

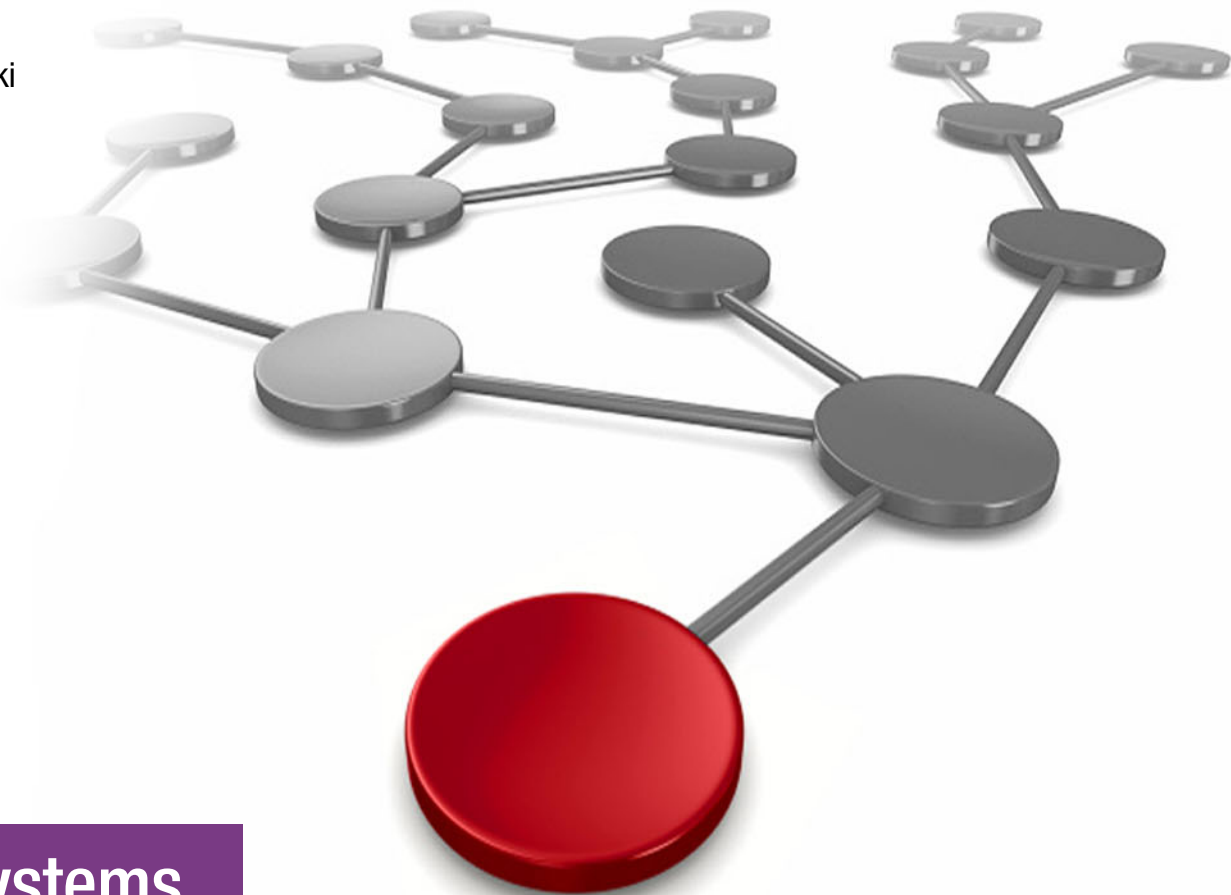
IBM Power System L922

Technical Overview and Introduction

Young Hoon Cho

Gareth Coates

Bartłomiej Grabowski



Power Systems



International Technical Support Organization

IBM Power System L922: Technical Overview and Introduction

July 2018

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

First Edition (July 2018)

This edition applies to the IBM Power System L922, machine type and model number 9008-22L.

© 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	x
Stay connected to IBM Redbooks	xi
Chapter 1. General description	1
1.1 Power L922 system overview	2
1.1.1 The operator panel	3
1.2 Operating environment	3
1.3 Physical package	5
1.4 Server features	5
1.4.1 Power L922 server features	5
1.4.2 Minimum features	7
1.4.3 Power supply features	7
1.5 Power L922 processor modules	7
1.5.1 Memory features	8
1.5.2 PCIe slots	8
1.6 Disk and media features	9
1.7 I/O drawers for the Power L922 server	12
1.7.1 PCIe3 I/O expansion drawer	13
1.7.2 I/O drawers and usable PCI slot	14
1.7.3 EXP24SX SAS Storage Enclosure (#ELLS) and EXP12SX SAS Storage Enclosure (#ELLL)	15
1.8 System racks	17
1.8.1 IBM 7014 Model T00 rack	17
1.8.2 IBM 7014 Model T42 rack	18
1.8.3 IBM 42U Slim Rack 7965-94Y	20
1.8.4 #0551	20
1.8.5 #0553	20
1.8.6 #ER05	20
1.8.7 The AC power distribution unit and rack content	21
1.8.8 Rack-mounting rules	23
1.8.9 Useful rack additions	23
1.8.10 Original equipment manufacturer rack	26
1.9 Hardware Management Console	27
1.9.1 New features	27
1.9.2 Hardware Management Console overview	28
1.9.3 Hardware Management Console code level	29
1.9.4 Two architectures of Hardware Management Console	29
1.9.5 Hardware Management Console connectivity to POWER9 processor-based systems' service processors	30
1.9.6 High availability Hardware Management Console configuration	31
Chapter 2. Architecture and technical overview	33
2.1 The IBM POWER9 processor	35

2.1.1	POWER9 processor overview	35
2.1.2	POWER9 processor features	36
2.1.3	POWER9 processor core	36
2.1.4	Simultaneous multithreading	37
2.1.5	Processor feature codes	38
2.1.6	Memory access	38
2.1.7	On-chip L3 cache innovation and Intelligent Cache	39
2.1.8	Hardware transactional memory	40
2.1.9	IBM Coherent Accelerator Processor Interface 2.0	40
2.1.10	Power management and system performance	42
2.1.11	Comparison of the POWER9, POWER8, and POWER7+ processors	42
2.2	Memory subsystem	43
2.2.1	Memory placement rules	43
2.2.2	Memory bandwidth	45
2.3	System bus	47
2.4	Internal I/O subsystem	48
2.4.1	Slot configuration	48
2.4.2	System ports	49
2.5	PCIe adapters	49
2.5.1	PCI Express	49
2.5.2	LAN adapters	50
2.5.3	Graphics accelerator adapters	51
2.5.4	SAS adapters	51
2.5.5	Fibre Channel adapters	52
2.5.6	Fibre Channel over Ethernet	53
2.5.7	InfiniBand host channel adapter	53
2.5.8	Cryptographic coprocessor	54
2.5.9	Coherent Accelerator Processor Interface adapters	54
2.6	Internal storage	55
2.6.1	Backplane (#EL66)	55
2.6.2	Split backplane option (#EL68)	55
2.6.3	PCIe3 NVMe express carrier card w/2 M.2 module slots (#EC59)	56
2.6.4	400 GB SSD Non-Volatile Memory express M.2 module (#EC14)	57
2.6.5	RAID support	58
2.6.6	Easy Tier	59
2.7	External IO subsystems	60
2.7.1	PCIe3 I/O expansion drawer	60
2.7.2	PCIe3 I/O expansion drawer optical cabling	61
2.7.3	PCIe3 I/O expansion drawer system power control network cabling	63
2.8	External disk subsystems	64
2.8.1	EXP24SX SAS Storage Enclosure and EXP12SX SAS Storage Enclosure	64
2.8.2	IBM Storage	66
2.9	Operating system support	67
2.9.1	Linux operating system	67
2.10	POWER9 reliability, availability, and serviceability capabilities by operating system	68
	Chapter 3. Virtualization	71
3.1	POWER Hypervisor	72
3.1.1	Virtual SCSI	74
3.1.2	Virtual Ethernet	74
3.1.3	Virtual Fibre Channel	74
3.1.4	Virtual (TTY) console	74
3.2	POWER processor modes	75

3.3 Single Root I/O Virtualization	77
3.4 PowerVM	78
3.4.1 Multiple shared processor pools	79
3.4.2 Virtual I/O Server	80
3.4.3 Live Partition Mobility	81
3.4.4 Active Memory Sharing	81
3.4.5 Active Memory Deduplication	82
3.4.6 Remote Restart	82
Related publications	83
IBM Redbooks	83
Online resources	83
Help from IBM	84

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

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

Active Memory™	Power Architecture®	PowerPC®
AIX®	POWER Hypervisor™	PowerVM®
Easy Tier®	Power Systems™	Real-time Compression™
Electronic Service Agent™	POWER6®	Redbooks®
EnergyScale™	POWER6+™	Redpaper™
IBM®	POWER7®	Redbooks (logo)  ®
IBM Spectrum™	POWER7+™	RS/6000®
IBM Spectrum Virtualize™	POWER8®	Storwize®
Micro-Partitioning®	POWER9™	System Storage®
POWER®	PowerHA®	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.

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

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

Preface

This IBM® Redpaper™ publication is a comprehensive guide covering the IBM Power System L922 (9008-22L) server, which was designed for data-intensive workloads such as databases and analytics in the Linux operating system. The objective of this paper is to introduce the major innovative Power L922 offering and its relevant functions:

- ▶ The new IBM POWER9™ processor, available at frequencies of 2.7 - 3.8 GHz, 2.9 - 3.8 GHz, and 3.4 - 3.9 GHz.
- ▶ Significantly strengthened cores and larger caches.
- ▶ Two integrated memory controllers that allow double the memory footprint of IBM POWER8® processor-based servers.
- ▶ An integrated I/O subsystem and hot-pluggable Peripheral Component Interconnect Express (PCIe) Gen4 and Gen3 I/O slots.
- ▶ I/O drawer expansion options offer greater flexibility.
- ▶ Support for Coherent Accelerator Processor Interface (CAPI) 2.0.
- ▶ New feature IBM EnergyScale™ technology provides 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 IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power L922 system.

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.

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

Gareth Coates has been working with IBM AIX® since 1988 and with Linux since 1991. He has trained IBM employees and customers on POWER4 through POWER9 processor-based systems, and AIX and Linux. He was appointed to the Worldwide Board of the IBM Learning Profession and spent time in the Server Services Group. In 2009, he joined the IBM EMEA Advanced Technology Services team, where he specializes in new product introduction of IBM POWER® processor-based hardware, IBM PowerVM®, and the Hardware Management Console (HMC). He presents at technical conferences and has co-authored numerous IBM Certification Tests.

Bartłomiej Grabowski is an IBM Champion and a Principal Systems Specialist in DHL IT Services. He has over 14 years of experience in enterprise solutions. He holds a bachelor degree in computer science from the Academy of Computer Science and Management in Bielsko-Biala. His areas of expertise include IBM Power Systems, IBM i, PowerHA®, PowerVM, and storage solutions. Bartłomiej is also a developer of IBM certification exams. He is a Gold Redbooks® author.

The project that produced this publication was managed by:

Scott Vetter, PMP
IBM Austin,

Thanks to the following individuals for their contribution and support of this publication:

Ron Arroyo, Matthew Butterbaugh, Nigel Griffiths, Daniel Henderson, Jeanine Hinck, Ray Laning, Chris Mann, Benjamin Mashak, Stephen Mroz, Thoi Nguyen, Kanisha Patel, William Starke, Jeff Stuecheli, Justin Thaler, Brain W Thompto, Julie Villarreal
IBM

Now you can become a published author, too!

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:

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

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:

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

- ▶ Follow us on Twitter:

<http://twitter.com/ibmredbooks>

- ▶ Look for us on LinkedIn:

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

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

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

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

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



General description

The next generation of Power Systems servers with POWER9 technology is built with innovations that can help deliver security and reliability for the data-intensive workloads of today's enterprises. POWER9 technology is designed from the ground up for data-intensive workloads, such as databases and analytics.

The Power Systems L922 server supports two processor sockets that offer 8-core or 16-core typical 3.4 - 3.9 GHz (maximum), 10-core or 20-core typical 2.9 - 3.8 GHz (maximum), or 24-core typical 2.7 - 3.8 GHz (maximum) POWER9 cores in a 19-inch rack-mount, 2U (EIA units) drawer configuration. All the cores are active.

The server supports a maximum of 32 DDR4 DIMM slots. The memory features that are supported are 8 GB, 16 GB, 32 GB, 64 GB, and 128 GB, and run at different speeds of 2133, 2400, and 2666 Mbps, offering a maximum system memory of 4096 GB.

1.1 Power L922 system overview

The Power L922 (9008-22L) server is a powerful one- or two-socket server that includes up to 24 activated cores. If only one socket is populated at the time of ordering, the second can be populated later. It has the I/O configuration flexibility to meet today's growth and tomorrow's processing needs. This server supports two processor sockets, offering 8-core, 10-core, or 12-core processors running 2.7 - 3.9 GHz in a 19-inch rack-mount, 2U (EIA units) drawer configuration. All the cores are active.

The Power L922 server supports a maximum of 32 DDR4 Registered DIMM (RDIMM) slots. If only one processor socket is populated, then only 16 RDIMMs can be used. The memory features that are supported are 16 GB, 32 GB, 64 GB, and 128 GB, allowing for a maximum system memory of 2 TB if one socket is populated and 4 TB with both sockets populated.

Two different features are available for the storage backplane:

- ▶ EL66: Eight SFF-3s with optional split card EL68
- ▶ EC59: Optional PCIe3 Non-Volatile Memory express (NVMe) carrier card with two M.2 module slots

Each of these backplane options uses leading-edge, integrated SAS RAID controller technology that is designed and patented by IBM.

The NVMe option offers fast start times and is ideally suited to the location of rootvg of Virtual I/O Server (VIOS) partitions.

Figure 1-1 shows the Power L922 server.



Figure 1-1 The Power L922 server

Note: The server has no internal DVD option, although an external USB DVD drive is available with feature code (FC) EUA5. Customers are encouraged to use USB flash drives to install operating systems and VIOS whenever possible as because they are much faster than DVDs.

1.1.1 The operator panel

The operator panel is formed of two parts. All of the servers have the first part, which provides the power switch and LEDs, and are shown in Figure 1-2.



Figure 1-2 Operator panel: Power switch and LEDs

The second part is an LCD panel with three buttons, and is shown in Figure 1-3.

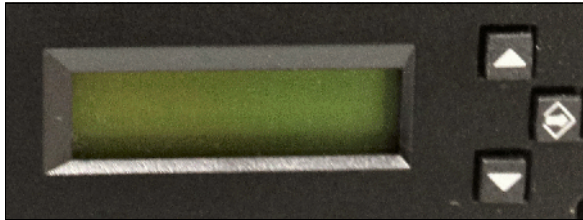


Figure 1-3 Operator panel: LCD and switches

In the Power L922 server, the LCD panel is optional, but if a rack contains any of the IBM POWER9 processor-based scale-out servers, one of them must have an LCD panel.

The LCD panel can be moved (by using the correct procedure) from one server to another to allow appropriate services to be carried out.

1.2 Operating environment

Table 1-1 lists the electrical characteristics for the Power L922 server.

Table 1-1 Electrical characteristics of the Power L922 server

Electrical characteristics	Properties
	Power L922 server
Operating voltage	1400 W power supply: 200 - 240 V AC
Operating frequency	47/63 Hz
Thermal output	6,416 Btu/hour (maximum)
Power consumption	1880 watts (maximum)
Power-source loading	1.94 kVa (maximum configuration)
Phase	Single

Note: The maximum measured value is the worst case power consumption that is expected from a fully populated server under an intensive workload. The maximum measured value also accounts for component tolerance and non-ideal operating conditions. Power consumption and heat load vary greatly by server configuration and utilization. The [IBM Systems Energy Estimator](#) should be used to obtain a heat output estimate based on a specific configuration.

Table 1-2 lists the environment requirements for the Power L922 server.

Table 1-2 Environment requirements for Power L922 server

Environment	Recommended operating	Allowable operating	Non-operating
Temperature	18 - 27°C (64.4 - 80.6°F)	5 - 40°C (41 - 104°F)	5 - 45°C (41 - 113°F)
Humidity range	5.5°C (42°F) dew point (DP) to 60% relative humidity (RH) or 15°C (59°F) dew point	8% - 85% RH	8% - 80% RH
Maximum dew point	N/A	24°C (75°F)	27°C (80°F)
Maximum operating altitude	N/A	3050 m (10000 ft)	N/A

Table 1-3 lists the noise emissions for the Power L922 server.

Table 1-3 Noise emissions for the Power L922 server

Product	Declared A-weighted sound power level, L_{WAd} (B)		Declared A-weighted sound pressure level, L_{pAm} (dB)	
	Operating	Idle	Operating	Idle
Power L922 server	7.8	6.9	61	53

Tip:

- ▶ Declared level L_{WAd} is the upper-limit A-weighted sound power level. Declared level L_{pAm} is the mean A-weighted emission sound pressure level that is measured at the 1-meter bystander positions.
- ▶ All measurements are made in conformance with ISO 7779 and declared in conformance with ISO 9296.
- ▶ 10 dB (decibel) equals 1 B (bel).

1.3 Physical package

Table 1-4 shows the physical dimensions of the Power L922 chassis. The server is available in a rack-mounted form factor and takes 2U (2 EIA units) of rack space.

Table 1-4 Physical dimensions of the rack-mounted Power L922 chassis

Dimension	Power L922 (9008-22L) server
Width	482 mm (18.97 in.)
Depth	766.5 mm (30.2 in.)
Height	86.7 mm (3.4 in.)
Weight	30.4 kg (67 lb)

Figure 1-4 show the front view of the Power L922 server.

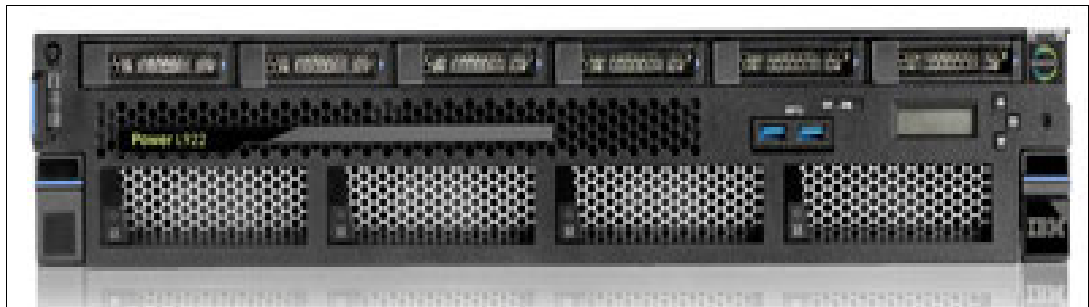


Figure 1-4 Front view of the Power L922 server

1.4 Server features

The Power L922 system chassis contains up to two processor modules. Each of the POWER9 processor chips in the server has a 64-bit architecture. All the cores are active.

1.4.1 Power L922 server features

This summary describes the standard features of the Power L922 server:

- ▶ POWER9 processor modules:
 - 8-core, typical 3.4 - 3.9 GHz (maximum) POWER9 processor.
 - 10-core, typical 2.9 - 3.8 GHz (maximum) POWER9 processor.
 - 12-core, typical 2.7 - 3.8 GHz (maximum) POWER9 processor.
- ▶ High-performance Mbps DDR4 error-correcting code (ECC) memory:
 - 8 GB, 16 GB, 32 GB, 64 GB, or 128 GB memory features with different sizes/configurations, which run at different frequencies of 2133, 2400, and 2666 Mbps.
 - Up to 4 TB of DDR4 memory with two POWER processors.
 - Up to 2 TB of DDR4 memory with one POWER processor.
- ▶ Storage feature: Eight small form factor (SFF) bays, one integrated SAS controller without cache, and JBOD RAID 0, 5, 6, or 10.

- ▶ Optionally, split the above SFF-3 bays and add a second integrated SAS controller without cache:
 - Expanded Function Storage Backplane 8 SFF-3 Bays/Single IOA with Write Cache.
 - Optionally, attach an EXP12SX/EXP24SX SAS HDD/solid-state drive (SSD) Expansion Drawer to the single IOA.
- ▶ Up to two PCIe3 NVMe carrier cards with two M.2 module slots (with up to four Mainstream 400 GB SSD NVMe M.2 modules). One PCIe3 NVMe carrier card can be ordered only with a storage backplane. If a PCIe3 NVMe carrier card is ordered with a storage backplane, then the optional split feature is not supported.
- ▶ Peripheral Component Interconnect Express (PCIe) slots with a single processor:
 - One x16 Gen4 low-profile, half-length (Coherent Accelerator Processor Interface (CAPI)).
 - One x8 Gen4 low-profile, half-length (with x16 connector) (CAPI).
 - Two x8 Gen3 low-profile, half-length (with x16 connectors).
 - Two x8 Gen3 low-profile, half-length. (One of these slots is used for the required base LAN adapter.)
- ▶ PCIe slots with two processors:
 - Three x16 Gen4 low-profile, half-length (CAPI).
 - Two x8 Gen4 low-profile, half-length (with x16 connectors) (CAPI).
 - Two x8 Gen3 low-profile, half-length (with x16 connectors).
 - Two x8 Gen3 low-profile, half-length. (One of these slots is used for the required base LAN adapter.)
- ▶ Integrated:
 - Service processor.
 - EnergyScale technology.
 - Hot-plug and redundant cooling.
 - Two front USB 3.0 ports.
 - Two rear USB 3.0 ports.
 - Two Hardware Management Console (HMC) 1 GbE RJ45 ports.
 - One system port with an RJ45 connector.
 - Two hot-plug, redundant power supplies.
 - 19-inch rack-mounting hardware (2U).

