service functions of the ASMI menu.

Repair and verify services

Repair and verify (R&V) services are automated service procedures that are used to guide a service provider step-by-step through the process of repairing a system and verifying that the problem was repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being serviced. The following scenarios are covered by R&V services:

- ▶ Replacing a defective FRU or a CRU
- ▶ Reattaching a loose or disconnected component
- ▶ Correcting a configuration error

- ▶ Removing or replacing an incompatible FRU
- ▶ Updating firmware, device drivers, operating systems, middleware components, and IBM applications after replacing a part

R&V procedures can be used by user engineers and IBM SSR providers who are familiar with the task and those engineers and providers who are not. Education-on-demand content is placed in the procedure at the appropriate locations. Throughout the R&V procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event to ensure that the guided maintenance procedures are operating correctly.

Clients can subscribe through the subscription services on the [IBM Support Portal](#) to obtain notifications about the latest updates that are available for service-related documentation.

4.9 Manageability

Several functions and tools help manageability so you can efficiently and effectively manage your system.

4.9.1 Service user interfaces

The service interface allows support personnel or the client to communicate with the service support applications in a server by using a console, interface, or terminal. Delivering a clear, concise view of available service applications, the service interface allows the support team to manage system resources and service information in an efficient and effective way. Applications that are available through the service interface are carefully configured and placed to give service providers access to important service functions.

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

- ▶ Guiding Light (see “Guiding Light” on page 148 and “Service labels” on page 148)
- ▶ Service processor and ASMI
- ▶ Operator panel
- ▶ Operating system service menu
- ▶ SFP on the HMC

Service processor

The service processor is a controller that is running its own operating system. It is a component of the service interface card. The service processor operating system includes specific programs and device drivers for the service processor hardware. The host interface is a processor support interface that is connected to the POWER processor.

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, which ensures the connection to the management console for manageability purposes and for accepting ASMI Secure Sockets Layer (SSL) network connections. The service processor can view and manage the machine-wide settings by using the ASMI. It also enables complete system and partition management from the HMC.

Analyzing a system that does not boot: The FSP can analyze a system that does not boot. Reference codes and detailed data is available in the ASMI and are transferred to the HMC.

The service processor uses two Ethernet ports that run at 1 Gbps speed. Consider the following points:

- ▶ Both Ethernet ports are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI. The ASMI options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- ▶ Both Ethernet ports support only auto-negotiation. Customer-selectable media speed and duplex settings are not available.
- ▶ The Ethernet ports have the following default IP addresses:
 - Service processor eth0 (HMC1 port) is configured as 169.254.2.147.
 - Service processor eth1 (HMC2 port) is configured as 169.254.3.147.

The following functions are available through the service processor:

- ▶ Call Home
- ▶ ASMI
- ▶ Error information (error code, part number, and location codes) menu
- ▶ View of guarded components
- ▶ Limited repair procedures
- ▶ Generate dump
- ▶ LED Management menu
- ▶ Remote view of ASMI menus
- ▶ Firmware update through a USB key

Advanced System Management Interface

ASMI is the interface to the service processor with which you manage the operation of the server, such as auto-power restart. You also can view information about the server, such as the error log and VPD. Various repair procedures require connection to the ASMI.

The ASMI is accessible through the management console. It is also accessible by using a web browser on a system that is connected directly to the service processor (in this case, a standard Ethernet cable or a crossed cable) or through an Ethernet network. ASMI can also be accessed from an ASCII terminal, but this option is available only while the system is in the platform powered-off mode.

Use the ASMI to change the service processor IP addresses or to apply certain security policies and prevent access from unwanted IP addresses or ranges.

You might use the service processor's default settings to operate your server. If the default settings are used, accessing the ASMI is not necessary. To access ASMI, use one of the following methods:

► **Management console**

If configured to do so, the management console connects directly to the ASMI for a selected system from this task.

To connect to the ASMI from a management console, complete the following steps:

- a. Open **Systems Management** from the navigation pane.
- b. From the work window, select one of the managed systems.
- c. From the System Management tasks list, click **Operations** → **Launch Advanced System Management (ASM)**.

