

IBM Power System AC922

Technical Overview and Introduction

Ritesh Nohria

Gustavo Santos

Volker Haug



 Analytics

Power Systems



International Technical Support Organization

IBM Power System AC922: Technical Overview and Introduction

July 2018

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

First Edition (July 2018)

This edition applies to the IBM Power System AC922, machine type and model number 8335-GTH and 8335-GTX.

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

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

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

Contents

Notices	v
Trademarks	vi
Preface	vii
Authors	vii
Now you can become a published author, too!	viii
Comments welcome	viii
Stay connected to IBM Redbooks	x
Chapter 1. Product summary	1
1.1 Key server features	2
1.2 Server models	4
1.2.1 Power AC922 server model GTH	4
1.2.2 Power AC922 server model GTX	6
1.2.3 Minimum features	8
Chapter 2. System architecture	9
2.1 Power AC922 server architecture	10
2.2 Processor subsystem	15
2.2.1 POWER9 processor overview	15
2.2.2 Processor features	16
2.2.3 Supported technologies	18
2.2.4 Processor feature codes	19
2.3 Memory subsystem	19
2.3.1 Memory feature codes and placement rules	20
2.3.2 Memory bandwidth	21
2.4 I/O subsystem	22
2.4.1 PCI Express Controller	22
2.4.2 IBM CAPI2	23
2.4.3 OpenCAPI	24
2.4.4 The NVIDIA Tesla V100	27
2.4.5 NVLink 2.0	30
2.5 PCI adapters	31
2.5.1 Slot configuration	32
2.5.2 Local area network adapters	33
2.5.3 Fibre Channel adapters	34
2.5.4 CAPI-enabled adapters	34
2.5.5 Compute-intensive accelerators	34
2.5.6 Flash storage adapters	35
2.6 System ports	36
2.7 Internal disks	36
2.7.1 Disk and media features	37
2.8 External I/O subsystems	37
2.9 Location codes	38
2.10 IBM System Storage	39
2.10.1 IBM Flash Storage	39
2.10.2 Software-defined storage	39
2.10.3 Hybrid storage	39
2.10.4 Storage area network	39

2.11 Operating system support	39
2.11.1 Ubuntu	39
2.11.2 Red Hat Enterprise Linux	40
2.12 Java	40
2.13 Reliability, availability, and serviceability	40
2.13.1 Error handling	41
2.13.2 Serviceability	42
Chapter 3. Physical infrastructure	45
3.1 Operating environment	46
3.1.1 Leak detection	48
3.1.2 Water pressure	49
3.2 Physical package	49
3.3 System power	49
3.4 System cooling	50
3.5 Rack specifications	54
3.5.1 IBM Enterprise Slim Rack 7965-S42	55
3.5.2 AC power distribution units	59
3.5.3 Rack-mounting rules	62
3.5.4 Original equipment manufacturer racks	62
Appendix A. IBM PowerAI	67
Deep learning frameworks requirements	68
IBM PowerAI Vision	68
Distributed Deep Learning	69
Large model support	70
Software download	71
IBM PowerAI Enterprise	71
PowerAI Enterprise features	72
Distributed training with IBM Fabric technology and Elastic Deep Learning	72
Monitoring and optimization with deep learning insights	73
Hyperparameter tuning	73
IBM Spectrum Conductor Deep Learning Impact	73
Importing, preparing, and transforming data faster	74
Automation to help with optimization and training	75
Dramatically reduce training times	75
Deploying, inferring, scoring, capturing organizational value, and reiterating	76
High-performance computing	76
Related publications	79
IBM Redbooks	79
Online resources	79
Help from IBM	81

Notices

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

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

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

AIX®
Easy Tier®
IBM®
IBM Cloud™
IBM Spectrum®
IBM Spectrum Conductor®

Interconnect®
LSF®
OpenCAPI™
POWER®
Power Architecture®
POWER9™

PowerHA®
PowerVM®
Redbooks®
Redbooks (logo) ®
Storwize®
System Storage™

The following terms are trademarks of other companies:

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

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

Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States 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 that covers the IBM Power System AC922 server (8335-GTH and 8335-GTX models). The Power AC922 server is the next generation of the IBM POWER® processor-based systems, which are designed for deep learning (DL) and artificial intelligence (AI), high-performance analytics, and high-performance computing (HPC).