1.4.2 Minimum features

The minimum Power L922 initial order must include a processor module, two 8 GB DIMMs, two power supplies and power cords, an operating system indicator, a cover set indicator, and a Language Group Specify. Also, it must include one of the storage options and one of the network options below:

- ▶ Storage options:
 - For boot from NVMe: One NVMe carrier and one NVMe M.2 module.
 - For boot from local SFF-3 HDD/SDD: One storage backplane and one SFF-3 HDD or SDD.
 - For boot from SAN: Internal HDD or SSD and RAID cards are not required if #0837 (Boot from SAN) is selected. A Fibre Channel adapter must be ordered if #0837 is selected.
- ▶ Network options:
 - One PCIe2 4-port 1 Gb Ethernet adapter.
 - One of the supported 10 Gb Ethernet adapters.

1.4.3 Power supply features

The Power L922 server supports two 1400 watts 200 - 240-Volt power supplies (#EL1B). Two power supplies are always installed. One power supply is required for normal system operation; the second is for redundancy.

1.5 Power L922 processor modules

A maximum of two processors with eight processor cores (#ELPV), two processors with 10 processor cores (#ELPW), or two processors with 12 cores (#ELPX) is allowed. All processor cores must be activated. The following list defines the allowed quantities of processor activation entitlements:

- ▶ One 8-core, typical 3.4 - 3.9 GHz (maximum) processor (#ELPV) requires that eight processor activation codes be ordered. A maximum of eight processor activations (#ELAV) is allowed.
- ▶ Two 8-core, typical 3.4 - 3.9 GHz (maximum) processors (#ELPV) require that 16 processor activation codes be ordered. A maximum of 16 processor activations (#ELAV) is allowed.
- ▶ One 10-core, typical 2.9 - 3.8 GHz (maximum) processor (#ELPW) requires that 10 processor activation codes be ordered. A maximum of 10 processor activation code features (#ELAW) is allowed.
- ▶ Two 10-core, typical 2.9 - 3.8 GHz (maximum) processors (#ELPW) require that 20 processor activation codes be ordered. A maximum of 20 processor activation code features (#ELAW) is allowed.
- ▶ Two 12-core, typical 2.7 - 3.8 GHz (maximum) processors (#ELPX) require that 24 processor activation codes be ordered. A maximum of 24 processor activation code features (#ELAX) is allowed.

Table 1-5 summarizes the processor features that are available for the Power L922 server.

Table 1-5 Processor features for the Power L922 server

Feature code	Processor module description
ELPV	8-core Typical 3.4 - 3.9 GHz (maximum) POWER9 processor
ELPW	10-core Typical 2.9 - 3.8 GHz (maximum) POWER9 processor
ELPX	12-core Typical 2.7 - 3.8 GHz (maximum) POWER9 processor

1.5.1 Memory features

A minimum of 32 GB of memory is required on the Power L922 server. Memory upgrades require memory pairs. The base memory is two 8 GB DDR4 memory modules (#EM60).

Table 1-6 lists the memory features that are available for the Power L922 server.

Table 1-6 Summary of memory features for the Power L922 server

Feature code	DIMM capacity	Minimum quantity	Maximum quantity
EM60	8 GB	0	32
EM62	16 GB	0	32
EM63	32 GB	0	32
EM64	64 GB	0	32
EM65	128 GB	0	32

Note: Different sizes/configurations run at different frequencies of 2133, 2400, and 2666 Mbps.

1.5.2 PCIe slots

The Power L922 server has up to nine PCIe hot-plug slots, providing excellent configuration flexibility and expandability.

The following section describes the available PCIe slots of the Power L922 server:

- ▶ With two POWER9 processor single-chip modules (SCMs), nine PCIe slots are available: Three are x16 Gen4 low-profile, half-length slots (CAPI), two are x8 Gen4 low-profile, half-length slots (with x16 connectors) (CAPI), two are x8 Gen3 low-profile, half-length slots (with x16 connectors), and two are x8 Gen3 low-profile, half-length slots (one of these slots is used for the required base LAN adapter).
- ▶ With one POWER9 processor SCM, six PCIe slots are available: One is x16 Gen4 low-profile, half-length slots (CAPI), one is x8 Gen4 low-profile, half-length slots (with x16 connector) (CAPI), two are x8 Gen3 low-profile, half-length slots (with x16 connectors), and two are x8 Gen3 low-profile, half-length slots (one of these slots is used for the required base LAN adapter).

The x16 slots can provide up to twice the bandwidth of x8 slots because they offer twice as many PCIe lanes. PCIe Gen4 slots can support up to twice the bandwidth of a PCIe3 slot, and PCIe3 slots can support up to twice the bandwidth of a PCIe Gen2 slot, assuming an equivalent number of PCIe lanes.

At least one PCIe Ethernet adapter is required on the server by IBM to ensure proper manufacture, test, and support of the server. One of the x8 PCIe slots is used for this required adapter.

The Power L922 server is smarter about energy efficiency when cooling the PCIe adapter environment. They sense which IBM PCIe adapters are installed in their PCIe slots and, if an adapter requires higher levels of cooling, they automatically speed up fans to increase airflow across the PCIe adapters. Faster fans increase the sound level of the server.

1.6 Disk and media features

Three backplane options are available for the Power L922 servers:

- ▶ Storage Backplane 8 SFF-3 Bays (#EL66)
- ▶ 4 + 4 SFF-3 Bays split backplane (#EL68)
- ▶ Expanded Function Storage Backplane 8 SFF-3 Bays/Single IOA with Write Cache (#EL67)

#EL66 provides eight SFF-3 bays and one SAS controller with zero write cache.

By optionally adding #EL68, a second integrated SAS controller with no write cache is provided, and the eight SFF-3 bays are logically divided into two sets of four bays. Each SAS controller independently runs one of the four-bay sets of drives.

The backplane options provide SFF-3 SAS bays in the system unit. These 2.5-inch or SFF SAS bays can contain SAS drives (HDD or SSD) mounted on a Gen3 tray or carrier. Thus, the drives are designated SFF-3. SFF-1 or SFF-2 drives do not fit in an SFF-3 bay. All SFF-3 bays support concurrent maintenance or hot-plug capability.

These backplane options use leading-edge, integrated SAS RAID controller technology that is designed and patented by IBM. A custom-designed IBM PowerPC® based ASIC chip is the basis of these SAS RAID controllers and provides RAID 5 and RAID 6 performance levels, especially for SSD. Internally, SAS ports are implemented and provide plenty of bandwidth. The integrated SAS controllers are placed in dedicated slots and do not reduce the number of available PCIe slots.

This backplane option supports HDDs or SSDs or a mixture of HDDs and SSDs in the SFF-3 bays. Mixing HDDs and SSDs applies even within a single set of six bays of the split backplane option.

Note: If mixing HDDs and SSDs, they must be in separate arrays (unless you use the IBM Easy Tier® function).

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

This backplane option is supported by Linux and VIOS. It is highly recommended but not required that the drives be protected.

Unlike the hot-plug PCIe slots and SAS bays, concurrent maintenance is not available for the integrated SAS controllers. Scheduled downtime is required if a service action is required for these integrated resources.

In addition to supporting HDDs and SSDs in the SFF-3 SAS bays, the Expanded Function Storage Backplane (#EJ1G) supports the optional attachment of an EXP12SX/EXP24SX drawer. All bays are accessed by both of the integrated SAS controllers. The bays support concurrent maintenance (hot plug).

Table 1-7 shows the available disk drive FCs that can be installed in the Power L922 server.

Table 1-7 Disk drive feature code description for the Power L922 server

Feature code	CCIN	Description	Maximum	OS support
ELD3	59CD	1.2 TB 10K RPM SAS SFF-2 Disk Drive (Linux)	672	Linux
ELF3	59DA	1.2 TB 10K RPM SAS SFF-2 Disk Drive 4 KB Block	672	Linux
ELF9	59DB	1.2 TB 10K RPM SAS SFF-3 Disk Drive 4 KB Block	8	Linux
ELGP	5B12	1.55 TB Enterprise SAS 4 KB SFF-2 SSD for Linux	336	Linux
ELGR	5B15	1.55 TB Enterprise SAS 4 KB SFF-3 SSD for Linux	8	Linux
EL8F	5B12	1.55 TB SFF-2 SSD 4 KB eMLC4 for Linux	336	Linux
EL8V	5B15	1.55 TB SFF-3 SSD 4 KB eMLC4 for Linux	8	Linux
ELFT	59DD	1.8 TB 10K RPM SAS SFF-2 Disk Drive 4 KB Block	672	Linux
ELFV	59DE	1.8 TB 10K RPM SAS SFF-3 Disk Drive 4 KB Block	8	Linux
EL96	5B21	1.86 TB Mainstream SAS 4 KB SFF-2 SSD for Linux	336	Linux
EL92	5B20	1.86 TB Mainstream SAS 4 KB SFF-3 SSD for Linux	8	Linux
ELHL	5B21	1.86 TB Mainstream SAS 4k SFF-2 SSD for Linux	336	Linux
ELHU	5B20	1.86 TB Mainstream SAS 4k SFF-3 SSD for Linux	8	Linux
EL80	5B21	1.9 TB Read Intensive SAS 4 KB SFF-2 SSD for Linux	336	Linux
EL8J	5B20	1.9 TB Read Intensive SAS 4 KB SFF-3 SSD for Linux	8	Linux
ELE7	5B2D	3.72 TB Mainstream SAS 4 KB SFF-2 SSD for Linux	336	Linux
ELE1	5B2C	3.72 TB Mainstream SAS 4 KB SFF-3 SSD for Linux	8	Linux
ELM8	5B2D	3.72 TB Mainstream SAS 4k SFF-2 SSD for Linux	336	Linux
ELMQ	5B2C	3.72 TB Mainstream SAS 4k SFF-3 SSD for Linux	8	Linux
EL62	5B1D	3.82 - 4.0 TB 7200 RPM 4 KB SAS LFF-1 Nearline Disk Drive	336	Linux
ELHN	5B2F	7.45 TB Mainstream SAS 4k SFF-2 SSD for Linux	336	Linux
ELHW	5B2E	7.45 TB Mainstream SAS 4k SFF-3 SSD for Linux	0	Linux
EL64	5B1F	7.72 - 8.0 TB 7200 RPM 4 KB SAS LFF-1 Nearline Disk Drive	336	Linux
ELEZ	59C9	300 GB 15K RPM SAS SFF-2 4 KB Block	672	Linux
ESRM	5B43	300 GB 15K RPM SAS SFF-2 4 KB Block Cached Disk Drive	672	Linux
EL1P	19B1	300 GB 15K RPM SAS SFF-2 Disk Drive	672	Linux
ELFB	59E1	300 GB 15K RPM SAS SFF-3 4 KB Block	8	Linux
ESRL	5B41	300 GB 15K RPM SAS SFF-3 4 KB Block Cached Disk Drive	8	Linux

Feature code	CCIN	Description	Maximum	OS support
ELDB	59E0	300 GB 15K RPM SAS SFF-3 Disk Drive	8	Linux
ELGB	5B10	387 GB Enterprise SAS 4 KB SFF-2 SSD for Linux	336	Linux
ELGD	5B13	387 GB Enterprise SAS 4 KB SFF-3 SSD for Linux	8	Linux
ELG5	5B16	387 GB Enterprise SAS 5xx SFF-2 SSD for Linux	336	Linux
ELG9	5B19	387 GB Enterprise SAS 5xx SFF-3 SSD for Linux	8	Linux
EL85	5B10	387 GB SFF-2 SSD 4 KB eMLC4 for Linux	336	Linux
EL78	5B16	387 GB SFF-2 SSD 5xx eMLC4 for Linux	336	Linux
EL8N	5B13	387 GB SFF-3 SSD 4 KB eMLC4 for Linux	8	Linux
EL7K	N/A	387 GB SFF-3 SSD 5xx eMLC4 for Linux	8	Linux
ES14	N/A	400 GB Mainstream SSD NVMe M.2 module	2	Linux
EL1Q	19B3	600 GB 10K RPM SAS SFF-2 Disk Drive	672	Linux
ELEV	59D2	600 GB 10K RPM SAS SFF-2 Disk Drive 4 KB Block	672	Linux
ELD5	59D0	600 GB 10K RPM SAS SFF-3 Disk Drive	8	Linux
ELF5	59D3	600 GB 10K RPM SAS SFF-3 Disk Drive 4 KB Block	8	Linux
ELFP	59CC	600 GB 15K RPM SAS SFF-2 4 KB Block	672	Linux
ESRR	5B47	600 GB 15K RPM SAS SFF-2 4 KB Block Cached Disk Drive	672	Linux
ELDP	59CF	600 GB 15K RPM SAS SFF-2 Disk Drive - 5xx Block	672	Linux
ELFF	59E5	600 GB 15K RPM SAS SFF-3 4 KB Block	8	Linux
ESRP	5B45	600 GB 15K RPM SAS SFF-3 4 KB Block Cached Disk Drive	8	Linux
ELGK	5B11	775 GB Enterprise SAS 4 KB SFF-2 SSD for Linux	336	Linux
ELGM	5B14	775 GB Enterprise SAS 4 KB SFF-3 SSD for Linux	8	Linux
ELGF	5B17	775 GB Enterprise SAS 5xx SFF-2 SSD for Linux	336	Linux
ELGH	5B1A	775 GB Enterprise SAS 5xx SFF-3 SSD for Linux	8	Linux
EL8C	5B11	775 GB SFF-2 SSD 4 KB eMLC4 for Linux	336	Linux
EL7E	5B17	775 GB SFF-2 SSD 5xx eMLC4 for Linux	336	Linux
EL8Q	5B14	775 GB SFF-3 SSD 4 KB eMLC4 for Linux	8	Linux
EL7P	5B1A	775 GB SFF-3 SSD 5xx eMLC4 for Linux	8	Linux
EL8Y	5B29	931 GB Mainstream SAS 4 KB SFF-2 SSD for Linux	336	Linux
EL83	5B2B	931 GB Mainstream SAS 4 KB SFF-3 SSD for Linux	8	Linux
ELHJ	5B29	931 GB Mainstream SAS 4k SFF-2 SSD for Linux	336	Linux
ELHS	5B2B	931 GB Mainstream SAS 4k SFF-3 SSD for Linux	8	Linux
ELQP	19B1	Quantity 150 of EL1P	4	Linux
ELQQ	19B3	Quantity 150 of EL1Q	4	Linux

Feature code	CCIN	Description	Maximum	OS support
ELRQ	59E8	Quantity 150 of EL4Q 387 GB SFF-2 4k SSD (Linux)	2	Linux
ELQS	59C3	Quantity 150 of EL4S 775 GB SFF-2 4k SSD (Linux)	2	Linux
ELR2	5B1D	Quantity 150 of EL62 3.86 - 4.0 TB 7200 RPM 4 KB LFF-1 Disk	2	Linux
ELR4	5B1F	Quantity 150 of EL64 7.72 - 8.0 TB 7200 RPM 4 KB LFF-1 Disk	2	Linux
ELQ8	5B16	Quantity 150 of EL78	2	Linux
ELQE	5B17	Quantity 150 of EL7E	2	Linux
ELR0	5B21	Quantity 150 of EL80	2	Linux
ELQ5	5B10	Quantity 150 of EL85	2	Linux
ELQC	5B11	Quantity 150 of EL8C	2	Linux
ELQF	5B12	Quantity 150 of EL8F	2	Linux
ELQY	5B29	Quantity 150 of EL8Y (931 GB SFF-2)	2	Linux
ELQ6	5B21	Quantity 150 of EL96	2	Linux
ELQ3	59CD	Quantity 150 of ELD3 (1.2 TB 10k SFF-2)	4	Linux
ELQ0	59CF	Quantity 150 of ELDP 600 GB 15k RPM SFF-2 Disk	4	Linux
ELQ7	5B2D	Quantity 150 of ELE7	2	Linux
ELQV	59D2	Quantity 150 of ELEV	4	Linux
ELQZ	59C9	Quantity 150 of ELEZ	4	Linux
ELQ2	59DA	Quantity 150 of ELF3	4	Linux
EDQ1	59CC	Quantity 150 of ELFP	4	Linux
ELQT	59DD	Quantity 150 of ELFT	4	Linux
ELR5	5B16	Quantity 150 of ELG5	2	Linux
ELRB	5B10	Quantity 150 of ELGB	2	Linux
ELRF	5B17	Quantity 150 of ELGF	2	Linux
ELRK	5B11	Quantity 150 of ELGK	2	Linux
ELRP	5B12	Quantity 150 of ELGP (1.55 TB SAS 4 KB)	2	Linux
ESVM	5B43	Quantity 150 of ESRM	4	Linux
ESVR	5B47	Quantity 150 of ESRR	4	Linux

The RDX docking station EUA4 accommodates RDX removable disk cartridges of any capacity. The disk is in a protective rugged cartridge enclosure that plugs into the docking station. The docking station holds one removable rugged disk drive/cartridge at a time. The rugged removable disk cartridge and docking station perform saves, restores, and backups similar to a tape drive. This docking station can be an excellent entry capacity/performance option.

The stand-alone USB DVD drive (EUA5) is an optional, stand-alone external USB-DVD device. It requires high current at 5 V and must use the front USB 3.0 port.

1.7 I/O drawers for the Power L922 server

If more Gen3 PCIe slots beyond the system node slots are required, PCIe3 I/O drawers can be attached to the Power L922 server.

EXP24SX /EXP12SX SAS Storage Enclosures (ELLS or ELLL) are also supported and provide storage capacity.

The 7226-1U3 model offers a 1U rack-mountable dual bay enclosure with storage device options of LTO5, 6, 7, and 8 tape drives with both SAS and Fibre Channel interfaces. The 7226 model also offers DVD-RAM SAS and USB drive features, and RDX 500 GB, 1 TB, and 2 TB drive options. Up to two drives (or four DVD-RAM drives) can be installed in any combination in the 7226 enclosure.

1.7.1 PCIe3 I/O expansion drawer

This 19-inch, 4U (4 EIA) enclosure provides PCIe3 slots outside of the system unit. It has two module bays. One 6-slot fan-out module (ELMF or ELMG) can be placed in each module bay. Two 6-slot modules provide a total of 12 PCIe3 slots. Each fan-out module is connected to a PCIe3 Optical Cable adapter in the system unit over an active optical CXP cable (AOC) pair or CXP copper cable pair.

The PCIe3 I/O Expansion Drawer has two redundant, hot-plug power supplies. Each power supply has its own separately ordered power cord. The two power cords plug into a power supply conduit that connects to the power supply. The single-phase AC power supply is rated at 1030 watts and can use 100 - 120 V or 200 - 240 V. If using 100 - 120 V, then the maximum is 950 watts. As a preferred practice, connect the power supply to a power distribution unit (PDU) in the rack. Power Systems PDUs are designed for a 200 - 240 V electrical source.

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

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

Figure 1-5 shows a PCIe3 I/O expansion drawer.



Figure 1-5 PCIe3 I/O expansion drawer

1.7.2 I/O drawers and usable PCI slot

Figure 1-6 shows the rear view of the PCIe3 I/O expansion drawer that is equipped with two PCIe3 6-slot fan-out modules with the location codes for the PCIe adapter slots.

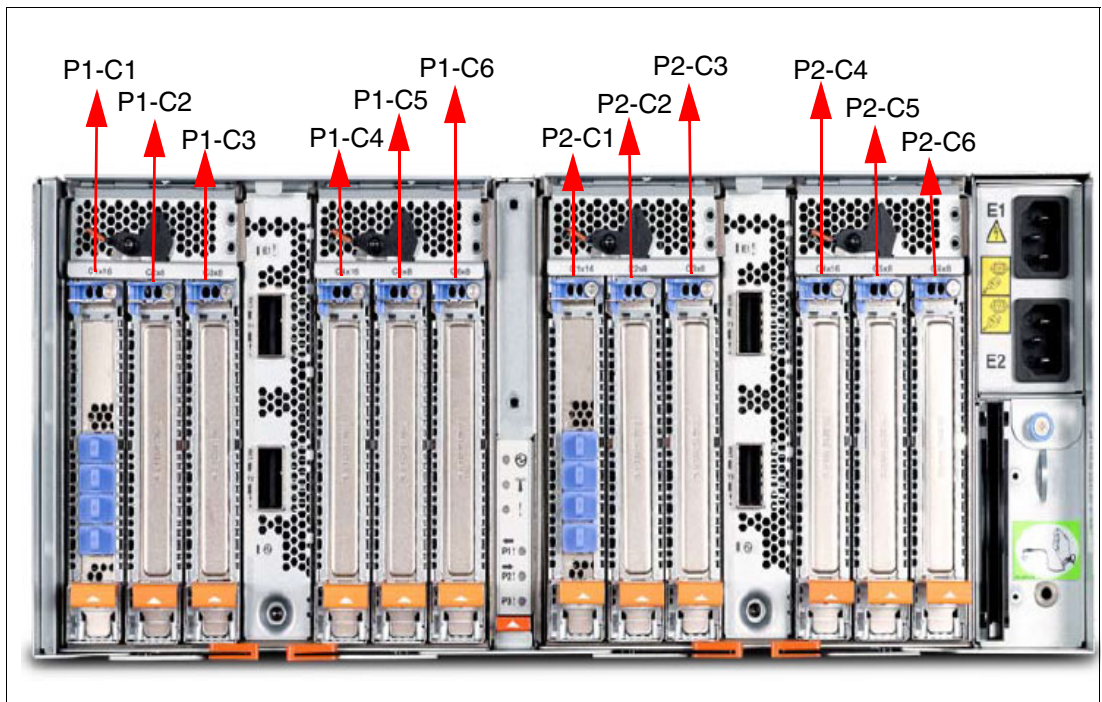


Figure 1-6 Rear view of a PCIe3 I/O expansion drawer with PCIe slots location codes

Table 1-8 provides details about the PCI slots in the PCIe3 I/O expansion drawer that is equipped with two PCIe3 6-slot fan-out modules.