► **Web browser**

At the time of writing, supported web browsers are Netscape 9.0.0.4, Microsoft Internet Explorer 7.0, Opera 9.24, and Mozilla Firefox 2.0.0.11. 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 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 Microsoft Internet Explorer, click **Tools Internet Options**. Clear the **Use TLS 1.0** option, and click **OK**.

► **ASCII terminal**

The ASMI on an ASCII terminal supports a subset of the functions that are provided by the web interface and is available only when the system is in the platform powered-off mode. The ASMI on an ASCII console is not available during several phases of system operation, such as the IPL and run time.

► **Command-line start of the ASMI**

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

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

On the HMC, a browser window opens automatically with the ASMI window and, when configured properly, a browser window opens on a remote system when issued from there.

Operator panel

The operator panel of the Power E950 server is composed of a base unit and a separate LCD unit that are individually concurrent maintainable. The base operator panel is always included with a Power E950 system, but the LCD unit is not needed if the system is under the control of an HMC.

The operator panel is used to present boot progress codes, which indicate advancement through the system power-on and initialization processes. The operator panel also is used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, which enable an IBM SSR or client to change various boot-time options and for other limited service functions.

The base operator panel provides LEDs and sensors:

- ▶ Power LED:
 - Color: Green.
 - Off: Enclosure is off (AC cord is not connected.)
 - On Solid: Enclosure is powered on.
 - On Blink: Enclosure is in the standby-power state.
- ▶ Enclosure Identify LED:
 - Color: Blue.
 - Off: Normal.
 - On Solid: Identify State.
- ▶ System Fault LED:
 - Color: Amber.
 - Off: Normal.
 - On Solid: Check Error Log.
- ▶ System Roll-up LED:
 - Color: Amber.
 - Off: Normal.
 - On Solid: Fault.
- ▶ Power Button
- ▶ System Reset Switch
- ▶ Two Thermal Sensors
- ▶ One Pressure/Altitude Sensor

The LCD operator panel features two rows of 16 characters and increment, decrement, and Enter buttons.

The following functions are available through the operator panel:

- ▶ Error information.
- ▶ Generate memory dump.
- ▶ View machine type, model, and serial number.
- ▶ View or change the IP addresses of the service processor.
- ▶ Limited set of repair functions.

Operating system service menu

The system diagnostic tests consist of IBM i service tools, stand-alone diagnostic tests that are loaded from the DVD drive, and online diagnostic tests (available in AIX).

When installed, online diagnostic tests are a part of the AIX or IBM i on the disk or server. They can be started in single-user mode (service mode), run in maintenance mode, or run concurrently (concurrent mode) with other applications. They can access the AIX error log and the AIX configuration data. IBM i has a service tools problem log, IBM i history log (QHST), and IBM i problem log.

The following modes are available:

- ▶ Service mode

This mode requires a service mode boot of the system and enables the checking of system devices and features. Service mode provides the most complete self-check of the system resources. All system resources (except the SCSI adapter and the disk drives that are used for paging) can be tested.

- ▶ Concurrent mode

This mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, certain devices might require more actions by the user or a diagnostic application before testing can be done.

- ▶ Maintenance mode

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

The System Management Services (SMS) error log is accessible from 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 Systems server, a DVD-ROM or DVD-RAM might be an option. An alternative method for maintaining and servicing the system must be available if you do not order the DVD-ROM or DVD-RAM.

With dedicated service tools (DST) or system service tools (SST), you can review various logs, run various diagnostic tests, or take several kinds of system memory dumps or other options.

Depending on the operating system, the following service-level functions are what you often see when you use the operating system service menus:

- ▶ Product activity log
- ▶ Trace Licensed Internal Code
- ▶ Work with communications trace
- ▶ Display/Alter/Dump
- ▶ Licensed Internal Code log
- ▶ Main storage memory dump manager
- ▶ Hardware service manager
- ▶ Call Home/Customer Notification
- ▶ Error information menu
- ▶ LED management menu
- ▶ Concurrent/Non-concurrent maintenance (within scope of the OS)
- ▶ Managing firmware levels:
 - Server
 - Adapter
- ▶ Remote support (access varies by OS)

Service Focal Point on the Hardware Management Console

Service strategies become more complicated in a partitioned environment. The Manage Serviceable Events task in the management console can help streamline this process.