This paper introduces the major innovative Power AC922 server features and their relevant functions:

- ▶ Different powerful IBM POWER9™ processors that offer up to 44 cores at up to 2.80 GHz (3.10 GHz turbo) performance with up to 2 TB of memory.
- ▶ IBM Coherent Accelerator Processor Interface (CAPI) 2.0, IBM OpenCAPI™, and second-generation NVIDIA NVLink 2.0 technology for exceptional processor to accelerator intercommunication.
- ▶ Up to six Tesla V100 with NVLink GPUs with 16 GB and 32 GB GPU memory.

This publication is for professionals who want to acquire a better understanding of IBM Power Systems products and is intended for the following audiences:

- ▶ Clients
- ▶ Sales and marketing professionals
- ▶ Technical support professionals
- ▶ IBM Business Partners
- ▶ Independent software vendors (ISVs)

This paper expands the set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power AC922 server.

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

Authors

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

Ritesh Nohria is a Client Technical Specialist for Power Systems products at IBM Systems, India. He has been with IBM since 2014. He has a total 23 years of experience in the field of IT. He holds a degree in Electronics and Controls from YMCA Institute of Engineering, Faridabad, India. He has served in various positions during his career, and provided support for IBM mainframes, Digital VAX, Control Data Cyber mainframe, SGI Origin systems, Cray Vector systems, and several distributions of Linux. His areas of expertise include HPC, modern data platforms, Internet of Things (IoT), and AI.

Gustavo Santos is an IBM Power Systems Consultant at IBM. He has been with IBM since 1997. He has 21 years of experience in Power Systems products and UNIX. He holds a degree in Systems Engineering from Universidad Abierta Interamericana. He also has 15 years of experience working in service delivery on IBM AIX®, Virtual I/O Server (VIOS), and Hardware Management Console (HMC) for multiple accounts in the United States and Latin America. During the last 4 years, he worked as a Power Systems Consultant deploying services and training courses. His areas of expertise include Power Systems products, AIX, VIOS, Live Partition Mobility (LPM), IBM GPFS, UNIX, and several IBM software products, such as IBM Cloud™ PowerVC Manager, IBM BigFix, the IBM Spectrum® family, and IBM System Mirror.

Volker Haug is an Executive IT Specialist & Open Group Distinguished IT Specialist in Ehningen, Germany. He holds a Diploma degree in Business Management from the University of Applied Studies in Stuttgart. His career includes more than 32 years of experience with Power Systems, AIX, and PowerVM® virtualization. He has written several IBM Redbooks® publications about Power Systems and PowerVM and is a Platinum Redbooks author. Volker is IBM POWER9 Champion and a member of the German Technical Expert Council, which is an affiliate of the IBM Academy of Technology.

The project that produced this publication was managed by:

Scott Vetter, PMP
IBM Austin

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

Ron Arroyo, Lokesh Bhatt, Matthew Butterbaugh, Daniel Henderson, Jeanine Hinck, Ray Laning, Chris Mann, Benjamin Mashak, Stephen Mroz, Thoi Nguyen, Kanisha Patel, Justin Thaler, Julie Villarrea
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>



Product summary

The IBM Power System AC922 is the next generation of the IBM POWER processor-based systems, which are designed for deep learning (DL) and artificial intelligence (AI), high-performance data analytics (HPDA), and high-performance computing (HPC).

The system is co-designed with OpenPOWER Foundation members and will be deployed at the most powerful supercomputer on the planet by a partnership between IBM, NVIDIA, Mellanox, and others. It provides the most current technologies that are available for HPC and improves the movement of data from memory to graphics processing units (GPUs) and back, which enables faster and lower latency data processing.

This massive computing capacity is packed into 2Us of rack space. There are special cooling systems to support the largest configurations: One is the air-cooled model GT, and the other is the water-cooled model GTX.

Among the new technologies that the system provides, the most significant are the following ones:

- ▶ Two IBM POWER9 processors with up to 40 cores for model GTH or 44 cores for model GTX, and improved buses.
- ▶ Up to 2 TB of DDR4 memory with improved speed.
- ▶ Up to six NVIDIA Tesla V100 (Volta) with NVLink GPUs, with 16 GB or 32 GB HMB2 memory.
- ▶ Second-generation NVLink technology with 2x throughput compared to the first generation.
- ▶ Four Peripheral Component Interconnect® Express (PCIe) x16 4.0 low-profile (LP) slots. Three are Coherent Accelerator Processor Interface (CAPI)-enabled for future CAPI-enabled devices (a maximum of three CAPI devices can be used concurrently).
- ▶ Two 2.5-inch SATA drives for a maximum of 4 TB hard disk drive (HDD) or 7.68 TB of solid-state drive (SSD) storage.
- ▶ Two integrated USB 3.0 ports.
- ▶ Two hot-swap and redundant power supplies: (maximum 4-GPU configuration) 2200 W 200 - 240 and 277 V AC.