Table 1-8 PCIe slot locations for the PCIe3 I/O expansion drawer with two fan-out modules

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 1-8 on page 15:

- ▶ All slots support full-length, regular-height adapters or short (low-profile) 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.

Table 1-9 summarizes the maximum number of I/O drawers that are supported and the total number of PCI slots that are available.

Table 1-9 Maximum number of I/O drawers that are supported and total number of PCI slots

System	Maximum number of I/O Exp Drawers	Maximum number of I/O fan-out modules	Maximum PCIe slots
Power L922 server (1-socket)	1	1	11
Power L922 server (2-socket)	1	2	19

1.7.3 EXP24SX SAS Storage Enclosure (#ELLS) and EXP12SX SAS Storage Enclosure (#ELLL)

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

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 Linux/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 bays (mode 1). It is possible to change the mode setting in the field by using software commands along with a documented procedure.

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.

Four mini-SAS HD ports on the EXP24SX or EXP12SX are attached to PCIe3 SAS adapters or attached to an integrated SAS controller in the Power L922 server.

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

Figure 1-7 shows the EXP24SX drawer.



Figure 1-7 The EXP24SX drawer

Figure 1-8 shows the EXP12SX drawer.



Figure 1-8 The EXP12SX drawer

1.8 System racks

The Power L922 server is designed to mount in the 36U 7014-T00 (#0551), the 42U 7014-T42 (#0553), or the IBM 42U Slim Rack (7965-94Y) rack. These racks are built to the 19-inch EIA 310D standard.

Order information: The racking approach for the initial order must be either a 7014-T00, 7014-T42, or 7965-94Y model. If an extra rack is required for I/O expansion drawers as a manufacturing execution system (MES) to an existing system, either an #0551, #0553, or #ER05 rack must be ordered.

You must leave 2U of space at either the bottom or top of the rack, depending on the client's cabling preferences, to allow for cabling to exit the rack.

If a system will be installed in a rack or cabinet that is not an IBM rack, ensure that the rack meets the requirements that are described in 1.8.10, "Original equipment manufacturer rack" on page 27.

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.

1.8.1 IBM 7014 Model T00 rack

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

- ▶ Has 36U (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 and DC configurations.
- ▶ Up to four PDUs can be mounted in the PDU bays, but other PDUs can fit inside the rack. For more information, see 1.8.7, "The AC power distribution unit and rack content" on page 22.
- ▶ For the T00 rack, three door options are available:
 - Front Door for 1.8 m Rack (#6068).

This feature 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.
 - A 1.8 m Rack Acoustic Door (#6248).

This feature provides a front and rear rack door that is designed to reduce acoustic sound levels in a general business environment.
 - A 1.8 m Rack Trim Kit (#6263).

If no front door is used in the rack, this feature provides a decorative trim kit for the front.

- ▶ Ruggedized Rack Feature.

For enhanced rigidity and stability of the rack, the optional Ruggedized Rack Feature (#6080) provides extra 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:

- T00 base empty rack: 244 kg (535 lb).
- T00 full rack: 816 kg (1795 lb).
- Maximum weight of drawers is 572 kg (1260 lb).
- Maximum weight of drawers in a zone 4 earthquake environment is 490 kg (1080 lb). This number equates to 13.6 kg (30 lb) per EIA.

Important: If more weight is added to the top of the rack, for example, adding #6117, the 490 kg (1080 lb) must be reduced by the weight of the addition. As an example, #6117 weighs approximately 45 kg (100 lb), so the new maximum weight of drawers that the rack can support in a zone 4 earthquake environment is 445 kg (980 lb). In the zone 4 earthquake environment, the rack must be configured starting with the heavier drawers at the bottom of the rack.

1.8.2 IBM 7014 Model T42 rack

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

- ▶ The T42 rack has 42U (EIA units) of usable space (6U of extra space).
- ▶ The model T42 supports AC power only.
- ▶ Weights are as follows:
 - T42 base empty rack: 261 kg (575 lb)
 - T42 full rack: 930 kg (2045 lb)

The available door options for the Model T42 rack are shown in Figure 1-9.

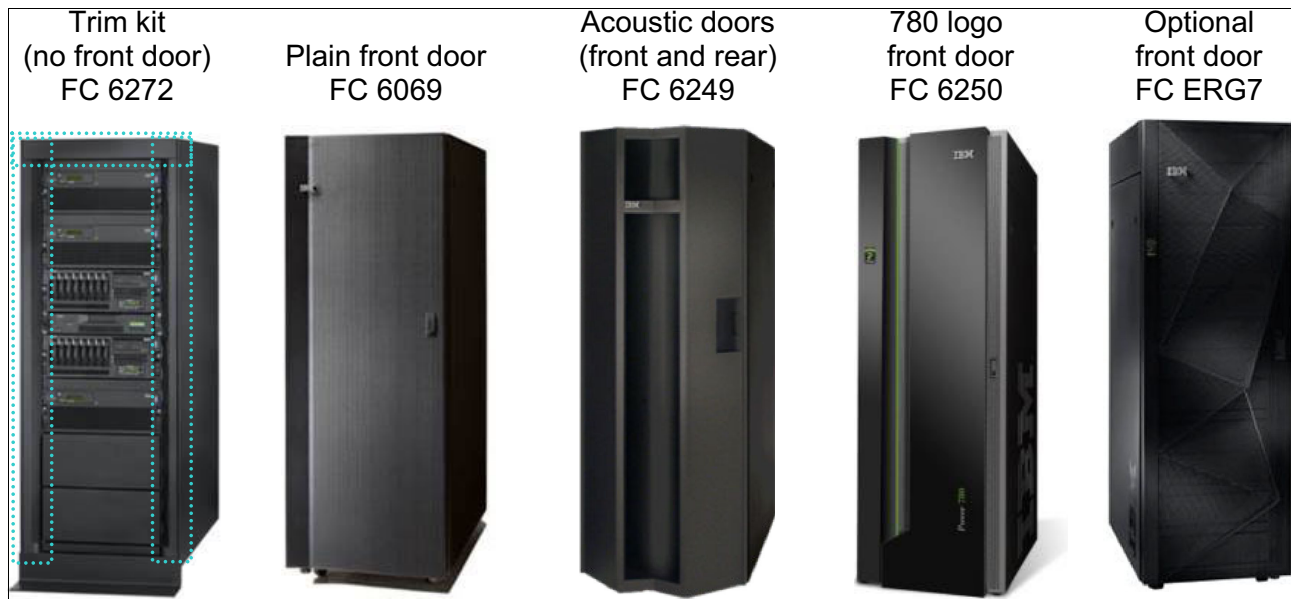


Figure 1-9 Door options for the T42 rack

In Figure 1-9:

- ▶ 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 and rear door to reduce noise by approximately 6 dB(A). It has a depth of approximately 191 mm (7.5 in.).
- ▶ The High-End Appearance Front Door (#6250) provides a front rack door with a field-installed Power 780 logo indicating that the rack contains a Power 780 system. The door is not acoustic and has a depth of about 90 mm (3.5 in.).

High end: For the High-End Appearance Front Door (#6250), use the High-End Appearance Side Covers (#6238) to make the rack appear as though it is a high-end server (but in a 19-inch rack format instead of a 24-inch rack).

- ▶ 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.).

Rear Door Heat Exchanger

To lead away more heat, a special door that is named the Rear Door Heat Exchanger (RDHX) (#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, which 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 RDHX, see [IBM Knowledge Center](#).

1.8.3 IBM 42U Slim Rack 7965-94Y

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

The IBM 42U Slim Rack has a lockable perforated front steel door, providing ventilation, physical security, and visibility of indicator lights in the installed equipment within. In the rear, either a lockable perforated rear steel door (#EC02) or a lockable RDHX(1164-95X) is used. Lockable optional side panels (#EC03) increase the rack's aesthetics, help control airflow through the rack, and provide physical security. Multiple 42U Slim Racks can be bolted together to create a rack suite (#EC04).

Up to six optional 1U PDUs can be placed vertically in the sides of the rack. More PDUs can be placed horizontally, but they each use 1U of space in this position.

1.8.4 #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 rack must be ordered separately with #0551.

1.8.5 #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 rack must be ordered separately with #0553.

1.8.6 #ER05

This feature 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 the #ER05:

- ▶ Front door (#EC01)
- ▶ Rear door (#EC02) or 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 that IBM Manufacturing can include the hardware in the rack.

1.8.7 The AC power distribution unit and rack content

For rack models T00 and T42, 12-outlet PDUs are available, which include the AC PDUs (#9188 and #7188), and the AC Intelligent PDU+ (#5889 and #7109). The Intelligent PDU+ (#5889 and #7109) is identical to #9188 and #7188 PDU, but are equipped with one Ethernet port, one console serial port, and one RS232 serial port for power monitoring.

The PDUs have 12 client-usable IEC 320-C13 outlets. There are six groups of two outlets that are fed by six circuit breakers. Each outlet is rated up to 10 amps, but each group of two outlets is fed from one 15 amp circuit breaker.

Four PDUs can be mounted vertically in the back of the T00 and T42 racks. Figure 1-10 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 T42 rack. The four vertical mounting locations are filled first in the T00 and T42 racks. Mounting PDUs horizontally uses 1U per PDU and reduces the space available for other racked components. When mounting PDUs horizontally, the best 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.

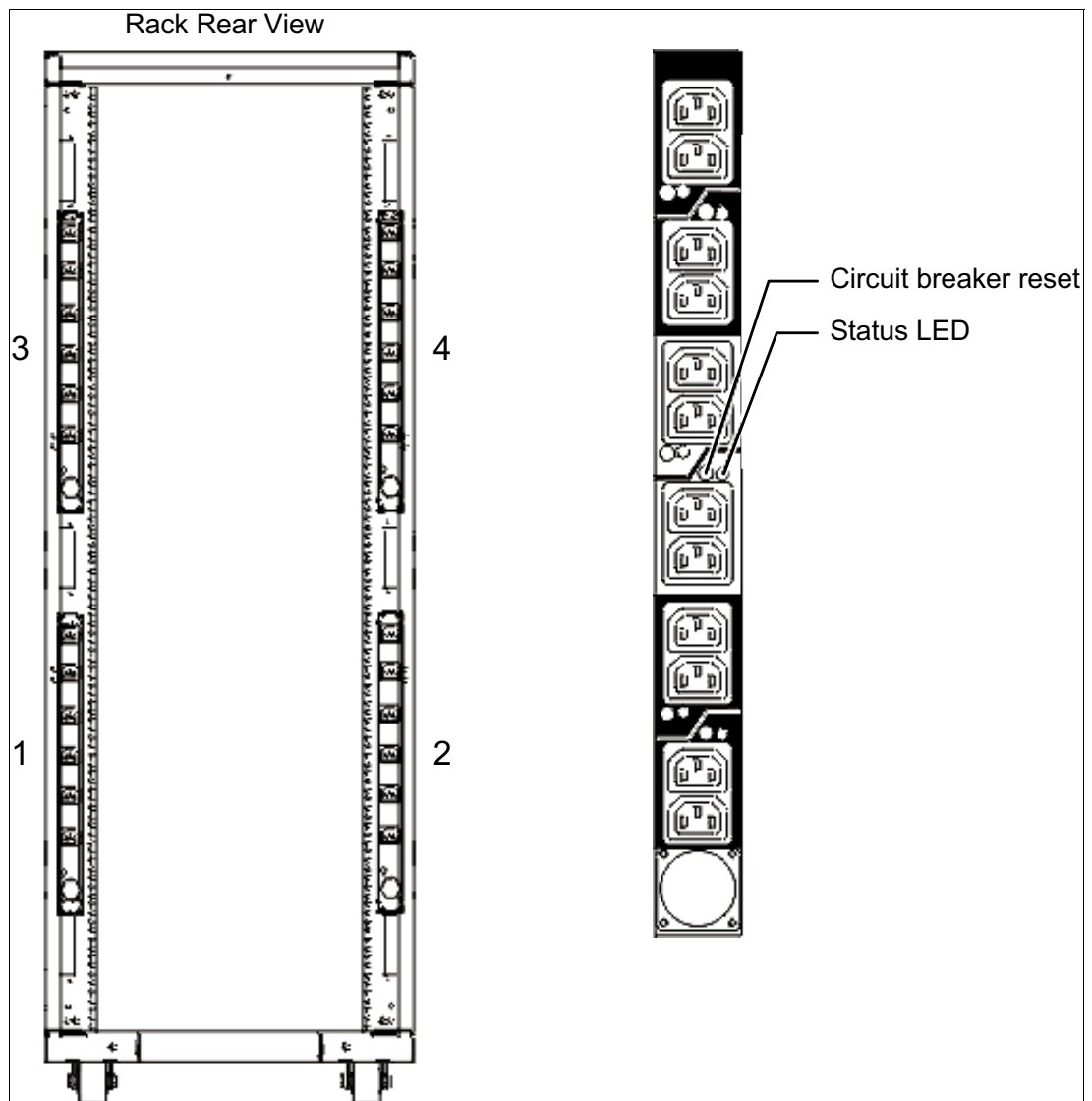


Figure 1-10 PDU placement and PDU view

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 1-10 shows the available wall power cord options for the PDU and intelligent power distribution unit (iPDU) features, which must be ordered separately.

Table 1-10 Wall power cord options for the PDU and iPDU features

Feature code	Wall plug	Rated voltage (Vac)	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 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 two 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 the “Planning for power” section in [IBM Knowledge Center](#).

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

1.8.8 Rack-mounting rules

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

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

Order information: The racking approach for the initial order must be either a 7014-T00, 7014-T42, or 7965-94Y. If an extra rack is required for I/O expansion drawers as an MES to an existing system, either #0551, #0553, or #ER05 must be ordered.

You must leave 2U of space at either the bottom or top of the rack, depending on the client's cabling preferences, to allow for cabling to exit the rack.

1.8.9 Useful rack additions

This section highlights several rack addition solutions for IBM 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 1-11.

Table 1-11 Supported drive features for the 7226-1U3 enclosure

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
#5763	DVD Front USB Port Sled with DVD-RAM USB Drive	Available
#5757	DVD RAM Slim USB Optical Drive	Available

Feature code	Description	Status
#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

Option descriptions are as follows:

- ▶ 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 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 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, POWER6+™, POWER7, 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 1-11 shows the IBM System Storage 7226 Multi-Media Enclosure.



Figure 1-11 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 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 [IBM 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 by using an 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.

1.8.10 Original equipment manufacturer rack

The system can be installed in a suitable original equipment manufacturer (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 (the horizontal width between the vertical columns of the holes on the two front-mounting flanges and on the two rear-mounting flanges). Figure 1-12 is a top view showing the specification dimensions.

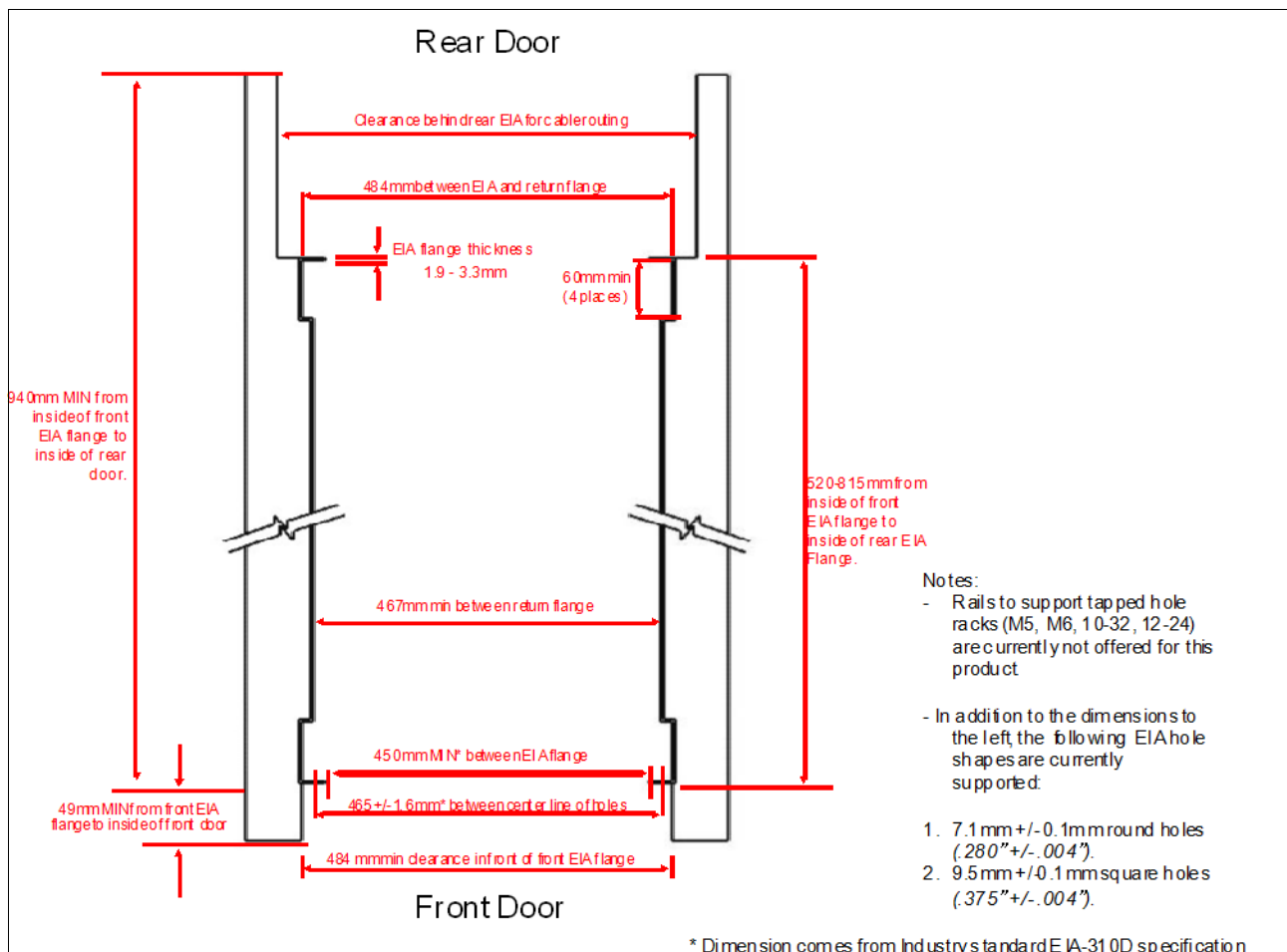


Figure 1-12 Top view of rack specification dimensions (not specific to IBM)

- ▶ 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 1-13 shows the top front specification dimensions.

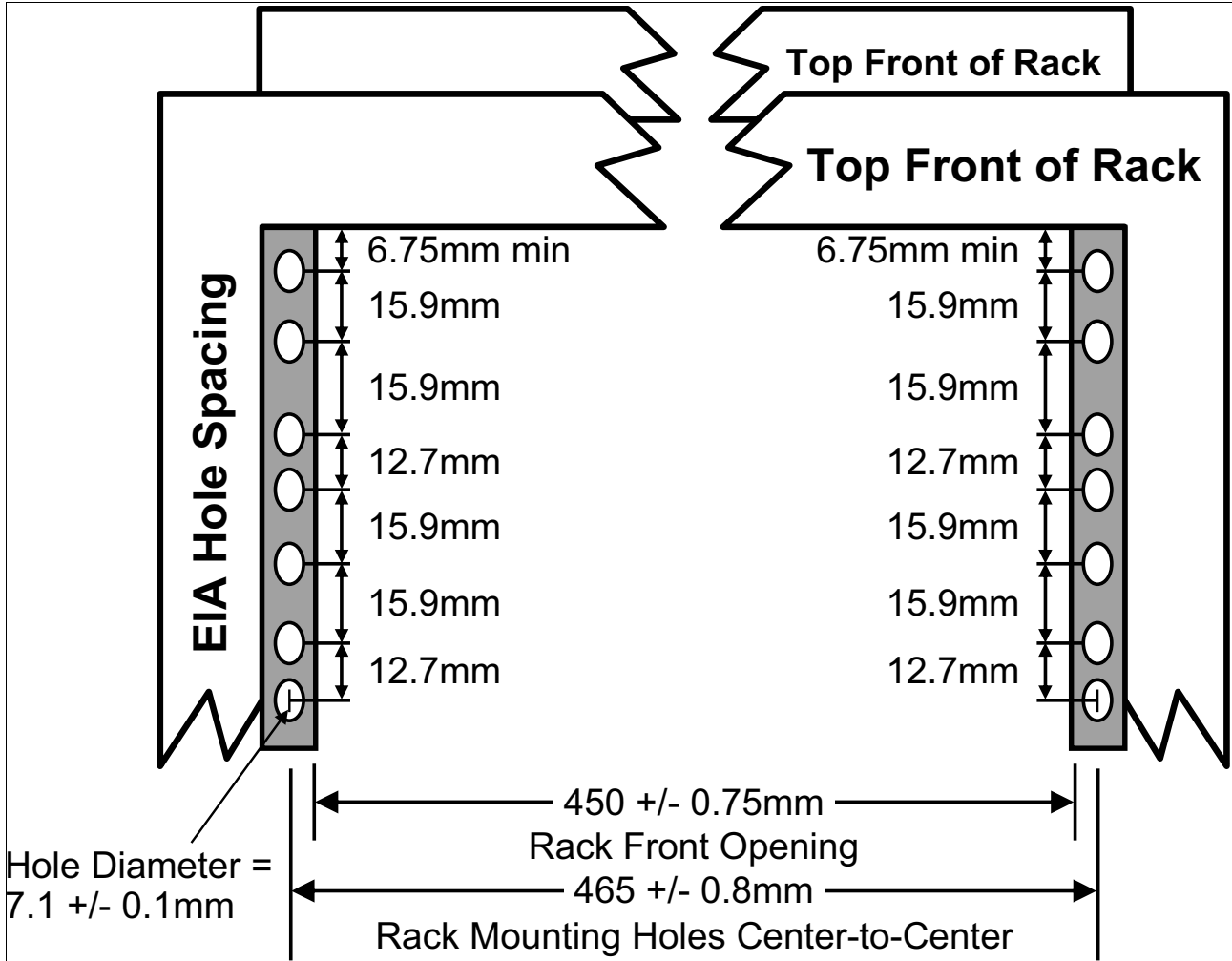


Figure 1-13 Rack specification dimensions: Top front view

1.9 Hardware Management Console

This section describes the Hardware Management Consoles (HMCs) that are available for Power Systems servers.