Each LPAR reports errors that it detects and forwards the event to the SFP application that is running on the management console without determining whether other LPARs also detect and report the errors. For example, if one LPAR reports an error for a shared resource, such as a managed system power supply, other active LPARs might report the same error.

By using the Manage Serviceable Events task in the management console, you can avoid long lists of repetitive Call Home information by recognizing that these errors are repeated errors and consolidating them into one error.

In addition, you can use the Manage Serviceable Events task to start service functions on systems and LPARs, including the exchanging of parts, configuring connectivity, and managing memory dumps.

4.9.2 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 HMC V8R8.1.0.0, the firmware installations are restricted to entitled servers. The customer must be registered with IBM and have the appropriate service contract. During the initial machine warranty period, the access key is installed in the machine by IBM Manufacturing. The key is valid for the regular warranty period plus some extra 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. The following new SRC packages are available:

- ▶ E302FA06: Acquisition entitlement check failed
- ▶ E302FA08: Installation entitlement check failed

Any firmware release that was made available during the entitled time frame can still be installed. For example, if the entitlement period ends on 31 December 2014 and a new firmware release is available before the end of that entitlement period, it can still be installed. If that firmware is downloaded after 31 December 2014 but it was made available before the end of the entitlement period, it can still be installed. Any newer release requires a new update access key.

Note: The update access key expiration date requires a valid entitlement of the system to perform firmware updates.

You can find an update access key at [IBM Capacity on Demand \(CoD\) Home](#).

For more information about entitled IBM Software Support, see [My Entitle Systems Support](#).

Firmware updates

System firmware is delivered as a release level or a service pack. Release levels support the general availability (GA) of new functions or features, and new machine types or models. Upgrading to a higher release level is disruptive to customer operations. These release levels are supported by service packs. Service packs are intended to contain only firmware fixes and not introduce new functions. A *service pack* is an update to a release level.

The management console is used for system firmware updates. By using the management console, you can use the concurrent firmware maintenance (CFM) option when concurrent service packs are available. CFM is the Power Systems Firmware updates that can be partially or wholly concurrent or nondisruptive. With the introduction of CFM, you can address the following concerns:

- ▶ A release level is approaching its end of service date (that is, it was available for approximately one year and service soon will not be supported).
- ▶ You want to move a system to a more standardized release level when there are multiple systems in an environment with similar hardware.
- ▶ A new release has a new function that is needed in the environment.
- ▶ A scheduled maintenance action causes a platform restart, which also provides an opportunity to upgrade to a new firmware release.

Updating and upgrading system firmware depends on several factors, including the current firmware that is installed and what operating systems are running on the system. These scenarios and the associated installation instructions are described in the Firmware section of [Fix Central](#).

You also might want to review the preferred practice white papers that are found at [Service and support best practices for Power Systems](#).

Firmware update steps

The system firmware consists of service processor microcode, Open Firmware microcode, and system power control network (SPCN) microcode.

The firmware and microcode can be downloaded and installed from the HMC or a running partition.

Power Systems servers include a permanent firmware boot side (A side) and a temporary firmware boot side (B side). New levels of firmware must be installed first on the temporary side to test the update's compatibility with applications. When the new level of firmware is approved, it can be copied to the permanent side.

For access to the initial websites that address this capability, see [POWER9 systems 9040-MR9](#).

For POWER9 processor-based servers, select **POWER9 systems**. Then, 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 Power Systems server has the following levels of server firmware and power subsystem firmware:

- ▶ **Installed level**
This level of server firmware or power subsystem firmware is installed on the temporary side of the system firmware. It also is installed into memory after the managed system is powered off and then powered on.
- ▶ **Activated level**
This level of server firmware or power subsystem firmware is active and running in memory.
- ▶ **Accepted level**
This level is the backup level of server or power subsystem firmware. You can return to this level of server or power subsystem firmware if you decide to remove the installed level. It is installed on the permanent side of system firmware.