Note: Servers with six GPUs configuration do not allow hot-swapping of the power supplies.

Figure 1-1 shows the front and rear views of the Power AC922 server with the water-cooling system (model GTX).

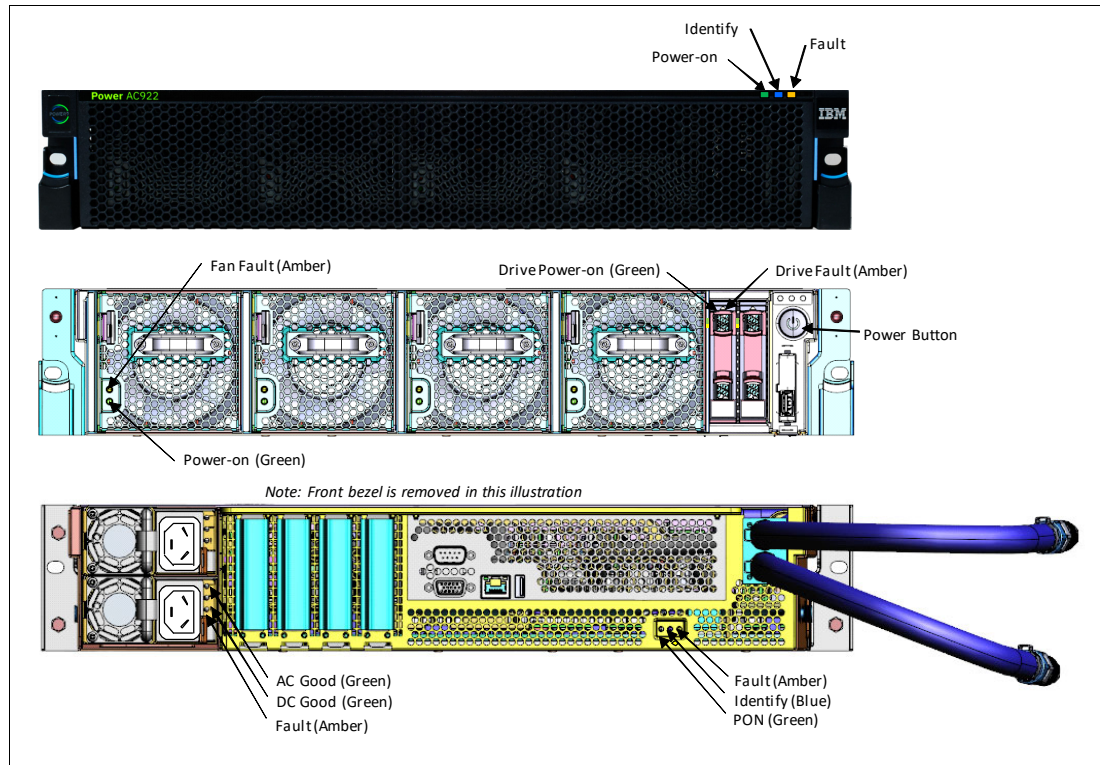


Figure 1-1 Front and rear views of the IBM Power AC922 server

1.1 Key server features

The Power AC922 server addresses the demanding needs of DL and AI, HPDA, and HPC.

An updated list of ported HPC applications that can use the IBM POWER technology is available at [IBM Power Systems HPC Applications Summary](#).

The system includes several features to improve performance:

- ▶ POWER9 processors:
 - Each POWER9 processor module has either 16, 18, 20, or 22 cores that are based on a 64-bit architecture:
 - The clock speeds for the 16-core chip are 2.7 GHz (3.3 GHz turbo) for the model GTH.
 - The clock speeds for the 20-core chip are 2.4 GHz (3.0 GHz turbo) for the model GTH.