1.9.1 New features

Here are some of the new features of the HMCs:

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

1.9.2 Hardware Management Console overview

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

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

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

The HMC is available as a hardware appliance or as a vHMC. The Power L922 server supports several service environments, including attachment to one or more HMCs or vHMCs. This is the default configuration for servers supporting multiple logical partitions with dedicated resource or virtual I/O.

Here are the HMCs for various hardware architectures:

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

Hardware support for customer replaceable units (CRUs) come 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 L922 server.
- ▶ Integrated Virtual Management (IVM) is no longer supported.

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

Figure 1-14 shows the HMC model selections and tier updates.

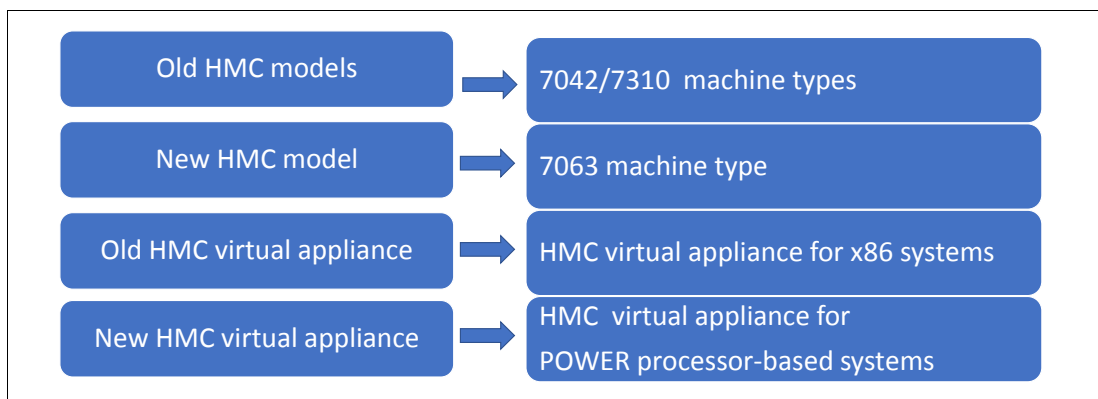


Figure 1-14 HMC selections

1.9.3 Hardware Management Console code level

The HMC code must be running at Version 9 Release 2 (V9R2) or later when you use the HMC with the Power L922 server.

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.
- ▶ HMC V9R1.911.0 added support for managing IBM OpenPOWER systems. The same HMC that is used to manage flexible service processor (FSP)-based enterprise systems can manage the baseboard management controller (BMC) based Power Systems AC and Power Systems LC servers. This support 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.

1.9.4 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. The x86-based HMCs are no longer available for ordering, but are supported as an option for managing the Power L922 server.

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

It is a preferred practice 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 by the Power L922 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 and 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
- ▶ Two 2-TB SATA LFF 3.5-inch HDD RAID 1 disks
- ▶ 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
- ▶ Four 10-Gb Ethernet Ports (RJ-45) (10 Gb/1 Gb/100 Mb)
- ▶ One 1-Gb Ethernet port for management (BMC)

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

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

1.9.5 Hardware Management Console connectivity to POWER9 processor-based systems' service processors

POWER9 processor-based systems 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 perform IBM Active Memory™ Sharing operations on PowerVM, you must have an Ethernet link to the managed partitions. A minimum of two Ethernet ports are needed on the HMC to provide such connectivity.

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

Figure 1-15 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.

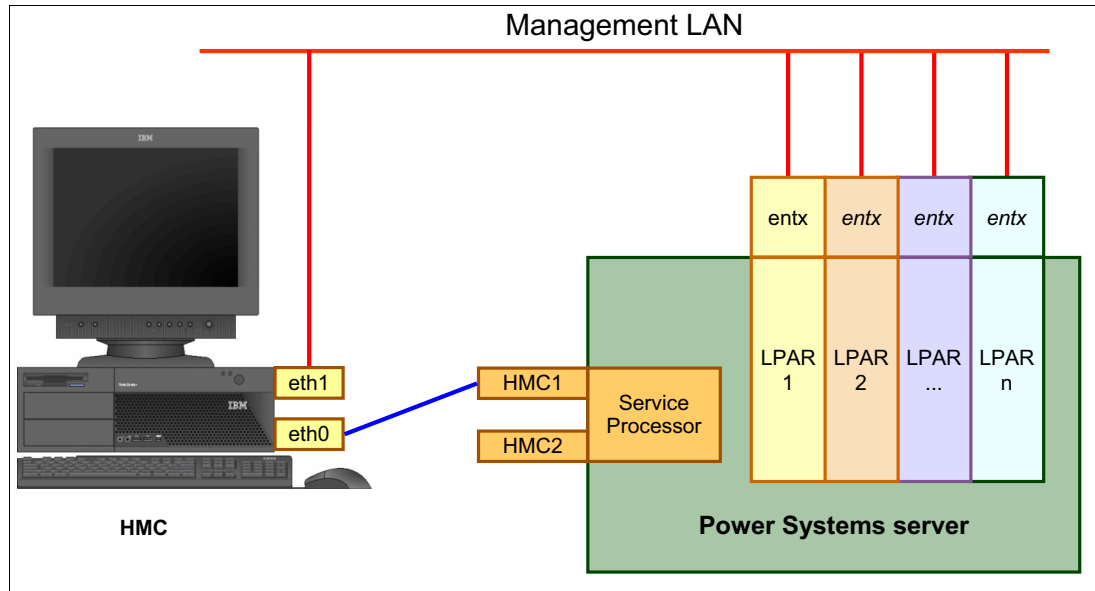


Figure 1-15 Network connections from the HMC to service processor and LPARs

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 timeout, predefined IP addresses are set up on both ports. Static IP address allocation is also an option and can be configured by using the 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. They 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

1.9.6 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 may have two HMCs attached to a system, which provides redundancy if 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 functionality.

Figure 1-16 shows one possible highly available HMC configuration that manages two servers. Each HMC is connected to one FSP port of each managed server.

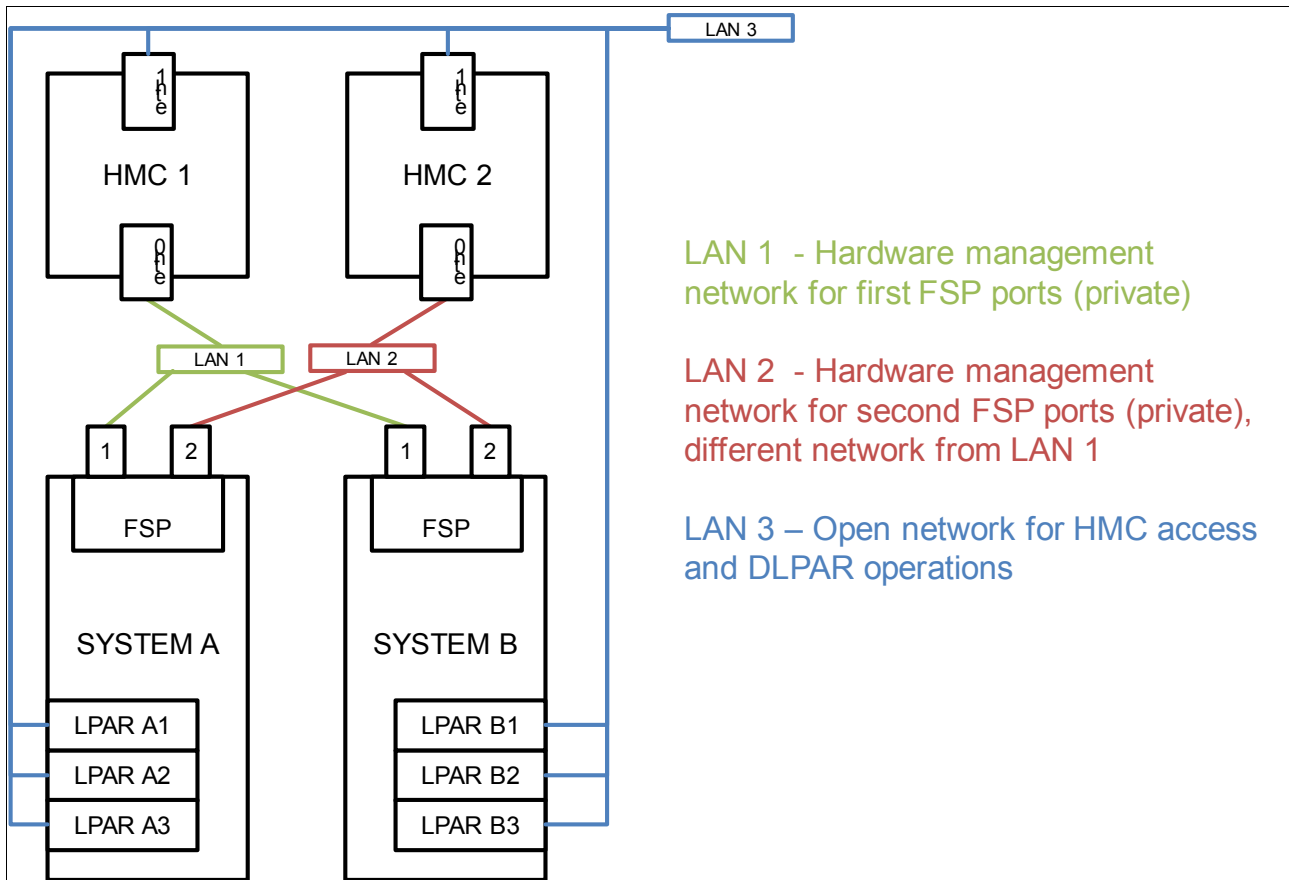


Figure 1-16 Highly available HMC networking example.

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 L922 (9008-22L) 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 IBM Power L922 logical system diagram.

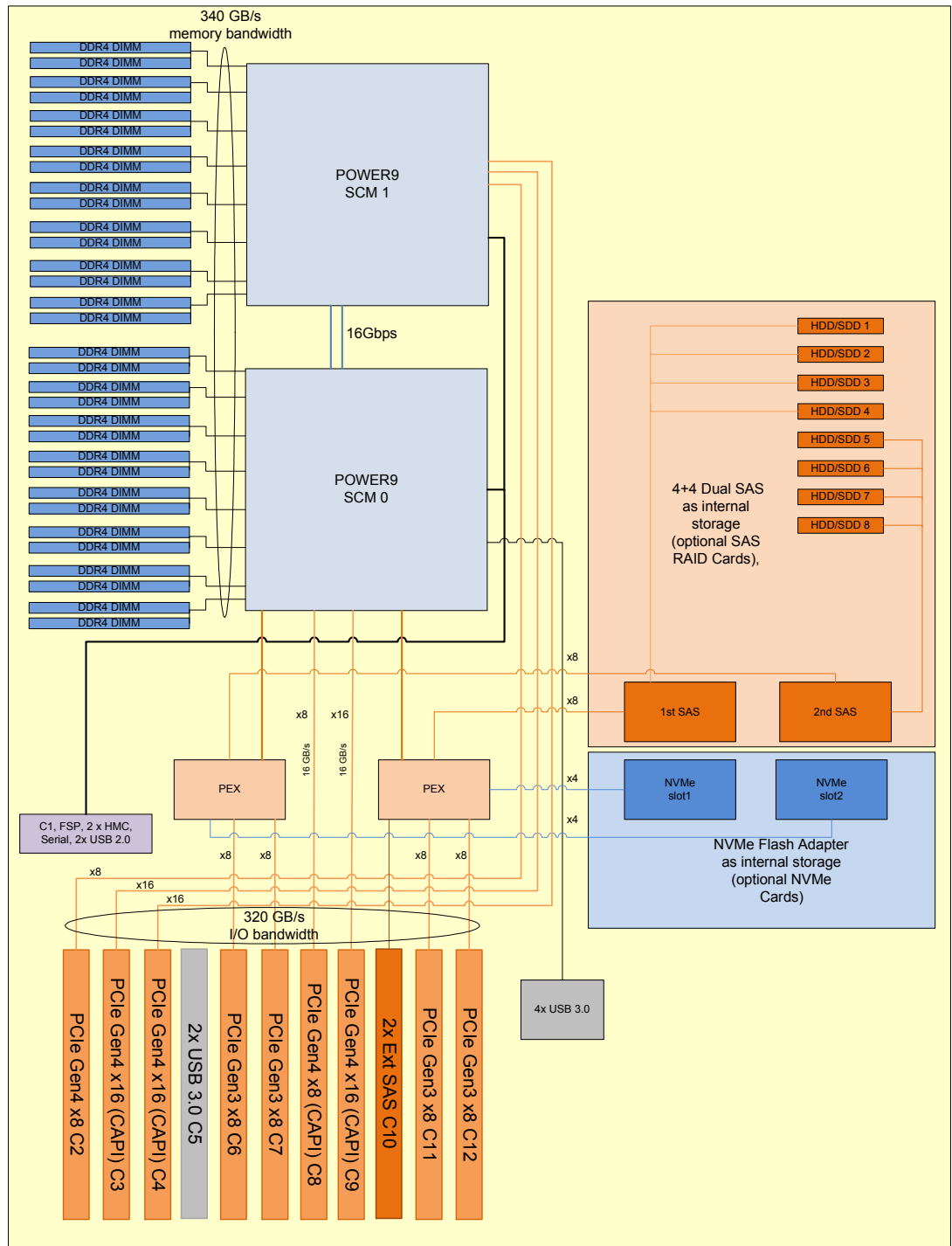


Figure 2-1 The IBM Power L922 logical system diagram

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

This is the architectural design of the POWER9 processor.

The servers are offered with various numbers of cores that are activated and a selection of clock frequencies so that IBM can make offerings at several price points, and allows customers to select a particular server (or servers) to fit their budget and performance requirements.

The POWER9 processor is single-chip module (SCM) that is 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 24 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 L2 cache, and 10 MB of L3 embedded DRAM (eDRAM). The interconnect also extends through module and system board technology to other POWER9 processors in addition to DDR4 memory and various I/O devices.

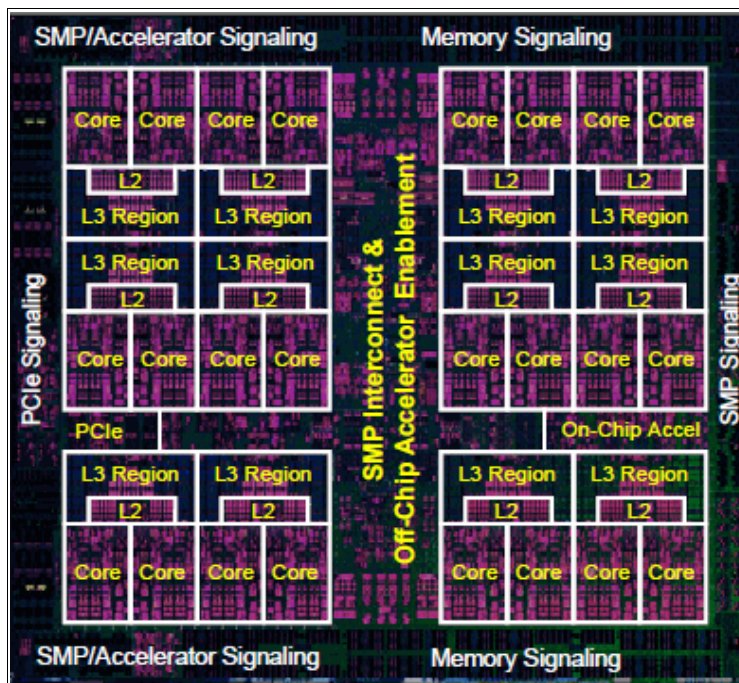


Figure 2-2 The 24-core POWER9 processor

The POWER9 processor has eight memory channels, and each channel supports up to two DDR4 DIMM slots. The Power L922 server in a two-SCM configuration can operate up to 4 TB of memory.

2.1.2 POWER9 processor features

The POWER9 chip provides embedded algorithms for the following features:

- ▶ External Interrupt Virtualization Engine. Reduces code overhead/path length and improves performance compared to the previous architecture.
- ▶ Gzip compression and decompression.
- ▶ PCIe Gen4 support.
- ▶ Two memory controllers that support direct-attached DDR4 memory.
- ▶ Cryptography through an advanced encryption standard (AES) engine.
- ▶ Random number generator (RNG).
- ▶ Secure Hash Algorithm (SHA) engine: SHA-1, SHA-256, SHA-512, Message Digest 5 (MD5).
- ▶ IBM Data Mover.

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

Table 2-1 Summary of 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

2.1.3 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 by using both local and global prediction tables with a selector table to choose the best predictor
- ▶ Improved out-of-order execution

- ▶ Two symmetric fixed-point execution units
- ▶ Two symmetric load/store units and two load units, all four of which can also run simple fixed-point instructions
- ▶ An integrated, multi-pipeline vector-scalar floating point unit for running both scalar and SIMD-type instructions, including the Vector Multimedia eXtension (VMX) instruction set and the improved Vector Scalar eXtension (VSX) instruction set, and capable of up to 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 [IBM Power ISA Version 3.0B](#).

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

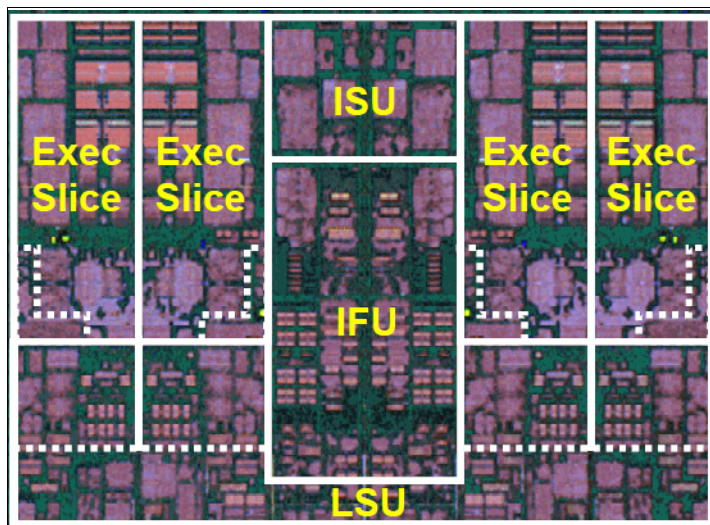


Figure 2-3 POWER9 processor chip

2.1.4 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 allows 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 (ST)	32
IBM POWER5	64	SMT2	128
IBM POWER6	64	SMT2	128
IBM POWER7	256	SMT4	1024
IBM POWER8	192	SMT8	1536
IBM POWER9	192	SMT8	1535

2.1.5 Processor feature codes

The Power L922 (9008-22L) server supports a two-processor socket with up to 24 cores. Table 2-3 shows the feature codes (FCs) for the server.

Table 2-3 Processor feature codes specification for the Power L922 server

Number of cores	Frequency	Feature code
Eight cores	3.4 - 3.9 GHz maximum	ELPV
Ten cores	2.9 - 3.8 GHz maximum	ELPW
Twelve cores	2.7 - 3.8 GHz maximum	ELPX

2.1.6 Memory access

The scale-out machines use industrial standard DDR4 DIMMs. Each POWER9 module has two memory controllers, which are connected to eight memory channels. Each memory channel can support up to two DIMMs. A single POWER9 module can support a maximum of 16 DDR4 DIMMs. The speed depends on the DIMM size and placement.

Table 2-4 shows the DIMM speeds.

Table 2-4 DIMM speeds

Registered DIMM (RDIMM) size	Mbps (One DIMM per port)	Mbps (Two DIMMs per port)
8 GB	2666	2133
16 GB	2666	2133
32 GB	2400	2133
64 GB	2400	2133
128 GB	2400	2133

The Power L922 server in a two-SCM configuration can operate up to 4 TB of memory.

Figure 2-4 shows an overview of the POWER9 direct attach memory.