Use the HMC-enhanced GUI to obtain information about the different firmware levels in effect by selecting **Resources** → **All Systems**, selecting the system or the systems of interest, selecting **Actions** → **Updates** → **View system information** → and selecting **None - Display current values**.

Figure 4-5 shows the information that is collected by the HMC.

EC Number	LIC Type	Machine Type/Model/Serial Number	Update Access Key Expiration	Installed Level	Activated Level	Accepted Level	Unactivated Deferred Level	Platform IPL Level	Update Control
01VM920	Managed System Primary	9040-MR9*13601FX	05/23/2021	FW920.00 (015)	FW920.00 (015)	FW920.00 (010)		FW920.00 (015)	Management Console

Figure 4-5 HMC Enhanced GUI system firmware information

IBM provides the CFM function on Power E950 servers. This function supports applying nondisruptive system firmware service packs to the system concurrently (without requiring a restart operation to activate changes).

The concurrent levels of system firmware can (on occasion) contain fixes that are known as *deferred*. These deferred fixes can be installed concurrently, but are not activated until the next IPL. Any deferred fixes are identified in the Firmware Update Descriptions table of the firmware document. For deferred fixes within a service pack, only the fixes in the service pack that cannot be concurrently activated are deferred.

The file-naming convention for the system firmware is listed in Table 4-3.

Table 4-3 Firmware naming convention

PPNNSSS_FFF_DDD			
PP	Package identifier	01	-
NN	Platform and class	SV	Low end
SSS	Release indicator		
FFF	Current fix pack		
DDD	Last disruptive fix pack		

For example, here is the naming convention for the current (as of this writing) Power E950 firmware:

01VM920_040_040

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.

4.9.3 Concurrent firmware maintenance improvements

Since POWER6, firmware service packs are concurrently applied and take effect immediately. Occasionally, a service pack is shipped where most of the features can be concurrently applied. However, a patch in this area required a system restart for activation because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during operation.

With PORE, the firmware can now dynamically power off processor components, change the registers, and reinitialize while the system is running without discernible affect to any applications that are running on a processor. This feature potentially allows concurrent firmware changes in POWER9 processor-based systems, which in earlier designs required a restart to take effect.

Activating new firmware functions requires installation of a firmware release level. This process is disruptive to server operations and requires a scheduled outage and full server restart.

4.9.4 Electronic Services and Electronic Service Agent

IBM transformed its delivery of hardware and software support services to help you achieve higher system availability. Electronic Services is a web-enabled solution that offers an exclusive, no extra charge enhancement to the service and support that is available for IBM servers. These services provide the opportunity for greater system availability with faster problem resolution and preemptive monitoring.

The Electronic Services solution consists of the following separate (but complementary) elements:

- ▶ Electronic Services news page
- ▶ Electronic Service Agent

Electronic Services news page

The Electronic Services news page is a single internet entry point that replaces the multiple entry points that traditionally are used to access IBM internet services and support. With the news page, you can gain easier access to IBM resources for assistance in resolving technical problems.

Electronic Service Agent

The ESA is software that is on your server. It monitors events and transmits system inventory information to IBM on a periodic, client-defined timetable. The ESA automatically reports hardware problems to IBM.

Early knowledge about potential problems enables IBM to deliver proactive service that can result in higher system availability and performance. In addition, information that is collected through the Service Agent is made available to IBM SSRs when they help answer your questions or diagnose problems. The installation and use of ESA for problem reporting enables IBM to provide better support and service for your IBM server.

For more information about how Electronic Services can work for you, see [IBM Electronic Support](#) (an IBM ID is required).

Electronic Services features the following benefits:

- ▶ **Increased uptime**

The ESA tool enhances the warranty or maintenance agreement by providing faster hardware error reporting and uploading system information to IBM Support. This benefit can lead to less time that is wasted monitoring the symptoms, diagnosing the error, and manually calling IBM Support to open a problem record.