- The clock speeds for the 18-core chip are 3.15 GHz (3.45 GHz turbo) for the model GTX.
- The clock speeds for the 22-core chip are 2.80 GHz (3.10 GHz turbo) for the model GTX.
- Up to four threads per core
- 512 KB of L2 cache and up to 10 MB NUCA of L3 cache that is shared by each pair of cores
- 108 GBps random and 170 GBps streaming peak memory bandwidth
- ▶ DDR4 memory:
 - Sixteen dual inline memory module (DIMM) memory slots
 - Maximum of 2048 GB DDR4 system memory
 - Increased clock speed from 1333 MHz to 2666 MHz for reduced latency
- ▶ NVIDIA Tesla V100 GPUs:
 - Up to six NVIDIA Tesla V100 GPUs, based on the NVIDIA SXM2 form factor connectors
 - 7.8 TFLOPs per GPU for double precision
 - 15.7 TFLOPs per GPU for single precision
 - 125 TFLOPs per GPU for DL, with new 640 Tensor Cores per GPU, which are designed for DL
 - 32 GB or 16 GB HBM2 internal memory with 900 GBps bandwidth, 1.5x the bandwidth compared to Pascal P100
 - Water cooling for six GPUs configurations to improve compute density
- ▶ NVLink 2.0:
 - Twice the throughput compared to the previous generation of NVLink
 - Up to 200 GBps of bidirectional bandwidth between GPUs
 - Up to 300 GBps of bidirectional bandwidth per POWER9 chip and GPUs, compared to 32 GBps of traditional PCIe Gen3
- ▶ OpenCAPI 3.0:
 - Open protocol bus to enable connections between the processor system bus in a high speed and cache coherent manner with OpenCAPI compatible devices, such as accelerators, network controllers, storage controllers, and advanced memory technologies
 - Up to 100 GBps of bidirectional bandwidth between CPUs and OpenCAPI devices
- ▶ Four PCIe Gen4 slots with up to 64 GBps bandwidth per slot, twice the throughput from PCIe Gen3, with three CAPI 2.0 capable slots
- ▶ An initial order of a 8335-GTH model server is a customer install system. A model upgrade from a model GTG to model GTH will be a CE installation. The 8335-GTX model is a customer install system.

Figure 1-2 shows the physical locations of the main server components.

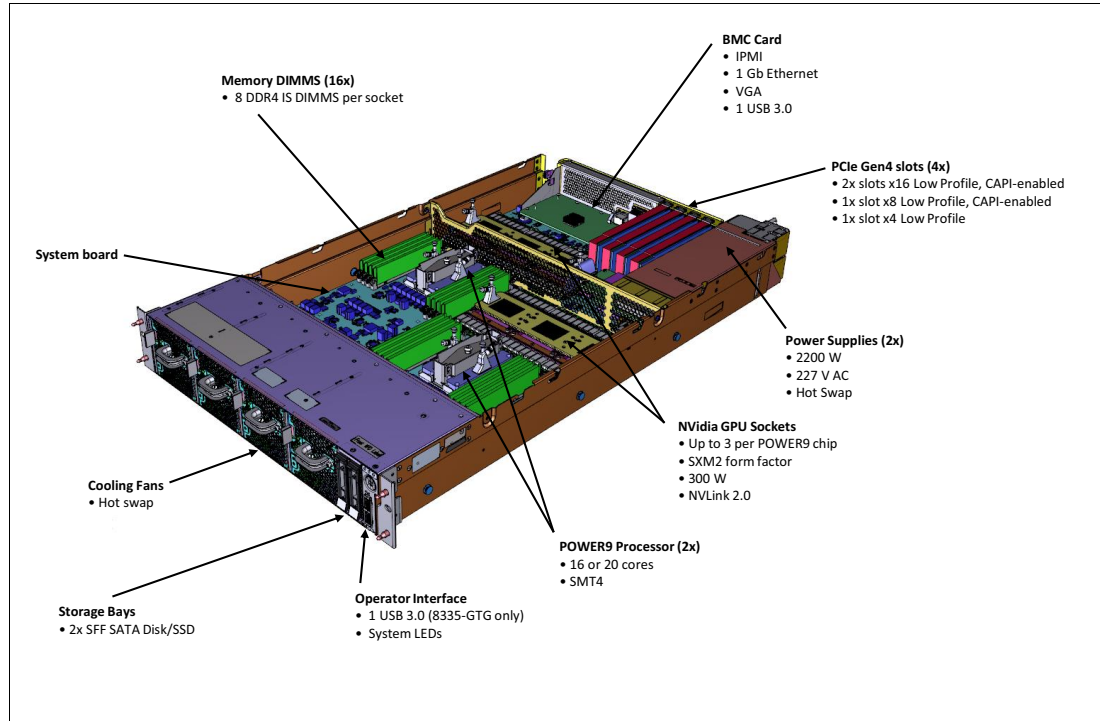


Figure 1-2 Location of server main components

1.2 Server models

The Power AC922 server is manufactured as two distinct models, as shown in Table 1-1.