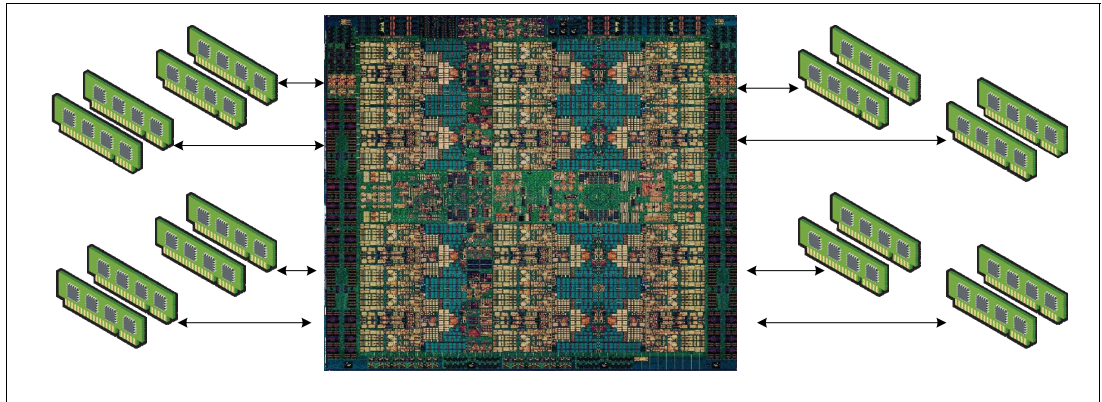


Figure 2-4 Overview of POWER9 direct attach memory

2.1.7 On-chip L3 cache innovation and Intelligent Cache

Similar to POWER8, the POWER9 processor uses a breakthrough in material engineering and microprocessor fabrication to implement the L3 cache in eDRAM and place it on the processor die. L3 cache is critical to a balanced design, as is the ability to provide good signaling between the L3 cache and other elements of the hierarchy, such as the L2 cache or SMP interconnect.

The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a fast 10 MB local region of L3 cache (FLR-L3), but also has access to other L3 cache regions as shared L3 cache. Additionally, each core can negotiate to use the FLR-L3 cache that is associated with another core, depending on the reference patterns. Data can also be cloned and stored in more than one core's FLR-L3 cache, again depending on the reference patterns. This Intelligent Cache management enables the POWER9 processor to optimize the access to L3 cache lines and minimize overall cache latencies.

Here is a list of features of the L3 cache on the POWER9 processor:

- ▶ Private 10 MB L3 cache/shared 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.

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

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

Most implementations of transactional memory are based on software. The 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 allowing the system to operate in maximum performance.

2.1.9 IBM Coherent Accelerator Processor Interface 2.0

IBM Coherent Accelerator Processor Interface (CAPI) 2.0 is the evolution of CAPI and defines a coherent accelerator interface structure for attaching special processing devices to the POWER9 processor bus. As with the original CAPI, CAPI2 can attach accelerators that have coherent shared memory access with the processors in the server and share full virtual address translation with these processors by using standard PCIe Gen4 buses with twice the bandwidth compared to the previous generation.

Applications can have customized functions in Field Programmable Gate Arrays (FPGAs) and queue work requests directly in shared memory queues to the FPGA. Applications can also have customized functions by using the same effective addresses (pointers) they use for any threads running on a host processor. From a practical perspective, CAPI enables a specialized hardware accelerator to be seen as an extra processor in the system with access to the main system memory and coherent communication with other processors in the system.

Figure 2-5 shows a comparison of the traditional model, where the accelerator must go through the processor to access memory with CAPI.

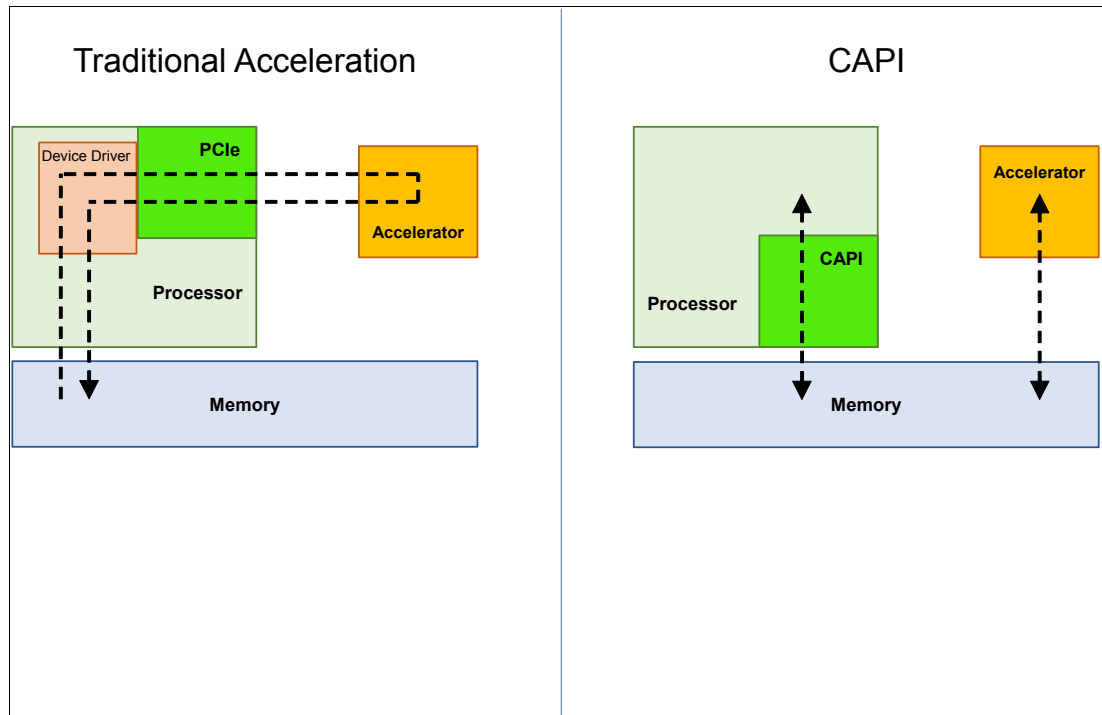


Figure 2-5 CAPI accelerator that is attached to the POWER9 processor

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

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

The implementation of CAPI on the POWER9 processor enables hardware companies to develop solutions for specific application demands. Companies use the performance of the POWER9 processor for general applications and the custom acceleration of specific functions by using a hardware accelerator with a simplified programming model and efficient communication with the processor and memory resources.

2.1.10 Power management and system performance

The POWER9 scale-out models introduced new features for EnergyScale including new variable processor frequency modes that provide a significant performance boost beyond the static nominal frequency. The following modes can be modified or disabled. The Power L922 (9008-22A) Maximum Performance mode is the default.

Disable all modes option

The processor clock frequency is set to its nominal value and the power that is used by the system remains at a nominal level. The Disable all modes option was the default for all systems before POWER9.

Static Power Save mode

Reduces the power consumption by lowering the processor clock frequency and the voltage to fixed values. This option also reduces the power consumption of the system while still delivering predictable performance.

Dynamic Power Performance mode

Causes the processor frequency to vary based on the processor use. During periods of high use, the processor frequency is set to the maximum value that is allowed, which might be above the nominal frequency. Additionally, the frequency is lowered below the nominal frequency during periods of moderate and low processor use.

Maximum Performance mode

The mode allows the system to reach the maximum frequency under certain conditions. The power consumption will increase. The maximum frequency is approximately 20% better than nominal.

The controls for all of these modes are available on the Advanced System Management Interface (ASMI) and can be dynamically modified.

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

Table 2-5 shows comparable characteristics between the POWER9, POWER8, and POWER7 processors.

Table 2-5 Comparison of the POWER9 processor and prior generations

Characteristics	POWER9	POWER8	POWER7+
Technology	14 nm	22 nm	32 nm
Die size	68.5 mm x 68.5 mm	649 mm ²	567 mm ²
Number of transistors	8 billion	4.2 billion	2.1 billion
Maximum cores	24	12	8
Maximum SMT threads per core	Four threads	Eight threads	Four threads
Maximum frequency	3.8 - 4.0 GHz	4.15 GHz	4.4 GHz
L2 cache	512 KB shared per two cores	512 KB per core	256 KB per core

Characteristics	POWER9	POWER8	POWER7+
L3 cache	10 MB of FLR-L3 cache per two cores with each core having access to the full 120 MB of L3 cache, on-chip eDRAM	8 MB of FLR-L3 cache per core with each core having access to the full 96 MB of L3 cache, on-chip eDRAM	10 MB of FLR-L3 cache per core with each core having access to the full 80 MB of L3 cache, on-chip eDRAM
Memory support	DDR4	DDR3 and DDR4	DDR3
I/O bus	PCIe Gen4	PCIe3	GX++

2.2 Memory subsystem

The Power L922 server is a two-socket server that supports up to two POWER9 processor modules. The servers support a maximum of 32 x DDR4 DIMM slots, with 16 DIMM slots per installed processor. The memory features that are supported are 8 GB, 16 GB, 32 GB, 64 GB, and 128 GB, allowing for a maximum system memory of 4 TB. Memory speeds vary depending on the DIMM size and modules placement, as shown in Table 2-6.

Table 2-6 POWER9 memory speed

RDIMM size	Mbps (1 DIMM per port)	Mbps (2 DIMMs per port)
8 GB	2400	2133
16 GB	2666	2133
32 GB	2400	2133
64 GB	2400	2133
128 GB	2400	2133

The maximum theoretical memory bandwidth for POWER9 processor module is 170 GBps. The total maximum theoretical memory bandwidth for a two-socket system is 340 GBps.

2.2.1 Memory placement rules

The following memory options are orderable:

- ▶ 8 GB DDR4 DRAM (EM60)
- ▶ 16 GB DDR4 DRAM (EM62)
- ▶ 32 GB DDR4 DRAM (EM63)
- ▶ 64 GB DDR4 DRAM (EM64)
- ▶ 128 GB DDR4 DRAM (EM65)

All memory must be ordered in pairs, with a minimum 32 GB per system per processor module that is installed.

The supported maximum memory is as follows for the Power L922 server:

- ▶ One processor module installed: 2 TB (Sixteen 128 GB DIMMs)
- ▶ Two processors modules installed: 4 TB (Thirty-two 128 GB DIMMs)

The basic rules for memory placement are:

- ▶ Each FC equates to a single physical DIMM.
- ▶ All memory features must be ordered in pairs.
- ▶ All memory DIMMs must be installed in pairs.
- ▶ Each DIMM within a pair must be of the same capacity.

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

Figure 2-6 shows the physical memory DIMM topology for the Power L922 server.

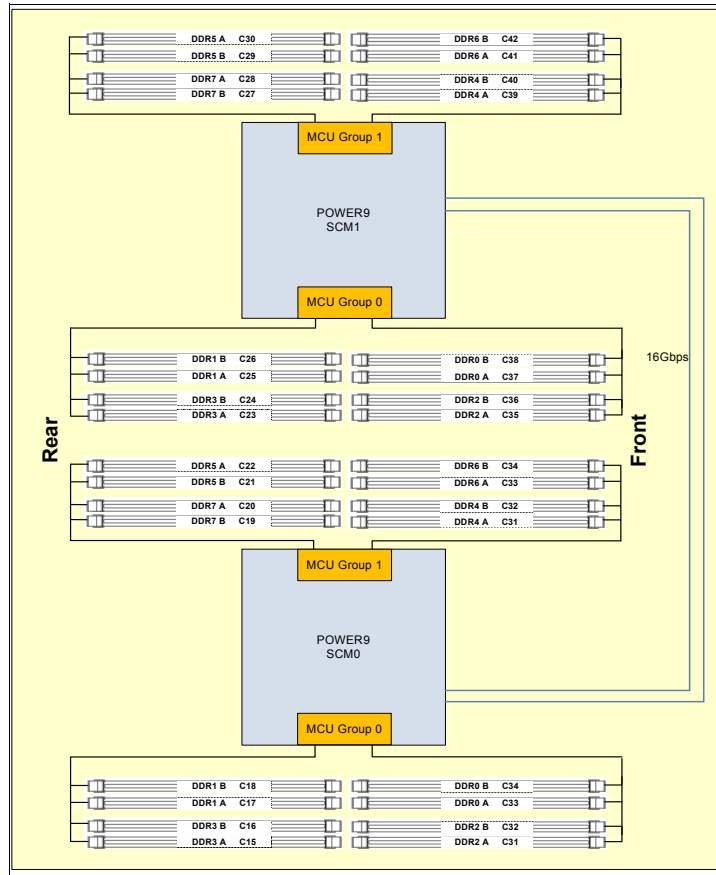


Figure 2-6 Memory DIMM topology for the Power L922 server

Figure 2-7 shows the DIMM plug sequence for the Power L922 server.

MCU Group 0				MCU Group 1				MCU Group 0				MCU Group 1			
DDR0	DDR1	DDR2	DDR3	DDR4	DDR5	DDR6	DDR7	DDR0	DDR1	DDR2	DDR3	DDR4	DDR5	DDR6	DDR7
A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
C33	C34	C17	C18	C31	C32	C15	C16	C36	C35	C22	C21	C38	C37	C20	C19
C41	C42	C25	C26	C39	C40	C23	C24	C44	C43	C30	C29	C46	C45	C28	C27
1	1			3	3			2	2			4	4		
		5	5			7	7		6	6			8	8	
	9	9			11	11			10	10			12	12	
		9	9			11	11			10	10			12	12

Figure 2-7 DIMM plug sequence for the Power L922 server

For the pair installation, follow this sequence:

- ▶ The first DIMM pair is installed at Red 1 (C33 DDR0-A and C17 DDR1-A) of SCM-0.
- ▶ The second DIMM pair is installed at Green 2 (C41 DDR0-A and C25 DDR1-A) of SCM-1.
- ▶ The third DIMM pair is installed at Gold 3 (C36 DDR4-A and C22 DDR5-A) of SCM-0.
- ▶ The fourth DIMM pair is installed at Purple 4 (C44 DDR4-A and C30 DDR5-A) of SCM-1
- ▶ The fifth DIMM pair is installed at Cyan 5 (C31 DDR2-A and C15 DDR3-A) of SCM-0.
- ▶ The sixth DIMM pair is installed at Pink 6 (C39 DDR2-A and C23 DDR3-A) of SCM-1.
- ▶ The seventh DIMM pair is installed at Gray 7 (C38 DDR6-A and C20 DDR7-A) of SCM-0.
- ▶ The eighth DIMM pair is installed at Yellow 8 (C46 DDR6-A and C28 DDR7-A) of SCM-1.

For the quad installation, follow this sequence:

- ▶ Two ninth DIMM pairs (or quad) are installed at Red 9 (C34 DDR0-B and C18 DDR1-B) and at Cyan 9 (C32DDR2-B and C16 DDR3-B) of SCM-0.
- ▶ Two 10th DIMM pairs (or quad) are installed at Green 10 (C42 DDR0-B and C26 DDR1-B) and at Pink 10(C40 DDR2-B and C24 DDR3-B) of SCM-1.
- ▶ Two 11th DIMM pairs (or quad) are installed at Gold 11 (C35 DDR4-B and C21 DDR5-B) and at Gray 11(C37 DDR6-B and C19 DDR7-B) of SCM-0.
- ▶ Two 12th DIMM pairs (or quad) are installed at Purple 12 (C43 DDR4-B and C29 DDR5-B) and at Yellow 12(C45 DDR6-B and C27 DDR7-B) of SCM-1.

More restrictions:

- ▶ You may not mix 1R DIMMs and 2R DIMMs on a single channel within an MCU group because they run at different DIMM data rates.
- ▶ DIMMs in the same color cells must be identical (same size and rank).

2.2.2 Memory bandwidth

The POWER9 processor has exceptional cache, memory, and interconnect bandwidths. The next sections show the bandwidth of the Power L922 server.

The bandwidth figures for the caches are calculated as follows:

- ▶ L1 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 clock of the core. The formulas are as follows:
 - 3.8 GHz core: $(4 * 16 \text{ B} + 2 * 16 \text{ B}) * 3.8 \text{ GHz} = 364.8 \text{ GBps}$
 - 3.9 GHz core: $(4 * 16 \text{ B} + 2 * 16 \text{ B}) * 3.9 \text{ GHz} = 374.4 \text{ GBps}$
- ▶ L2 cache: In one clock cycle, one 64-byte load operation and two 16-byte store operations can be accomplished. The value varies depending on the clock of the core. The formulas are as follows:
 - 3.8 GHz core: $(1 * 64 \text{ B} + 2 * 16 \text{ B}) * 3.8 \text{ GHz} = 364.8 \text{ GBps}$
 - 3.9 GHz core: $(1 * 64 \text{ B} + 2 * 16 \text{ B}) * 3.9 \text{ GHz} = 374.4 \text{ GBps}$
- ▶ L3 cache: One 32-byte load operation and one 32-byte store operation can be accomplished at one clock cycle. The formulas are as follows:
 - 3.8 GHz core: $(1 * 32 \text{ B} + 1 * 32 \text{ B}) * 3.8 \text{ GHz} = 243.2 \text{ GBps}$
 - 3.9 GHz core: $(1 * 32 \text{ B} + 1 * 32 \text{ B}) * 3.9 \text{ GHz} = 249 \text{ GBps}$

Power L922 server maximum bandwidths estimates

Table 2-7 shows the maximum bandwidth estimates for a single core on the Power L922 server.

Table 2-7 The Power L922 server: Single core bandwidth estimates

Single core	Power L922 server One core @ 3.8 GHz (maximum)	Power L922 server One core @ 3.9 GHz (maximum)
L1 (data) cache	364.8 GBps	374.4 GBps
L2 cache	364.8 GBps	374.4 GBps
L3 cache	243.2 GBps	249 GBps

For an entire Power L922 server that is populated with two processor modules, Table 2-8 shows the overall bandwidths.

Table 2-8 The Power L922 server: Total bandwidth maximum estimates

Total bandwidths	Power L922 server	Power L922 server	Power L922 server
	Sixteen cores @ 3.9 GHz (maximum)	Twenty cores @ 3.8 GHz (maximum)	Twenty-four cores @ 3.8 GHz (maximum)
L1 (data) cache	5990.4 GBps	7296 GBps	8755.2 GBps
L2 cache	5990.4 GBps	7296 GBps	8755.2 GBps
L3 cache	3993.6 GBps	4864 GBps	5836.8 GBps
Total memory	340 GBps	340 GBps	340 GBps
PCIe Interconnect	320 GBps	320 GBps	320 GBps

Note: There are several POWER9 design points to consider when comparing hardware designs by using SMP communication bandwidths as a unique measurement. POWER9 provides:

- ▶ More cores per socket, which leads to lower inter-CPU communication
- ▶ More RAM density (up to 2 TB per socket), allowing for less inter-CPU communication
- ▶ Greater RAM bandwidth for less dependence on the L3 cache
- ▶ Intelligent hypervisor scheduling that places RAM usage close to the CPU
- ▶ New SMP routing so that multiple channels are available when congestion occurs

2.3 System bus

This section provides more information about the internal buses.

The Power L922 server has internal I/O connectivity through PCIe Gen4 and Gen3 (PCI Express Gen4/Gen3, or PCIe Gen4/Gen3) slots, and external connectivity through SAS adapters.

The internal I/O subsystem on the Power L922 server is connected to the PCIe controllers on a POWER9 processor in the system. The IBM Power System in a two-socket configuration has a bus that has 80 PCIe G4 lanes running at maximum 16 Gbps full-duplex, and provides 320 GBps of I/O connectivity to the PCIe slots, SAS internal adapters, and USB ports.

Some PCIe slots are connected directly to the PCIe Gen4 buses on the processors, and PCIe3 devices are connected to these buses through PCIe3 Switches. For more information about which slots are connected directly to the processor and which ones are attached to a PCIe3 Switch (referred as PEX), see Figure 2-1 on page 36.

Figure 2-8 compares the POWER8 and POWER9 I/O bus architectures.

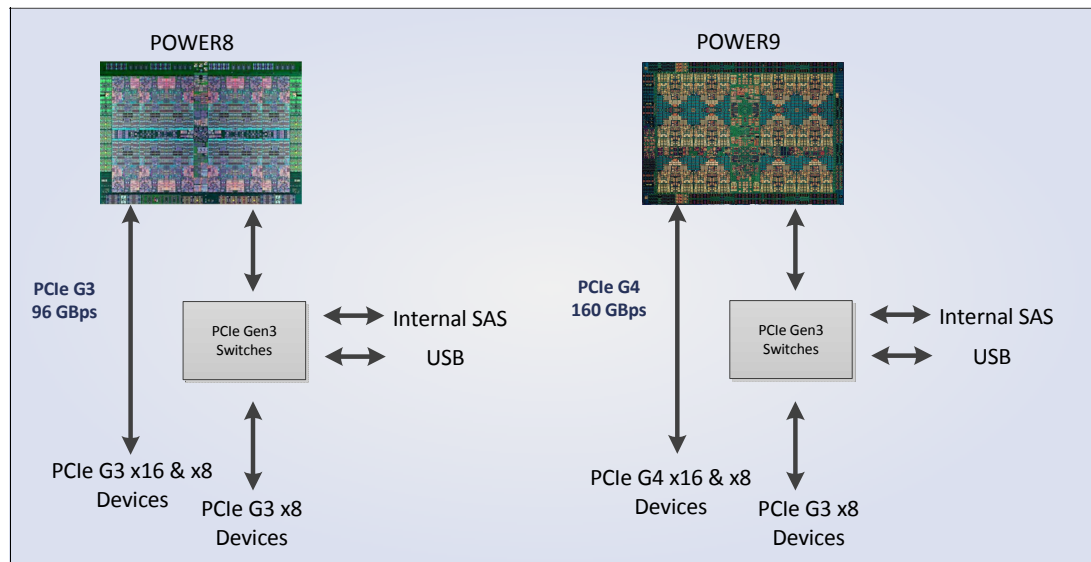


Figure 2-8 Comparison of POWER and POWER9 I/O bus architectures

Table 2-9 lists the I/O bandwidth of Power L922 processor configurations.