Its 24x7 monitoring and reporting mean no more dependence on human intervention or off-hours customer personnel when errors are encountered in the middle of the night.

- ▶ **Security**

The ESA tool is secure in monitoring, reporting, and storing the data at IBM. The ESA tool securely transmits through the internet (HTTPS or VPN) or modem. It can be configured to communicate securely through gateways to provide customers a single point of exit from their site.

Communication is one way. Activating ESA does not enable IBM to call into a customer's system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. It is viewable only by the customer and IBM. The customer's business applications or business data is *never* transmitted to IBM.

- ▶ **More accurate reporting**

Because system information and error logs are automatically uploaded to the IBM Support Center with the service request, customers are not required to find and send system information, which decreases the risk of misreported or misdiagnosed errors.

When inside IBM, problem error data is run through a data knowledge management system and knowledge articles are appended to the problem record.

- ▶ Customized support

By using the IBM ID that you enter during activation, you can view system and support information by selecting **My Systems** at [IBM Electronic Support](#).

My Systems provides valuable reports about 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, which provides advanced search of the technical support knowledge base. By using Premium Search and the ESA information that was collected from your system, your clients can see search results that apply specifically to their systems.

For more information about how to use the power of IBM Electronic Services, contact your IBM SSR, or see [IBM Electronic Support](#).

Service Event Manager

The Service Event Manager (SEM) allows the user to decide which of the Serviceable Events are called home with the ESA. Certain events can be locked. Some customers might not allow data to be transferred outside their company. After the SEM is enabled, the analysis of the possible problems might take longer.

Consider the following points:

- ▶ The SEM can be enabled by running the following command:

```
chhmc -c sem -s enable
```

- ▶ You can disable SEM mode and specify what state in which to leave the Call Home feature by running the following commands:

```
chhmc -c sem -s disable --callhome disable
```

```
chhmc -c sem -s disable --callhome enable
```

The basic configuration of the SEM can be accomplished by using the HMC Enhanced GUI. Select **Serviceability** → **Event Manager for Call Home** (Figure 4-6) to get access to the Events Manager for Call Home menu.