Table 1-1 Summary of Power AC922 server available models

Server model	POWER9 chips	Maximum memory	Maximum GPU cards	Cooling
8335-GTH	2	2 TB	4	Air-cooled
8335-GTX	2	2 TB	6	Water-cooled

1.2.1 Power AC922 server model GTH

This summary describes the standard features of the Power AC922 model 8335-GTH server:

- ▶ 19-inch rack-mount (2U) chassis
- ▶ Two POWER9 processor modules:
 - 16-core 2.7 GHz (3.3 GHz turbo) processor module (#EP0P)
 - 20-core 2.4 GHz (3.0 GHz turbo) processor module (#EP0R)
 - Up to 2048 GB of 2666 MHz DDR4 error-correcting code (ECC) memory
- ▶ Two small form factor (SFF) bays for HDDs or SSDs that support:
 - Two 1 TB 7200 RPM NL SATA disk drives (#ELD0)
 - Two 2 TB 7200 RPM NL SATA disk drives (#ES6A)

- Two 960 GB SATA SSDs (#ELU4)
- Two 1.92 TB SATA SSDs (#ELU5)
- Two 3.84 TB SATA SSDs (#ELU6)
- ▶ Integrated SATA controller
- ▶ Four PCIe Gen4 slots:
 - Two PCIe x16 4.0 LP slots, CAPI-enabled
 - One PCIe x16 4.0 LP slot, CAPI-enabled or one PCIe x8 shared 4.0 slot, CAPI-enabled
 - One PCIe x4 4.0 LP slot
- ▶ NVIDIA Tesla V100 GPU options:
 - Zero, two, or four 16 GB SXM2 NVIDIA Tesla V100 GPUs with NVLink Air-Cooled (#EC4J)
 - Zero, two, or four 32 GB SMX2 NVIDIA Tesla V100 GPUs with NVLink Air-Cooled (#EJ4L)
- ▶ Integrated features:
 - IBM EnergyScale technology
 - Hot-swap and redundant cooling
 - Two 1 Gb RJ45 ports
 - One front USB 3.0 port for general use
 - One rear USB 3.0 port for general use
 - One system port with RJ45 connector
- ▶ Two power supplies (both are required.)

The internal view of the fully populated Power AC922 model 8335-GTH server with four GPUs is shown in Figure 1-3. In this figure, the air baffles were removed to better show the major components.