Table 2-9 I/O bandwidth

I/O	I/O bandwidth (maximum theoretical)
Total I/O bandwidth	Power L922 server with two processors: <ul style="list-style-type: none"> ▶ 160 GBps simplex ▶ 320 GBps duplex

Each POWER9 processor module has 40 PCIe lanes running at 16 Gbps full-duplex. The bandwidth formula is calculated as follows:

Forty lanes * 2 processors * 16 Gbps * 2 = 320 GBps

2.4 Internal I/O subsystem

The internal I/O subsystem is on the system board, which supports PCIe slots. PCIe adapters on the Power L922 server are hot-pluggable.

All PCIe slots support enhanced error handling (EEH). PCI EEH-enabled adapters respond to a special data packet that is generated from the affected PCIe slot hardware by calling system firmware, which examines the affected bus, allows the device driver to reset it, and continues without a system restart.

2.4.1 Slot configuration

The Power L922 server provides PCIe3 and PCIe Gen4 slots. The number of PCIe slots that are available on the Power L922 server depends on the number of installed processors. Table 2-10 provides information about the PCIe slots in the Power L922 server.

Table 2-10 PCIe slot locations and descriptions for the Power L922 server

Slot availability	Description	Adapter size
Two slots (P1-C6, P1-C12)	PCIe3 x8	Half-height, half-length
Two slots (P1-C7, P1-C11)	PCIe3 x8	Half-height, half-length
Three slots (P1-C3 ^a , P1-C4 ^a , P1-C9)	PCIe Gen4 x16	Half-height, half-length
Two slots (P1-C2 ^a , P1-C8)	PCIe Gen4 x8 with x16 connector	Half-height, half-length

a. The slot is available when the second processor slot is populated.

Table 2-11 lists the PCIe adapter slot locations and details for the Power L922 server.

Table 2-11 PCIe slot locations and details for the Power L922 server

Location code	Description	Slot capabilities		
		CAPI	Single Root I/O Virtualization (SR-IOV)	I/O adapter enlarged capacity enablement order ^a
P1-C2 ^b	PCIe Gen4 x8 with x16 connector	No	Yes	5
P1-C3 ^b	PCIe Gen4 x16	Yes	Yes	2
P1-C4 ^b	PCIe Gen4 x16	Yes	Yes	3
P1-C6	PCIe3 x8 with x16 connector	No	Yes	6
P1-C7	PCIe3 x8	No	Yes	10
P1-C8 ^b	PCIe Gen4 x8 with x16 connector	Yes	Yes	4

Location code	Description	Slot capabilities		
		CAPI	Single Root I/O Virtualization (SR-IOV)	I/O adapter enlarged capacity enablement order ^a
P1-C9 ^b	PCIe Gen4 x16	Yes	Yes	1
P1-C11	PCIe3 x8 (default LAN slot)	No	Yes	11
P1-C12	PCIe3 x8 with x16 connector	No	Yes	7

a. Enabling the I/O adapter enlarged capacity option affects only Linux partitions.

b. A high-performance slot that is directly connected to the processor module. The connectors in these slots are differently colored than the slots from the PCIe3 switches.

Figure 2-9 shows the rear view of the Power L922 server with the location codes for the PCIe adapter slots.

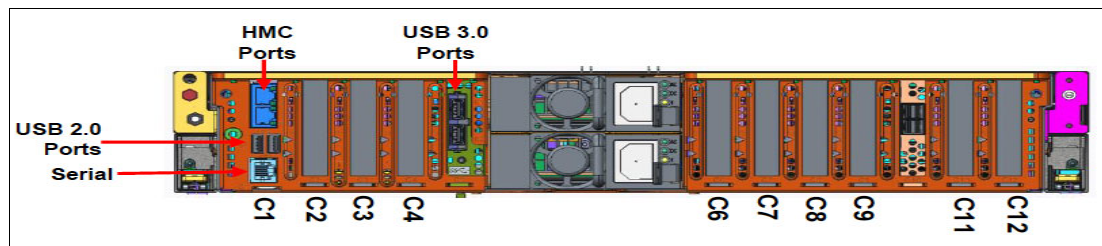


Figure 2-9 Rear view of a rack-mounted Power L922 server with PCIe slots location codes

2.4.2 System ports

The system board has one serial port that is called a *system port*. The one system port is RJ45, and is supported by Linux for attaching serial devices such as an asynchronous device. If the device does not have an RJ45 connection, a converter cable such as 3930 can provide a 9-pin D-shell connection.

2.5 PCIe adapters

This section covers the various types and functions of the PCI adapters that are supported by the Power L922 servers.

2.5.1 PCI Express

PCIe uses a serial interface and allows for point-to-point interconnections between devices (by using a directly wired interface between these connection points). A single PCIe serial link is a dual-simplex connection that uses two pairs of wires, one pair for transmit and one pair for receive, and can transmit only 1 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 interface that is supported on this server is PCIe Gen4, which is capable of 16 GBps simplex (32 GBps duplex) on a single x16 interface. PCIe Gen4 slots also support previous generations (Gen2 and Gen1) adapters, which operate at lower speeds, according to the following rules:

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

All adapters support EEH. PCIe adapters use a different type of slot than PCI adapters. If you attempt to force an adapter into the wrong type of slot, you might damage the adapter or the slot.

IBM Power L922 servers use PCIe low profile (LP) cards.

PCIe full height and full high cards are not compatible with Power L922 servers.

Before adding or rearranging adapters, use the [IBM System Planning Tool \(SPT\)](#) 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. Use [IBM Prerequisites](#).

The following sections describe the supported adapters and provide tables of orderable FCs.

2.5.2 LAN adapters

To connect the Power L922 servers to a local area network (LAN), you can use the LAN adapters that are supported in the PCIe slots of the system unit.

Table 2-12 lists the available LAN adapters for Power L922 servers.

Table 2-12 Available LAN adapters in a Power L922 server

Feature code	CCIN	Description	Minimum	Maximum
EL55	2CC4	PCIe2 2-port 10/1 GbE BaseT RJ45 Adapter	0	12
EN0U	2CC3	PCIe2 4-port (10 Gb+1 GbE) Copper SFP+RJ45 Adapter	0	12
EN0S	2CC3	PCIe2 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter	0	12
5899	576F	PCIe2 4-port 1 GbE Adapter	0	6
EL4L	576F	PCIe2 4-port 1 GbE Adapter	0	12
EL3Z	2CC4	PCIe2 LP 2-port 10/1 GbE BaseT RJ45 Adapter	0	9
EN0V	2CC3	PCIe2 LP 4-port (10 Gb+1 GbE) Copper SFP+RJ45 Adapter	0	9
EN0T	2CC3	PCIe2 LP 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter	0	9
EL4M	576F	PCIe2 LP 4-port 1 GbE Adapter	0	9
EC2S	58FA	PCIe3 2-Port 10 Gb NIC & ROCE SR/Cu Adapter	0	2
EL53	57BC	PCIe3 2-port 10 GbE NIC&RoCE SFP+ Copper Adapter	0	12

Feature code	CCIN	Description	Minimum	Maximum
EC2U	58FB	PCIe3 2-Port 25/10 Gb NIC & ROCE SR/Cu Adapter	0	4
EC3B	57BD	PCIe3 2-Port 40 GbE NIC RoCE QSFP+ Adapter	0	8
EL57	2CC1	PCIe3 4-port (10 Gb FCoE & 1 GbE) SFP+Copper & RJ45	0	12
EL56	2B93	PCIe3 4-port (10 Gb FCoE & 1 GbE) SR & RJ45	0	12
EN15	2CE3	PCIe3 4-port 10 GbE SR Adapter	0	12
EC3T	2CEB	PCIe3 LP 1-port 100 Gb EDR IB Adapter x16	0	3
EC2R	58FA	PCIe3 LP 2-Port 10 Gb network interface card (NIC) & ROCE SR/Cu Adapter	0	5
EL3X	57BC	PCIe3 LP 2-port 10 GbE NIC&RoCE SFP+ Copper Adapter	0	9
EC3E	2CEA	PCIe3 LP 2-port 100 Gb EDR IB Adapter x16	0	3
EC3L	2CEC	PCIe3 LP 2-port 100 GbE (NIC & RoCE) QSFP28 Adapter x16	0	3
EC2T	58FB	PCIe3 LP 2-Port 25/10 Gb NIC & ROCE SR/Cu Adapter	0	8
EC3A	57BD	PCIe3 LP 2-Port 40 GbE NIC RoCE QSFP+ Adapter	0	8
EN0N	2CC0	PCIe3 LP 4-port (10 Gb FCoE & 1 GbE) LR & RJ45 Adapter	0	8
EL3C	2CC1	PCIe3 LP 4-port (10 Gb FCoE & 1 GbE) Service Focal Point (SFP) + Copper & RJ45	0	9
EL38	2B93	PCIe3 LP 4-port (10 Gb Fibre Channel over Ethernet (FCoE) & 1 GbE) SRIOV SR & RJ45	0	9
EC62	2CF1	PCIe4 LP 1-port 100 Gb EDR InfiniBand CAPI adapter	0	3
EC64	2CF2	PCIe4 LP 2-port 100 Gb EDR InfiniBand CAPI adapter	0	3
EC67	2CF3	PCIe4 LP 2-port 100 Gb ROCE EN LP adapter	0	3

2.5.3 Graphics accelerator adapters

An adapter can be configured to operate in either 8-bit or 24-bit color modes. The adapter supports both analog and digital monitors.

Table 2-13 lists the available graphics accelerator adapter for the Power L922 server.

Table 2-13 The graphics accelerator card that is supported in the Power L922 servers

Feature code	CCIN	Description	Minimum	Maximum
5269	5269	PCIe LP POWER GXT145 Graphics Accelerator	0	6

2.5.4 SAS adapters

Table 2-14 lists the SAS adapters that are available for Power L922 servers.

Table 2-14 The PCIe SAS adapters that are supported in Power L922 servers

Feature code	CCIN	Description	Minimum	Maximum
EJ10	57B4	PCIe3 SAS Tape/DVD Adapter Quad-port 6 Gb x8	0	8
EJ14	57B1	PCIe3 12 GB Cache RAID PLUS SAS Adapter Quad-port 6 Gb x8	0	8
EL3B	57B4	PCIe3 LP RAID SAS Adapter Quad-Port 6 Gb x8	0	7
EL59	57B4	PCIe3 RAID SAS Adapter Quad-port 6 Gb x8	0	8
EL60	57B4	PCIe3 LP SAS Tape/DVD Adapter Quad-port 6 Gb x8	0	7

2.5.5 Fibre Channel adapters

The server supports direct or SAN connection to devices that use Fibre Channel adapters.

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

Table 2-15 summarizes the available Fibre Channel adapters for Power L922 systems. They all have LC connectors.

Table 2-15 The PCIe Fibre Channel adapters that are available for Power L922 servers

Feature code	CCIN	Description	Minimum	Maximum
5729	5729	PCIe2 8 Gb 4-port Fibre Channel Adapter	0	12
EL2N	577D	PCIe LP 8 Gb 2-Port Fibre Channel Adapter	0	8
EL43	577F	PCIe3 LP 2-port 16 Gb Fibre Channel Adapter	0	8
EL58	577D	PCIe 8 Gb 2-port Fibre Channel Adapter	0	12
EL5B	577F	PCIe3 16 Gb 2-port Fibre Channel Adapter	0	12
EL5U	578F	PCIe3 32 Gb 2-port Fibre Channel Adapter	0	12
EL5V	578F	PCIe3 LP 2-port 32 Gb Fibre Channel Adapter	0	8
EL5W	578E	PCIe3 16 Gb 4-port Fibre Channel Adapter	0	12
EL5X	578E	PCIe3 LP 16 Gb 4-port Fibre Channel Adapter	0	8
EL5Y	578D	PCIe2 LP 8 Gb 2-Port Fibre Channel Adapter	0	8
EL5Z	578D	PCIe2 8 Gb 2-Port Fibre Channel Adapter	0	12
EN0Y	N/A	PCIe2 LP 8 Gb 4-port Fibre Channel Adapter	0	8

Note: The usage of N_Port ID Virtualization (NPIV) through the Virtual I/O Server (VIOS) requires an NPIV-capable Fibre Channel adapter, such as 5729.

2.5.6 Fibre Channel over Ethernet

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

Figure 2-10 compares existing Fibre Channel and network connections and FCoE connections.

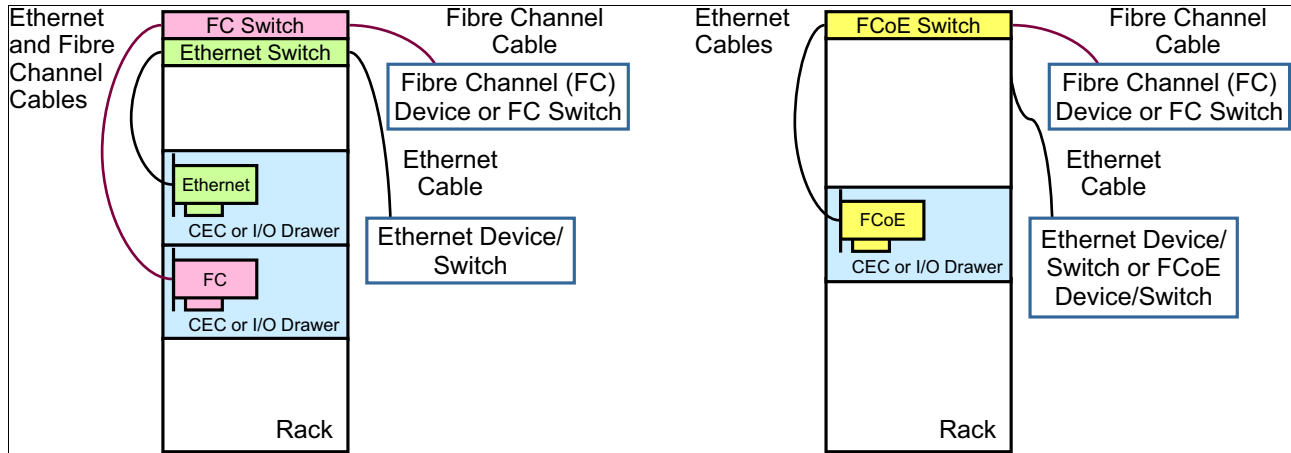


Figure 2-10 Comparison between an existing Fibre Channel and network connection and FCoE connection

FCoE adapters are high-performance, Converged Network Adapters (CNAs) that use SR optics. Each port can simultaneously provide NIC traffic and Fibre Channel functions.

Table 2-16 lists the available FCoE adapters that are available for Power L922 servers.

Table 2-16 The FCoE adapters that available for Power L922 servers

Feature code	CCIN	Description	Minimum	Maximum
EL38	2B93	PCIe3 LP 4-port (10 Gb FCoE & 1 GbE) SRIOV SR & RJ45	0	9
EL3C	2CC1	PCIe3 LP 4-port (10 Gb FCoE & 1 GbE) SFP+Copper & RJ45	0	9
EL56	2B93	PCIe3 4-port (10 Gb FCoE & 1 GbE) SR & RJ45	0	12
EL57	2CC1	PCIe3 4-port (10 Gb FCoE & 1 GbE) SFP+Copper & RJ45	0	12
EN0N	2CC0	PCIe3 LP 4-port (10 Gb FCoE & 1 GbE) LR & RJ45 Adapter	0	8

2.5.7 InfiniBand host channel adapter

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

InfiniBand is an open set of interconnect standards and specifications. The main InfiniBand specification is published by the IBTA and is available at the [IBTA website](#).

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

The InfiniBand physical connection consists of multiple byte lanes. Each individual byte lane is a four-wire, 2.5, 5.0, or 10.0 Gbps bidirectional connection. Combinations of link width and byte lane speed allow for overall link speeds of 2.5 - 120 Gbps. The architecture defines a layered hardware protocol and also a software layer to manage initialization and the communication between devices. Each link can support multiple transport services for reliability and multiple prioritized virtual communication channels.

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

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

Table 2-17 lists the InfiniBand adapters that are available for Power L922 servers.

Table 2-17 InfiniBand adapters that are available for Power L922 servers

Feature code	CCIN	Description	Minimum	Maximum
EC62	2CF1	PCIe4 LP 1-port 100 Gb EDR InfiniBand CAPI adapter	0	3
EC64	2CF2	PCIe4 LP 2-port 100 Gb EDR InfiniBand CAPI adapter	0	3

2.5.8 Cryptographic coprocessor

The cryptographic coprocessor card that is supported for the Power L922 server is shown in Table 2-18.

Table 2-18 Cryptographic coprocessor that is available for the Power L922 server

Feature code	CCIN	Description	Minimum	Maximum
EJ33	4767	PCIe3 Crypto Coprocessor BSC-Gen3 4767	0	12

2.5.9 Coherent Accelerator Processor Interface adapters

The CAPI-capable adapters that are available for Power L922 servers are shown in Table 2-19.

Table 2-19 CAPI-capable adapters that are available for Power L922 servers

Feature code	CCIN	Description	Minimum	Maximum
EC62	2CF1	PCIe4 LP 1-port 100 Gb EDR InfiniBand CAPI adapter	0	3
EC64	2CF2	PCIe4 LP 2-port 100 Gb EDR InfiniBand CAPI adapter	0	3

2.5.10 USB adapters

The USB adapters that are available for Power L922 servers are shown in Table 2-19.

Table 2-20 USB adapters that are available for Power L922 servers

Feature code	CCIN	Description	Minimum	Maximum
EC45	58F9	PCIe2 LP 4-Port USB 3.0 Adapter	0	8
EC46	58F9	PCIe2 4-Port USB 3.0 Adapter	0	12

2.6 Internal storage

The internal storage on the Power L922 servers depends on the DASD/media backplane that is used. The servers support various DASD/media backplanes:

Two different features are available for the storage backplane:

- ▶ #EL66: Eight SFF-3s with an optional split card #EL68
- ▶ #EC59: Optional PCIe3 Non-Volatile Memory express (NVMe) carrier card with two M.2 module slots

2.6.1 Backplane (#EL66)

This backplane option provides SFF-3 SAS bays in the system unit. These 2.5-inch or small form factor (SFF) SAS bays can contain SAS drives (hard disk drives (HDDs) or solid-state drives (SSDs)) that are mounted on a Gen3 tray or carrier. Thus, the drives are designated SFF-3. SFF-1 or SFF-2 drives do not fit in an SFF-3 bay. All SFF-3 bays support concurrent maintenance or hot-plug capability.

This backplane option uses leading-edge, integrated SAS RAID controller technology that is designed and patented by IBM. A custom-designed PowerPC based ASIC chip is the basis of these SAS RAID controllers, and provides industry-leading RAID 5 and RAID 6 performance levels, especially for SSDs. Internally, SAS ports are implemented and provide plenty of bandwidth. The integrated SAS controllers are placed in dedicated slots and do not reduce the number of available PCIe slots.

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

This FC provides one integrated SAS adapter with no cache running eight SFF-3 SAS bays in the system unit.

For split backplane capability, add #EL68H.

The supported operating systems are:

- ▶ Red Hat Enterprise Linux
- ▶ SUSE Linux Enterprise Server
- ▶ Ubuntu Server

2.6.2 Split backplane option (#EL68)

This feature modifies the base storage backplane cabling and adds a second, high-performance SAS controller. The existing eight SFF-3 SAS bays are cabled to be split

into two sets of four bays, each with one SAS controller. Both SAS controllers are located in integrated slots and do not use a PCIe slot.

The high-performance SAS controllers each provide RAID 0, RAID 5, RAID 6, and RAID 10 support. JBOD support for HDDs is also supported. There is no write cache on either controller.

Both 5xx and 4 KB sector HDDs/SSDs are supported. 5xx and 4 KB drives cannot be mixed in the same array.

This FC provides a second integrated SAS adapter with no cache and internal cables to provide two sets of four SFF-3 bays in the system unit.

The supported operating systems are:

- ▶ Red Hat Enterprise Linux
- ▶ SUSE Linux Enterprise Server
- ▶ Ubuntu Server

2.6.3 PCIe3 NVMe express carrier card w/2 M.2 module slots (#EC59)

The NVMe option offers fast start times and is ideally suited as the location of the rootvg of VIOS partitions.

#EC59 is a carrier card for 400 GB Mainstream SSD (#ES14). You may have a maximum quantity of 2 #ES14s per #EC59.

This FC provides a PCIe3 NVMe card with 2 M.2 module slots. You must have an #ES14.

The supported operating systems are:

- ▶ SUSE Linux Enterprise Server 12 Service Pack 3, or later.
- ▶ SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 12 Service Pack 3, or later.
- ▶ Red Hat Enterprise Linux.
- ▶ Ubuntu Server.
- ▶ AIX is supported (for VIOS).

If #EC59 is selected, no disk units must be ordered. If you do not order #EC59 or #0837, then at least one disk unit must be ordered. If no HDD/SSD/SAN boot (#0837) is ordered, then #EC59 (with at least one #ES14) is the load source.

Figure 2-11 shows the location of #EC59 in a Power L922 server.