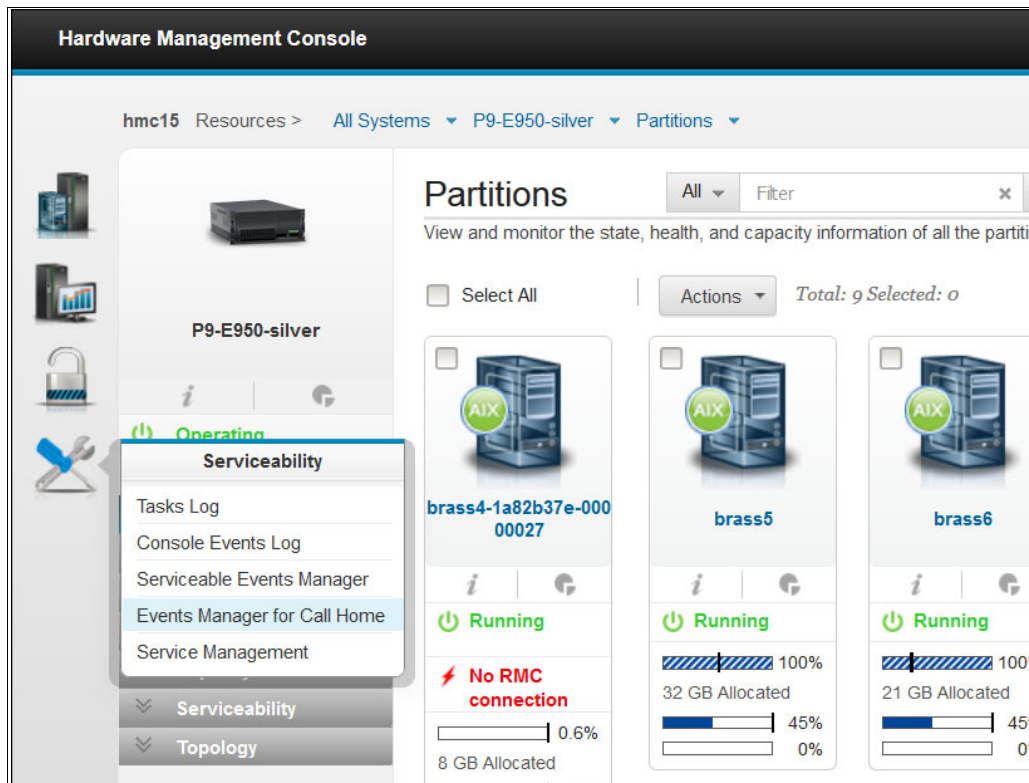


Figure 4-6 Service Event Manager configuration through the HMC Enhanced GUI

In the Events Manager for Call Home menu, you can add an HMC that is used to manage the serviceable events to the list of registered management consoles and proceed with further configuration steps, as shown in Figure 4-7.

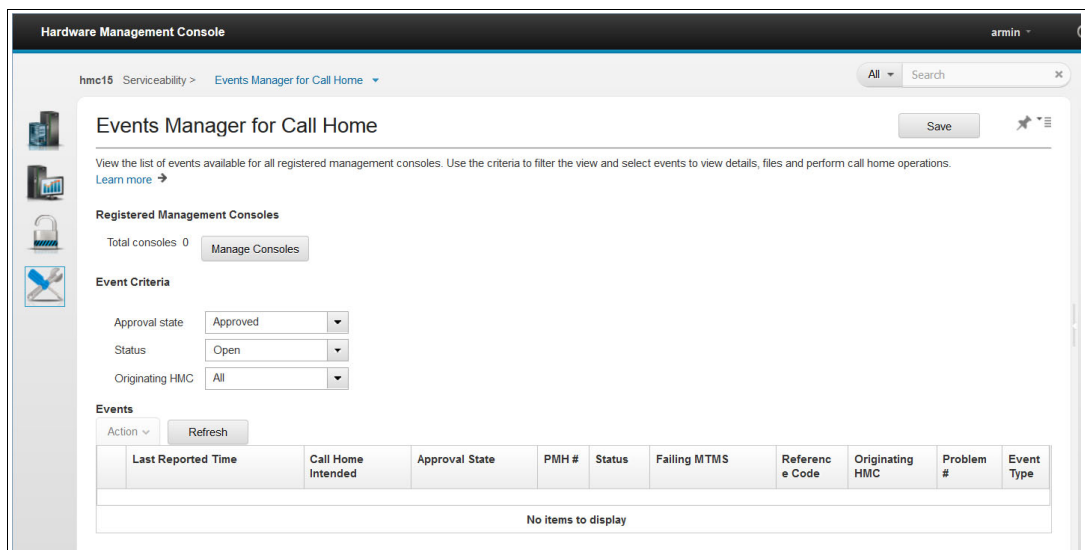


Figure 4-7 Event Manager for Call Home menu of the HMC Enhanced GUI

The following configurable options are available:

- ▶ Registered Management Consoles

“Total consoles” lists the number of consoles that are registered. Select **Manage Consoles** to manage the list of Registered Management Consoles.
- ▶ 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 the approval state to filter the list.
- ▶ Status

Select the value for the status to filter the list.
- ▶ Originating HMC

Select a single registered console or the **All consoles** option to filter the list.
- ▶ Serviceable Events

The Serviceable Events table shows the list of events that are based on the filters that are selected. To refresh the list, click **Refresh**.

The following menu options are available when you select an event in the table:

- ▶ View Details...

Shows the details of this event.
- ▶ View Files...

Shows the files that are associated with this event.
- ▶ Approve Call Home

Approves the Call Home of this event. This option is available only if the event is not yet approved.

The Help / Learn more function can be used to get more information about the other available windows for the Serviceable Event Manager.