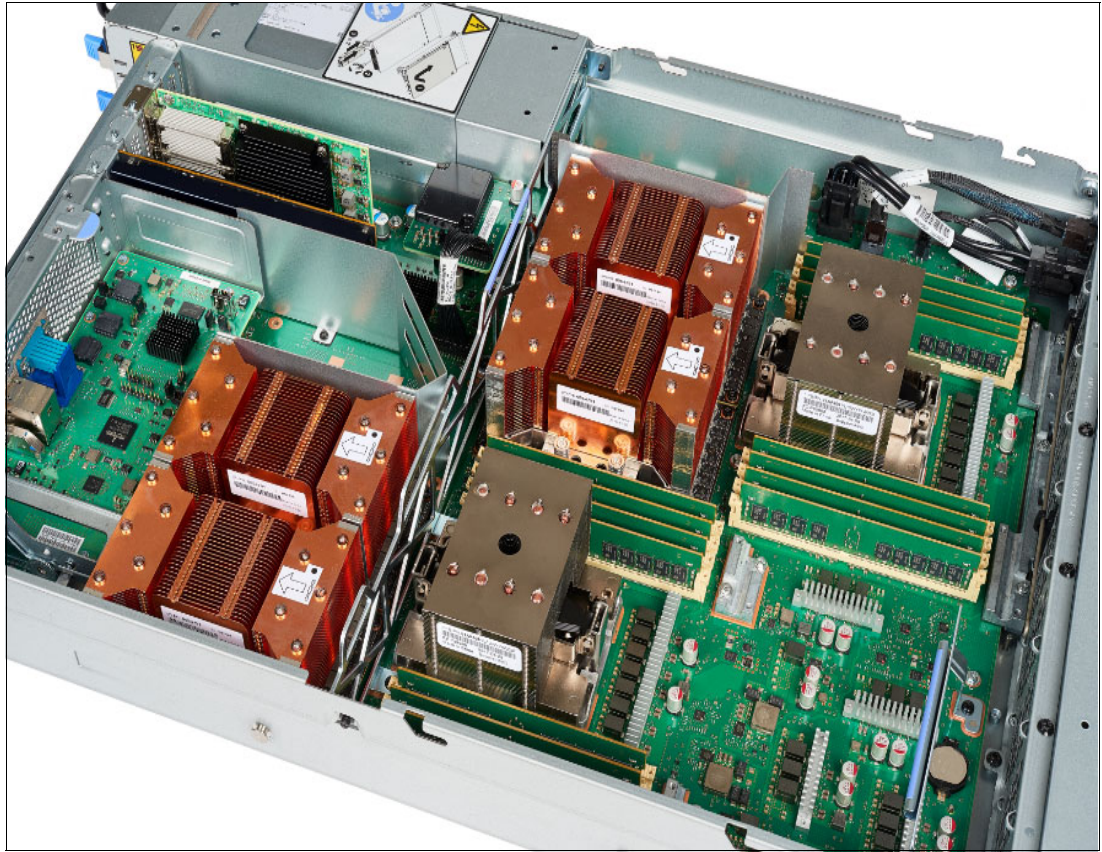


Figure 1-3 Power AC922 server model GTH fully populated with four GPUs

1.2.2 Power AC922 server model GTX

This summary describes the standard features of the Power AC922 model GTX server:

- ▶ 19-inch rack-mount (2U) chassis
- ▶ Two POWER9 processor modules:
 - 18-core 3.15 GHz (3.45 GHz turbo) processor module
 - 22-core 2.80 GHz (3.10 GHz turbo) processor module
 - Up to 2048 GB of 2666 MHz DDR4 ECC memory
- ▶ Two SFF bays for HDDs or SSD that support:
 - Two 1 TB 7200 RPM NL SATA disk drives (#ELD0)
 - Two 2 TB 7200 RPM NL SATA disk drives (#ES6A)
 - Two 960 GB SATA SSDs (#ELU4)
 - Two 1.92 TB SATA SSDs (#ELU5)
 - Two 3.84 TB SATA SSDs (#ELU6)
- ▶ Integrated SATA controller

- ▶ Four PCIe Gen4 slots:
 - Two PCIe x16 Gen4 LP slots, CAPI-enabled
 - One PCIe x8 Shared Gen4 LP slot, CAPI-enabled
 - One PCIe x4 Gen4 LP slot
- ▶ NVIDIA Tesla V100 GPU options:
 - Four or six 16 GB SXM2 NVIDIA Tesla V100 GPUs with NVLink Water-Cooled
 - Four or six 32 GB SMX2 NVIDIA Tesla V100 GPUs with NVLink Water-Cooled
- ▶ Integrated features:
 - IBM EnergyScale technology
 - Hot-swap and redundant cooling
 - Two 1 Gb RJ45 ports
 - One rear USB 3.0 port for general use
 - One system port with RJ45 connector
- ▶ Two power supplies (Both are required.)

The internal view of the fully populated Power AC922 model 8335-GTX server with six GPUs and the water-cooling system installed is shown in Figure 1-4. In this figure, the air baffles were removed to better show the major components.