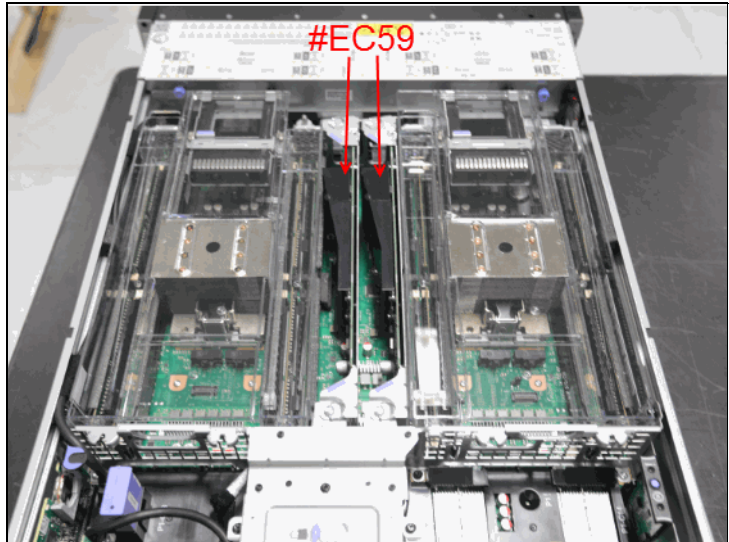


Figure 2-11 Two #EC59s in a Power L922 server

Figure 2-12 on page 59 shows an #EC59.

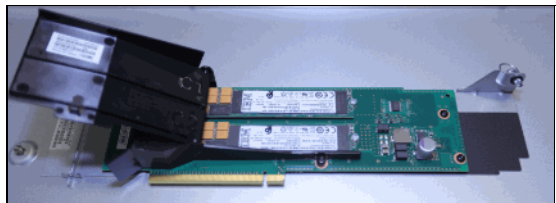


Figure 2-12 An #EC59 with the cover open showing two #EC14 modules

A nice feature of using this NVMe technology is that each #EC14 appears to the operating system as an individual disk.

Tip: If two #EC59s are configured, each with two #EC14s, it is possible to have the rootvg of the first VIOS mirrored to an #EC14 in each #EC59, and the second VIOS can be mirrored to the other two modules. This setup provides excellent performance and resilience.

2.6.4 400 GB SSD Non-Volatile Memory express M.2 module (#EC14)

This FC is a 400 GB Mainstream SSD that is formatted in 4096-byte sectors (4 KB). The drive is mounted on the PCIe NVMe Carrier Card W/ 2 M.2 Sockets (#EC59). The Drive Write Per Day (DWPD) rating is 1 calculated over a 5-year period. Approximately 1,095 TB of data can be written over the life of the drive, but depending on the nature of the workload, it might be larger. Use this FC for boot support and non-intensive workloads. Boot is supported.

Note: You must order, at a minimum, one #ES14 module with each #EC59 that is ordered. You may order a maximum of 2 #ES14s per #EC59.

Using this FC for anything than boot support and non-intensive workloads might result in throttled performance and high temperatures that lead to timeouts and critical thermal warnings.

The supported operating systems are:

- ▶ SUSE Linux Enterprise Server 12 Service Pack 3, or later.
- ▶ SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 12 Service Pack 3, or later.
- ▶ Red Hat Enterprise Linux.
- ▶ Ubuntu Server.
- ▶ AIX is supported (for VIOS).

Figure 2-13 shows two #EC14s.

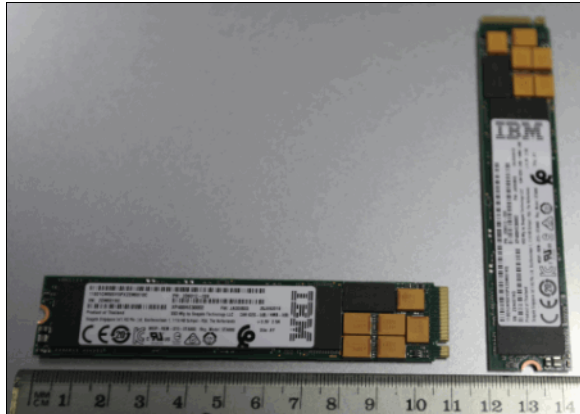


Figure 2-13 Two #EC14 NVMe modules

2.6.5 RAID support

There are multiple protection options for HDDs/SSDs in the Power L922 servers, whether they are contained in the SAS SFF bays in the system unit or drives in disk-only I/O drawers. Although protecting drives is always recommended, users can choose to leave a few or all drives unprotected at their own risk, and IBM supports these configurations.

Drive protection

HDD/SSD drive protection can be provided by AIX (VIOS), Linux, or by the HDD/SSD hardware controllers.

The default storage backplanes (#EL66) in the Power L922 server contains one SAS HDD/SSD controller, and supports JBOD and RAID 0, 5, 6, and 10 for AIX (VIOS) or Linux. A secondary non-redundant controller is added when using split backplane (#EL68), so each of the six disk bays has a separated disk controller.

When you choose the optional #EJ1D, #EJ1M, or #EL66, #EL67, and #EL68 storage backplane, the controller is replaced by a pair of high-performance RAID controllers with dual integrated SAS controllers with 1.8 GB of physical write cache. High-performance controllers run 18 SFF-3 SAS bays, 1.8-inch SSD cage bays dual controllers (also called dual I/O adapters or paired controllers), and their write caches are placed in integrated slots and do not use PCIe slots. Patented active/active configurations with at least two arrays are supported.

The write cache, which is responsible for increasing write performance by caching data before it is written to the physical disks, can have its data compression capabilities activated, providing up to 7.2 GB effective cache capacity. The write cache contents are protected against power loss by flash memory and super capacitors, which removes the need for battery maintenance.

The high-performance SAS controllers provide RAID 0, RAID 5, RAID 6, and RAID 10 support, and its Easy Tier variants (RAID 5T2, RAID 6T2, and RAID 10T2) if the server has both HDDs and SSDs installed.

The Easy Tier function is supported, so the dual controllers can automatically move hot data to an attached SSD and cold data to an attached HDD for Linux, and VIOS environments. If an EXP 24S SFF Gen2-bay Drawer (#5887) is attached to the adapters, the Easy Tier function is extended to the disks on this drawer. To learn more about Easy Tier, see 2.6.6, “Easy Tier” on page 62.

Linux can use disk drives that are formatted with 512-byte blocks when they are mirrored by the operating system. These disk drives must be reformatted to 528-byte sectors when they are used in RAID arrays. Although a small percentage of the drive's capacity is lost, extra data protection, such as error-correcting code (ECC) and bad block detection, is gained in this reformatting. For example, a 300 GB disk drive, when reformatted, provides approximately 283 GB.

Supported RAID functions

The base hardware supports RAID 0, 5, 6, and 10. When more features are configured, the server supports hardware RAID 0, 5, 6, 10, 5T2, 6T2, and 10T2:

- ▶ RAID 0 provides striping for performance, but does not offer any fault tolerance.
The failure of a single drive results in the loss of all data on the array. This version of RAID increases I/O bandwidth by simultaneously accessing multiple data paths.
- ▶ RAID 5 uses block-level data striping with distributed parity.
RAID 5 stripes both data and parity information across three or more drives. Fault tolerance is maintained by ensuring that the parity information for any given block of data is placed on a drive that is separate from the ones that are used to store the data itself. This version of RAID provides data resiliency if a single drive fails in a RAID 5 array.
- ▶ RAID 6 uses block-level data striping with dual distributed parity.
RAID 6 is the same as RAID 5 except that it uses a second level of independently calculated and distributed parity information for 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 an equivalent storage capacity. This version of RAID provides data resiliency if one or two drives fail in a RAID 6 array. When you work with large capacity disks, RAID 6 allows you to sustain data parity during the rebuild process.
- ▶ RAID 10 is a striped set of mirrored arrays.
It is a combination of RAID 0 and RAID 1. A RAID 0 stripe set of the data is created across a two-disk array for performance benefits. A duplicate of the first stripe set is then mirrored on another two-disk array for fault tolerance. This version of RAID provides data resiliency if a single drive fails, and it can provide resiliency for multiple drive failures.

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

2.6.6 Easy Tier

The server can handle both HDDs and SSDs that are attached to its storage backplane if they are on separate arrays.

When the HDDs and SSDs 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 setup 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 2-14 shows an Easy Tier array.

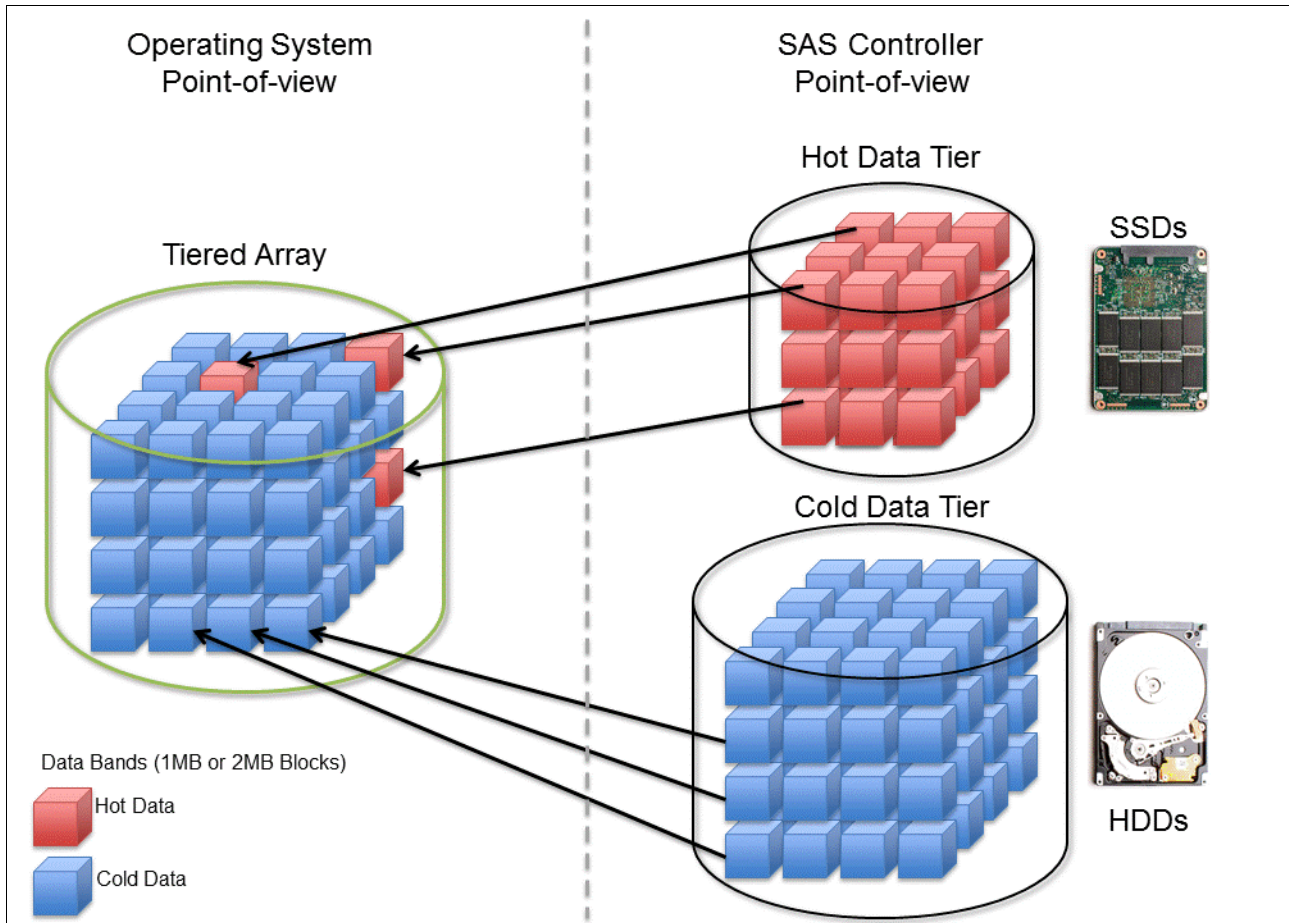


Figure 2-14 Easy Tier

2.7 External IO subsystems

This section describes the PCIe3 I/O expansion drawer that can be attached to the Power L922 server.

2.7.1 PCIe3 I/O expansion drawer

The PCIe3 I/O expansion (#ELMX) drawer is a 4U high, PCI Gen3-based, and rack-mountable I/O drawer. It offers two PCIe fan-out modules (#ELMF or #ELMG). The PCIe fan-out module provides six PCIe3 full-high, full-length slots (two x16 and four x8). The PCIe slots are hot pluggable.

The PCIe fan-out module has two CXP ports that are connected two CXP ports on a PCIe Optical Cable Adapter (#EJ05) on the server. A pair of AOCs or a pair of CXP copper cables are used for this connection.

Concurrent repair and add/removal of PCIe adapters are done by Hardware Management Console (HMC) guided menus or by operating system support utilities.

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

Figure 2-15 shows the back view of the PCIe3 I/O expansion drawer.

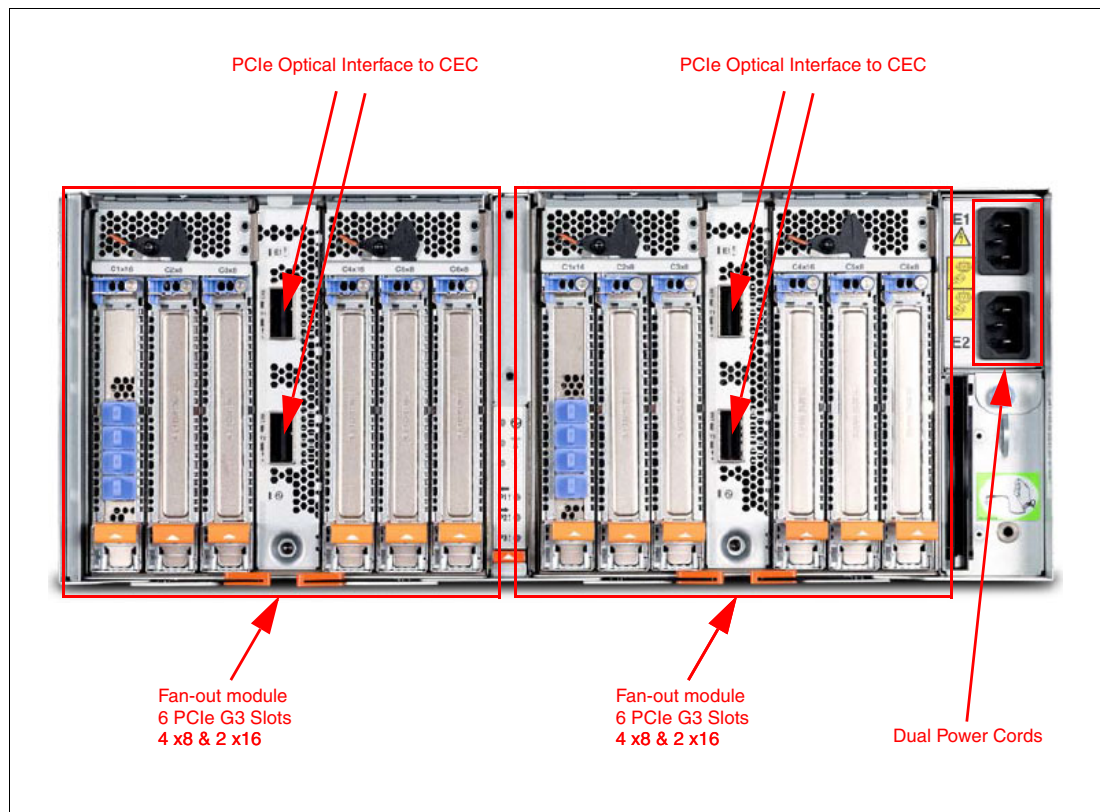


Figure 2-15 Rear view of the PCIe3 I/O expansion drawer

2.7.2 PCIe3 I/O expansion drawer optical cabling

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

- ▶ 3M Optical Cable Pair for PCIe3 Expansion Drawer (#ECC7)
- ▶ 10M Optical Cable Pair for PCIe3 Expansion Drawer (#ECC8)
- ▶ 3M Copper CXP Cable Pair for PCIe3 Expansion Drawer (#ECCS)

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

Limitation: You cannot mix copper and optical cables on the same PCIe3 I/O drawer. Both fan-out modules use copper cables or both use optical cables.

A minimum of one PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer is required to connect to the PCIe3 6-slot fan-out module in the I/O expansion drawer. The fan-out module has two CXP ports. The top CXP port of the fan-out module is cabled to the top CXP port of the PCIe3 Optical Cable Adapter. The bottom CXP port of the fan-out module is cabled to the bottom CXP port of the same PCIe3 Optical Cable Adapter.

To set up the cabling correctly, complete the following steps:

1. Connect an optical cable or copper CXP cable to connector T1 on the PCIe3 optical cable adapter in your server.
2. Connect the other end of the optical cable or copper CXP cable to connector T1 on one of the PCIe3 6-slot fan-out modules in your expansion drawer.
3. Connect another cable to connector T2 on the PCIe3 optical cable adapter in your server.
4. Connect the other end of the cable to connector T2 on the PCIe3 6-slot fan-out module in your expansion drawer.
5. Repeat steps 1 - 4 for the other PCIe3 6-slot fan-out module in the expansion drawer, if required.

Drawer connections: Each fan-out module in a PCIe3 Expansion Drawer can be connected only to a single PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer.

Figure 2-16 shows the connector locations for the PCIe3 I/O expansion drawer.

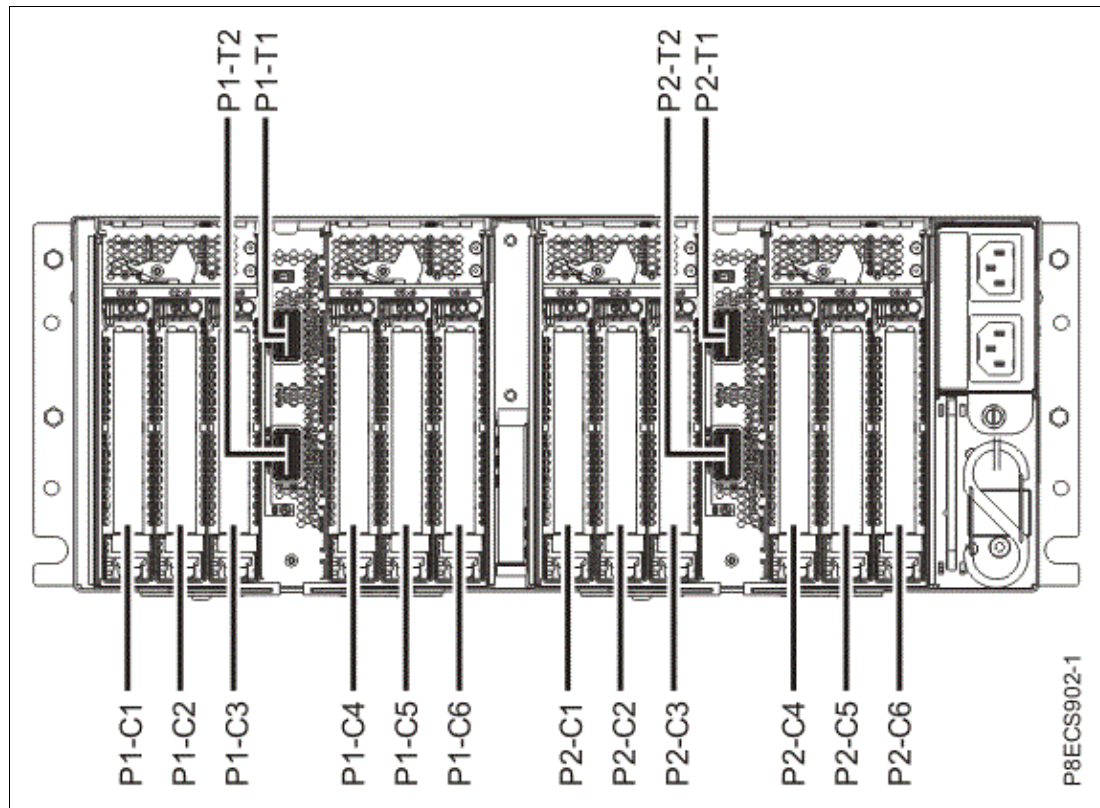


Figure 2-16 Connector locations for the PCIe3 I/O expansion drawer

Figure 2-17 shows typical optical cable connections.

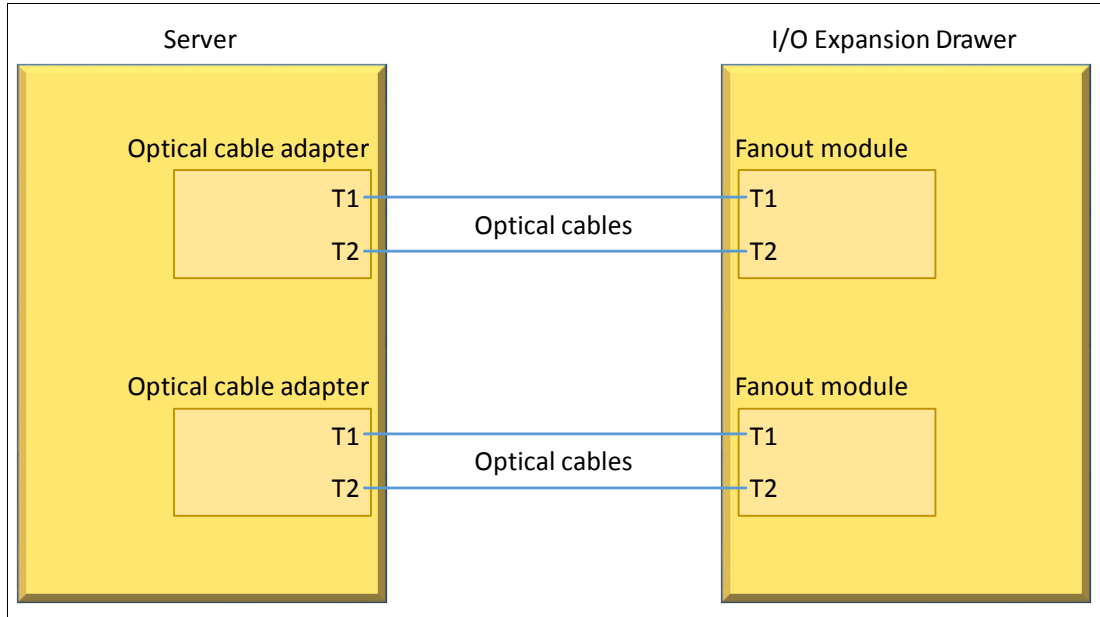


Figure 2-17 Typical optical cable connection

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

The PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ05) is supported in slots P1-C4 and P1-C9 for the Power L922 server. This is a double-wide adapter that requires two adjacent slots. If #EJ05 is installed in this slot, the external SAS port is not allowed in the system.