4.10 Selected POWER9 RAS capabilities by operating system

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

Table 4-4 Selected RAS features by operating system

RAS feature	AIX	Linux
Processor		
FFDC for fault detection/error isolation	X	X
Dynamic Processor Deallocation	X	X
I/O subsystem		
PCIe bus enhanced error detection	X	X
PCIe bus enhanced error recovery	X	X

RAS feature	AIX	Linux
PCIe card hot-swap	X	X
Memory availability		
Memory Page Deallocation	X	X
SUE handling	X	X
Fault detection and isolation		
Storage Protection Keys	X	Not used by OS
ELA	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
EED collection	X	X
SP/OS Call Home on non-HMC configurations	X	X
IO adapter/device stand-alone diagnostic tests with PowerVM	X	X
SP mutual surveillance with IBM POWER Hypervisor	X	X
Dynamic firmware update with HMC	X	X
Service Agent Call Home Application	X	X
Service Indicator LED support	X	X
System memory dump for memory, POWER Hypervisor, and SP	X	X
IBM Knowledge Center / IBM Systems Support Site service publications	X	X
System Service/Support 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 [HMC])	X	X
Concurrent Op Panel Display Maintenance	X	X
Redundant HMCs	X	X
High availability clustering support	X	X
R&V Guided Maintenance with HMC	X	X

RAS feature	AIX	Linux
PowerVM Live Partition / Live Application Mobility With PowerVM Enterprise Edition	X	X
EPOW		
EPOW errors handling	X	X

Related publications

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

IBM Redbooks

The following IBM Redbooks publications provide more information about the topics in this document. Some publications that are referenced in this list might be available in softcopy only.

- ▶ *IBM PowerAI: Deep Learning Unleashed on IBM Power Systems Servers*, SG24-8409
- ▶ *IBM Power System AC922 Technical Overview and Introduction*, REDP-5494
- ▶ *IBM Power System E980 Technical Overview and Introduction*, REDP-5510
- ▶ *IBM Power System L922 Technical Overview and Introduction*, REDP-5496
- ▶ *IBM Power System S822LC for High Performance Computing Introduction and Technical Overview*, REDP-5405
- ▶ *IBM Power Systems LC921 and LC922 Introduction and Technical Overview*, REDP-5495
- ▶ *IBM Power Systems S922, S914, and S924 Technical Overview and Introduction*, REDP-5497
- ▶ *IBM PowerVM Best Practices*, SG24-8062
- ▶ *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

You can search for, view, download, or order these documents and other Redbooks publications, Redpapers, web docs, drafts, and additional materials, at the following website:

ibm.com/redbooks

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 Knowledge Center: IBM Power Systems Hardware
<http://www-01.ibm.com/support/knowledgecenter/api/redirect/powersys/v3r1m5/index.jsp>
- ▶ IBM Knowledge Center: Migration combinations of processor compatibility modes for active Partition Mobility
<http://www-01.ibm.com/support/knowledgecenter/api/redirect/powersys/v3r1m5/topic/p7hc3/iphc3pcmcombosact.htm>

- ▶ IBM Portal for OpenPOWER - POWER9 Monza Module
https://www.ibm.com/systems/power/openpower/tgcmDocumentRepository.xhtml?aliasId=POWER9_Monza
- ▶ IBM Power Systems
<http://www.ibm.com/systems/power/>
- ▶ IBM Storage website
<http://www.ibm.com/systems/storage/>
- ▶ IBM Systems Energy Estimator
<http://www-912.ibm.com/see/EnergyEstimator/>
- ▶ IBM System Planning Tool website
<http://www.ibm.com/systems/support/tools/systemplanningtool/>
- ▶ NVIDIA Tesla V100
<https://www.nvidia.com/en-us/data-center/tesla-v100/>
- ▶ NVIDIA Tesla V100 Performance Guide
<http://images.nvidia.com/content/pdf/volta-marketing-v100-performance-guide-us-r6-web.pdf>
- ▶ OpenCAPI
<http://opencapi.org/technical/use-cases/>
- ▶ OpenPOWER Foundation
<https://openpowerfoundation.org/>
- ▶ 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?brandid=Hardware~Systems~Power>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



REDP-5509-00

ISBN 0738457094

Printed in U.S.A.

Get connected