Table 2-21 shows PCIe adapter slot priorities and the maximum adapters that are supported in the Power L922 system.

Table 2-21 PCIe adapter slot priorities and maximum adapters that are supported

System	Feature code	Slot priorities	Maximum number of adapters that are supported
Power L922 (One processor)	#EJ05	9	1
Power L22 (Two processors)	#EJ05	9/10, 3/4	2

2.7.3 PCIe3 I/O expansion drawer system power control network cabling

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

2.8 External disk subsystems

This section describes the following external disk subsystems that can be attached to the Power L922 server:

- ▶ EXP24SX SAS Storage Enclosure (#ELLS) and EXP12SX SAS Storage Enclosure (#ELLL)
- ▶ IBM Storage

2.8.1 EXP24SX SAS Storage Enclosure and EXP12SX SAS Storage Enclosure

The EXP24SX is a storage expansion enclosure with twenty-four 2.5-inch SFF SAS bays. It supports up to 24 hot-swap HDDs or SSDs in 2 EIA of space in a 19-inch rack. The EXP24SX SFF bays use SFF Gen2 (SFF-2) carriers/ trays that are identical to the carrier/trays in the previous EXP24S Drawer. With Linux/VIOS, the EXP24SX can be ordered with four sets of six bays (mode 4), two sets of 12 bays (mode 2, or one set of 24 bays (mode 1).

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 could hold SSDs and one set could hold HDDs, but you cannot mix SSDs and HDDs in the same set of 12 bays.

The EXP12SX is a storage expansion enclosure with twelve 3.5-inch large form factor (LFF) SAS bays. It supports up to 12 hot-swap HDDs in 2 EIA of space in a 19-inch rack. The EXP12SX SFF bays use LFF Gen1 (LFF-1) carriers/trays. 4-KB sector drives (4096 or 4224) are supported. With Linux/VIOS, the EXP12SX can be ordered with four sets of three bays (mode 4), two sets of six bays (mode 2) or one set of 12 bays (mode 1). Only 4-KB sector drives are supported in the EXP12SX drawer.

Four mini-SAS HD ports on the EXP24SX or EXP12SX are attached to PCIe3 SAS adapters or attached to an integrated SAS controller in the Power L922 server. The following PCIe3 SAS adapters support the EXP24SX and EXP 12SX:

- ▶ PCIe3 RAID SAS Adapter Quad-port 6 Gb x8 (#EJ0J, #EJ0M, #EL3B, or #EL59)
- ▶ PCIe3 12 GB Cache RAID Plus SAS Adapter Quad-port 6 Gb x8 (#EJ14)

Earlier generation PCIe2 or PCIe1 SAS adapters are not supported by the EXP24SX.

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 cable options are:

- ▶ X12 cable: 3-meter copper (#ECDJ)
- ▶ YO12 cables: 1.5-meter copper (#ECDT), 3-meter copper (#ECDU)
- ▶ 3M 100 GbE Optical Cable QSFP28 (AOC) (#EB5R)
- ▶ 5M 100 GbE Optical Cable QSFP28 (AOC) (#EB5S)
- ▶ 10M 100 GbE Optical Cable QSFP28 (AOC) (#EB5T)
- ▶ 15M 100 GbE Optical Cable QSFP28 (AOC) (#EB5U)
- ▶ 20M 100 GbE Optical Cable QSFP28 (AOC) (#EB5V)
- ▶ 30M 100 GbE Optical Cable QSFP28 (AOC) (#EB5W)
- ▶ 50M 100 GbE Optical Cable QSFP28 (AOC) (#EB5X)
- ▶ 100M 100 GbE Optical Cable QSFP28 (AOC) (#EB5Y)

There are six SAS connectors on the rear of the EXP24SX and EXP12SX to which to SAS adapters or controllers are attached. They are labeled T1, T2, and T3; there are two T1, two T2, and two T3 connectors.

- ▶ 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 T2s and two T3s, to access all SAS bays.

Figure 2-18 shows the connector locations for the EXP24SX and EXP12SX storage enclosures.

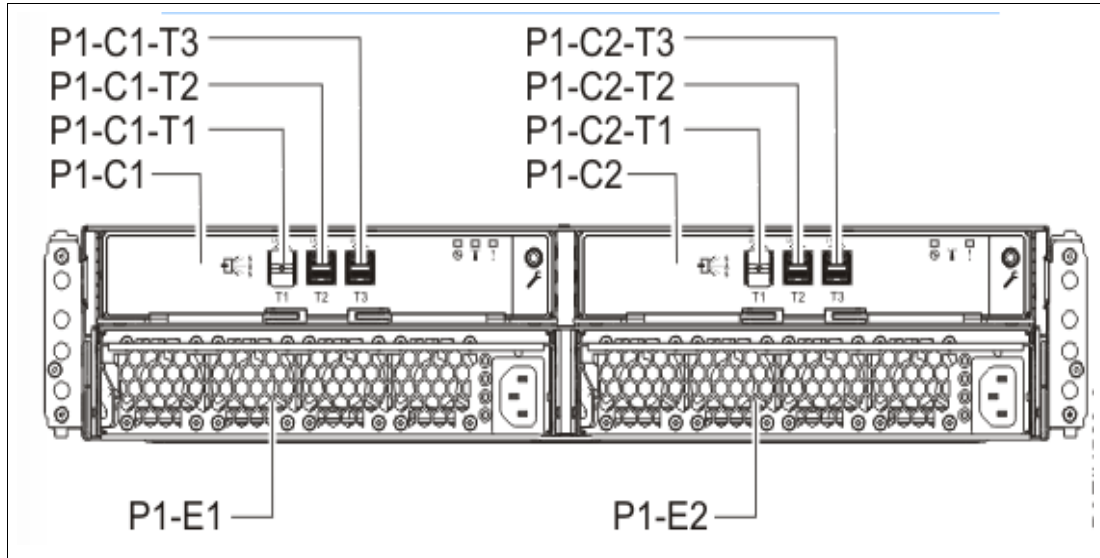


Figure 2-18 Connector locations for the EXP24SX and EXP12SX storage enclosures

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

The EXP24SX and EXP12SX drawers 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

2.8.2 IBM Storage

The IBM Storage 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 [Data Storage Solutions](#).

The following section highlights a few of the offerings.

IBM Flash Storage

The next generation of IBM Flash Storage delivers the extreme performance and efficiency that you need to succeed, with a new pay-as-you-go option to reduce your costs and scale-on-demand. For more information, see [Flash Storage and All Flash Arrays](#).

IBM DS8880 hybrid storage

IBM DS8880 Hybrid Storage is a family of storage systems that includes IBM DS8886 for high-performance functionality in a dense, expandable package, and IBM DS8884 to provide advanced functionality for consolidated systems or multiple platforms in a space-saving design. IBM DS8880 systems combine resiliency and intelligent flash performance to deliver microsecond application response times and more than six-nines availability. For more information, see [IBM DS8880 hybrid storage](#).

IBM XIV Storage System

IBM XIV® Gen3 is a high-end, grid-scale storage system that excels in tuning-free consistent performance, extreme ease of use, and exceptional data economics, including inline, field-proven IBM Real-time Compression™. IBM XIV is ideal for hybrid cloud, offering predictable service levels for dynamic workloads, simplified scale management (including in multi-tenant environments), flexible consumption models, and robust cloud automation and orchestration through OpenStack, the RESTful API, and VMware. It offers security and data protection through hot encryption, advanced mirroring and self-healing, and investment protection with perpetual licensing. For more information, see [IBM XIV Storage System](#).

IBM Storwize V7000

IBM Storwize® V7000 is an enterprise-class storage solution that offers the advantages of IBM Spectrum™ Virtualize software. It can help you lower capital and operational storage costs with heterogeneous data services while optimizing performance with flash storage. IBM Storwize V7000 enables you to take advantage of hybrid cloud technology without replacing your current storage. For more information, see [IBM Storwize V7000](#).

IBM Storwize V5000

IBM Storwize V5000 is a flexible storage solution that offers extraordinary scalability from the smallest to the largest system without disruption. Built with IBM Spectrum Virtualize™ software, it can help you lower capital and operational storage costs with heterogeneous data services. Storwize V5000 is an easily customizable and upgradeable solution for better investment protection, improved performance, and enhanced efficiency. For more information, see [IBM Storwize V5000](#).

2.9 Operating system support

The Power L922 servers support the Linux operating system.

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

2.9.1 Linux operating system

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

The supported versions of Linux on the Power L922 servers are as follows:

- ▶ If you are installing a Linux operating system logical partition (LPAR):
 - Red Hat Enterprise Linux 7 for Power LE, version 7.4, or later
 - SUSE Linux Enterprise Server 12 Service Pack 3, or later
 - Ubuntu Server 16.04.4, or later
 - SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 11 Service Pack 4
- ▶ If you are installing the Linux operating systems LPAR in nonproduction SAP implementations:
 - SUSE Linux Enterprise Server 12 Service Pack 3, or later
 - SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 12 Service Pack 3, or later
 - Red Hat Enterprise Linux 7 for Power LE, version 7.4, or later
 - Red Hat Enterprise Linux for SAP with Red Hat Enterprise Linux 7 for Power LE version 7.4, 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 the recommended packages for your Red Hat, SUSE Linux, or Fedora distribution. The packages are available from [Service and productivity tools for Linux on Power servers](#).

To learn about developing on the IBM Power Architecture®, find packages, get access to cloud resources, and discover tools and technologies, see 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 allows users to take leading-edge advantage of the IBM POWER hardware features on Linux.

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](#).

2.10 POWER9 reliability, availability, and serviceability capabilities by operating system

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

The elements of RAS can be described as follows:

- Reliability** Indicates how infrequently a defect or fault in a server occurs.
- Availability** Indicates how infrequently the functioning of a system or application is impacted by a fault or defect.
- Serviceability** Indicates how well faults and their effects are communicated to system managers, and how efficiently and nondisruptively the faults are repaired.

Table 2-22 provides a list of the Power Systems RAS capabilities by operating system. The HMC is an optional feature on scale-out Power Systems servers.

Table 2-22 Selected RAS features by operating system

RAS feature	Linux
Processor	
First failure data capture (FFDC) for fault detection/error isolation	X
Dynamic Processor Deallocation	X
I/O subsystem	
PCI Express bus enhanced error detection	X
PCI Express bus enhanced error recovery	X
PCI Express card hot-swap	X
Memory availability	
Memory Page Deallocation	X
Special Uncorrectable Error Handling	X
Fault detection and isolation	
Storage Protection Keys	Not used by OS
Error log analysis	X
Serviceability	
Boot-time progress indicators	X
Firmware error codes	X
Operating system error codes	X
Inventory collection	X
Environmental and power warnings	X
Hot-swap DASD / media	X
Dual disk controllers / Split backplane	X

RAS feature	Linux
EED collection	X
SP/OS "Call Home" on non-HMC configurations	X
IO adapter/device stand-alone diagnostic tests with PowerVM	X
SP mutual surveillance with IBM POWER Hypervisor™	X
Dynamic firmware update with HMC	X
Service Agent Call Home Application	X
Service Indicator LED support	X
System dump for memory, POWER Hypervisor, and SP	X
IBM Knowledge Center / IBM Systems Support Site service publications	X
System Service/Support Education	X
Operating system error reporting to HMC Service Focal Point (SFP) application	X
RMC secure error transmission subsystem	X
Healthcheck scheduled operations with HMC	X
Operator panel (real or virtual [HMC])	X
Concurrent Op Panel Display Maintenance	X
Redundant HMCs	X
High availability clustering support	X
Repair and Verify Guided Maintenance with HMC	X
PowerVM Live Partition / Live Application Mobility With PowerVM Enterprise Edition	X
EPOW	
EPOW errors handling	X



Virtualization

Virtualization is a key factor for productive and efficient use of IBM Power Systems servers. In this chapter, you can find a brief description of virtualization technologies that are available for POWER9. 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 POWER Hypervisor

IBM 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 your 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 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 technology, virtualized processors, IEEE VLAN-compatible virtual switches, 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 layer between the physical hardware resources and the logical partitions (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 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 and whether it is connected to the managed console. It requires memory to support the resource assignment to the LPARs on the server. The amount of memory that is required by the POWER Hypervisor firmware varies according to several factors:

- ▶ Memory is required for hardware page tables (HPTs).
- ▶ Memory is required to support I/O devices.
- ▶ Memory is required 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 Virtual I/O Server (VIOS) and Linux partitions of the maximum memory size of the partition. VIOS and Linux use larger page sizes (16 KB and 64 KB) instead of using 4 KB pages. Using larger page sizes reduces the overall number of pages that must be tracked so that the overall size of the HPT can be reduced. As an example, for an Linux 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 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 of the **chsyscfg** Hardware Management Console (HMC) command can be issued to define the HPT ratio to be 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 may incur more CPU usage 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 Entries (TCEs), which provide an information path between I/O devices and partitions. The TCEs provide the address of the I/O buffer, an indication 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 TCE entries 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.

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 are 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 usage by the POWER hypervisor

The IBM System Planning Tool (SPT) is a resource that you can use to estimate the amount of hypervisor memory that is required for a specific server configuration. After the SPT executable file is downloaded and installed, a configuration can be defined by selecting the appropriate hardware platform, selecting installed processors and memory, and defining partitions and partition attributes. Given a configuration, the SPT can estimate the amount of memory that will be assigned to the hypervisor, which helps 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

3.1.1 Virtual SCSI

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

3.1.2 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).

3.1.3 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.

3.1.4 Virtual (TTY) console

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

3.2 POWER processor modes

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

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

- ▶ 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 available processor modes on a POWER9 processor-based machine.

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

Processing mode

Dedicated
 Shared

Processing units

Total managed system processing units : 16.00
 Minimum shared processing units :
 Desired shared processing units :
 Maximum shared processing units :
 Shared processor pool:

Virtual processors

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

Sharing mode

Uncapped Weight :
 Processor compatibility mode:

OK Cancel Help

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 other generation must be activated in a specific compatibility mode.

Table 3-1 shows an example about which processor mode must be selected when 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 wanted processor mode is not supported on the destination.		Fails because the wanted proc mode is not supported on the destination.	
POWER9	POWER8	Fails because wanted processor mode is not supported on the destination.		Fails because wanted processor mode is not supported on the destination.	
Default	POWER9	Fails because wanted processor mode is not supported on the destination.		Default	POWER8
POWER8	POWER8	POWER8	POWER8	POWER8	POWER8
Default	POWER8	Default	POWER8	Default	POWER8
POWER7	POWER7	POWER7	POWER7	POWER7	POWER7

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 the need for 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: The target has 60 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) Service Focal Point (SFP) + Copper and RJ4 Adapter (#EN0K)

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 IBM 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: IBM PowerVM (#EC22) is required for all activated processors.

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 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 allocate to the LPAR. An LPAR is the equivalent of a virtual machine (VM).

You can assign processors, memory, and input/output devices to LPARs. You can run 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 feature code (FC). 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 0.05processing units. This number represents 0.05 of a physical core. Each physical core can be shared by up to 20 shared processor partitions, and the partition's entitlement can be incremented fractionally by as little as 0.01 of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC.

The Power L922 server supports up to 24 cores in a single system. Here are the maximum numbers:

- ▶ 24 dedicated partitions
- ▶ 480 micropartitions (maximum of 20 micropartitions per physical active core)

An important point is that 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 technology-based servers, you can configure dedicated partitions to become processor donors for idle processors that they own, allowing for the donation of spare CPU cycles from dedicated processor partitions to 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

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 the processor capacity is managed according to user-defined attributes.

If the Power Systems server is under heavy load, each micropartition within an SPP is ensured its processor entitlement plus any capacity that it might be allocated from the reserved pool capacity if the micropartition is uncapped.

If certain micropartitions in an SPP do not use their capacity entitlement, the unused capacity is ceded, and other uncapped micropartitions within the same SPP are allocated the extra 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 MSPP 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 a 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 adapter, 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.

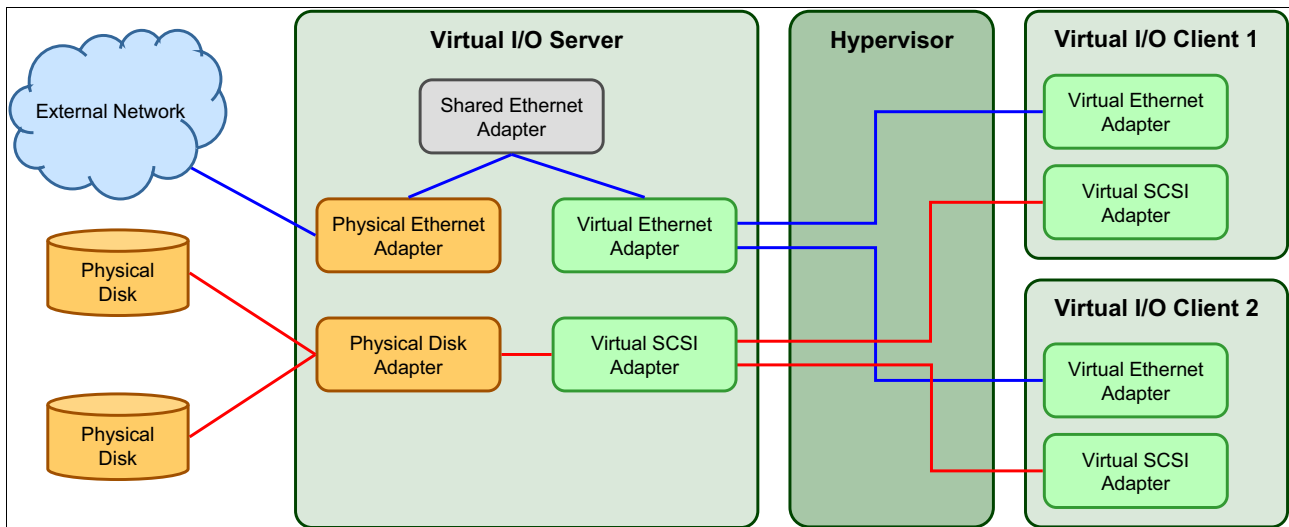


Figure 3-2 Architectural view of the VIOS

As a preferred practice, 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 with the VLANs on the external switches. Because the SEA processes packets at Layer 2, the original MAC address and VLAN tags of the packet are visible to other systems on the physical network. IEEE 802.1 VLAN tagging is supported.

The SEA also provides the ability for several client partitions to share one physical adapter. With an SEA, 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 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 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 (virtual SCSI server adapter and 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.

LPM provides systems management flexibility and improves system availability through the following functions:

- ▶ Avoid planned outages for hardware upgrade or firmware maintenance.
- ▶ Avoid 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 NPIV, 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 an IBM Power System 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 the 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 the 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 in memory after each partition has its own operating system, and some of them might even share the same kinds of applications. On heavily loaded systems, this behavior might lead to a shortage of the available memory resources, forcing paging by the Active Memory Sharing partition operating systems, the Active Memory Deduplication pool, or both, which might decrease overall system performance.

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

Active Memory Deduplication depends on the Active Memory Sharing feature being available, and it 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. In the event of 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 action can be done faster than restarting the server that failed 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 hosts the partition.

HMC V9R1 brings 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 powering 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 by using the REST API.

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 Introduction and Technical Overview*, REDP-5494
- ▶ *IBM Power System S822LC for High Performance Computing Introduction and Technical Overview*, REDP-5405
- ▶ *IBM Power Systems H922 and H924 Introduction and Technical Overview*, REDP-5498
- ▶ *IBM Power Systems LC921 and LC922 Introduction and Technical Overview*, REDP-5495
- ▶ *IBM Power Systems S922, S914, and S924 Introduction and Technical Overview*, 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 Portal for OpenPOWER - POWER9 Monza Module
https://www.ibm.com/systems/power/openpower/tgcmDocumentRepository.xhtml?aliasId=POWER9_Monza
- ▶ IBM Fix Central website
<http://www.ibm.com/support/fixcentral/>
- ▶ IBM Knowledge Center
<http://www.ibm.com/support/knowledgecenter/>
- ▶ IBM Power Systems
<http://www.ibm.com/systems/power/>
- ▶ IBM Power Systems Hardware: IBM Knowledge Center
<https://www.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm>

- ▶ IBM Storage website
<http://www.ibm.com/systems/storage/>
- ▶ IBM System Planning Tool (SPT) website
<http://www.ibm.com/systems/support/tools/systemplanningtool/>
- ▶ IBM Systems Energy Estimator
<http://www-912.ibm.com/see/EnergyEstimator/>
- ▶ Migration combinations of processor compatibility modes for active Partition Mobility
<https://www.ibm.com/support/knowledgecenter/POWER7/p7hc3/iphc3pcmcombosact.htm>
- ▶ 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/0verview?brandind=Hardware~Systems~Power>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



REDP-5496-00

ISBN 0738456985

Printed in U.S.A.

Get connected