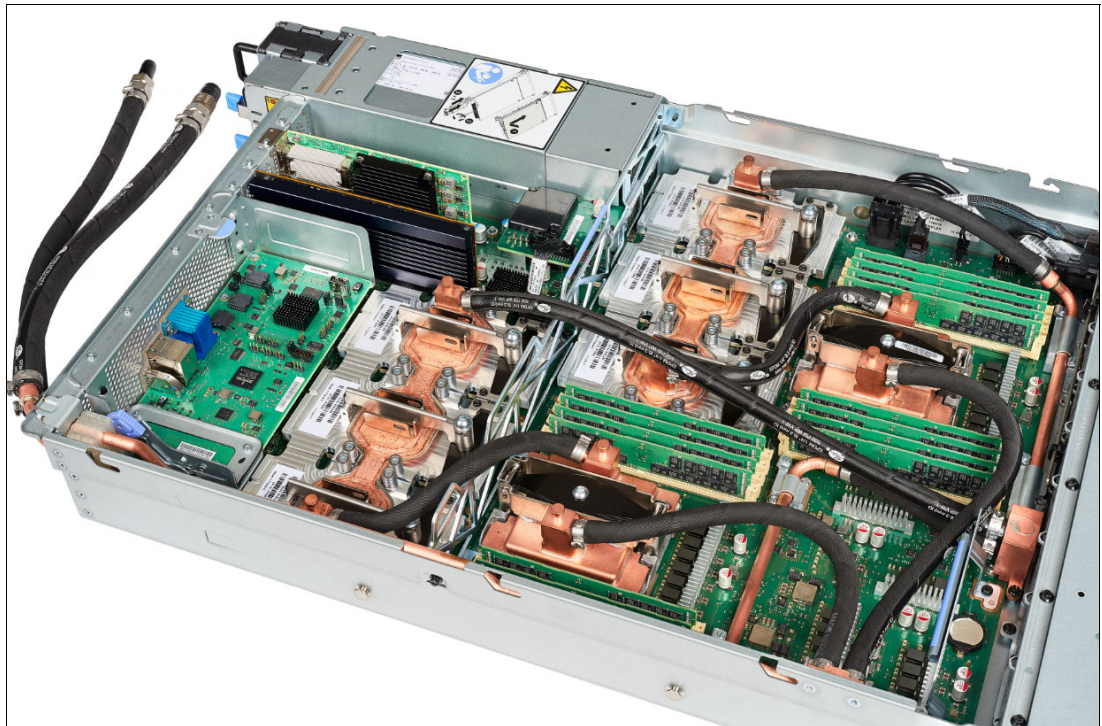


Figure 1-4 Power AC922 server model GTX fully populated with six GPUs

Note: A Hardware Management Console (HMC) is not supported on the Power AC922 servers.

1.2.3 Minimum features

The minimum initial order for the Power AC922 model 8335-GTH must include the following features:

- ▶ Two processor modules with at least 16 cores each
- ▶ 128 GB of memory (sixteen 8 GB memory DIMMs)
- ▶ Two power supplies and power cords (Both are required.)
- ▶ Rack-mounting hardware
- ▶ A Linux operating system (OS) indicator
- ▶ A rack integration indicator
- ▶ A Language Group Specify

The minimum initial order for the Power AC922 model 8335-GTX must include the following features:

- ▶ Two processor modules with at least 16 cores each
- ▶ 128 GB of memory (sixteen 8 GB memory DIMMs)
- ▶ Four NVIDIA Tesla V100 GPUs with NVLink Water-Cooled
- ▶ Two power supplies and power cords (Both are required.)
- ▶ Rack-mounting hardware
- ▶ A Linux OS indicator
- ▶ A rack integration indicator
- ▶ A Language Group Specify



System architecture

This chapter describes the overall system architecture for the IBM Power System AC922 server. The bandwidths that are provided throughout the section are theoretical maximums that are used for reference.

Note: The speeds that are shown are at an individual component level. Multiple components and application implementation are key to achieving the preferred performance. Always do the performance sizing at the application-workload environment level and evaluate performance by using real-world performance measurements and production workloads.

2.1 Power AC922 server architecture

The Power AC922 server is a two single-chip module (SCM) system. Each SCM is attached to eight memory registered DIMM (RDIMM) slots. The server has a maximum capacity of 16 memory dual inline memory modules (DIMMs) that enable up to 2 TB of memory.

The system board has sockets for four or six graphics processing units (GPUs) depending on the model, each of which is 300 watts capable. Additionally, the server has a total of four Peripheral Component Interconnect Express (PCIe) Gen4 slots, and three of these slots are Coherent Accelerator Processor Interface (CAPI)-capable.

Figure 2-1 shows the location of the processors, memory DIMMs, GPUs, and PCIe slots.



Figure 2-1 Component location for 4-GPU and 6-GPU system board

An integrated SATA controller is routed through a dedicated PCI bus on the main system board and enables up to two SATA hard disk drives (HDDs) or solid-state drives (SSDs) to be installed.

Figure 2-2 shows the location of the integrated SATA connector. This bus also drives the integrated Ethernet and USB ports.

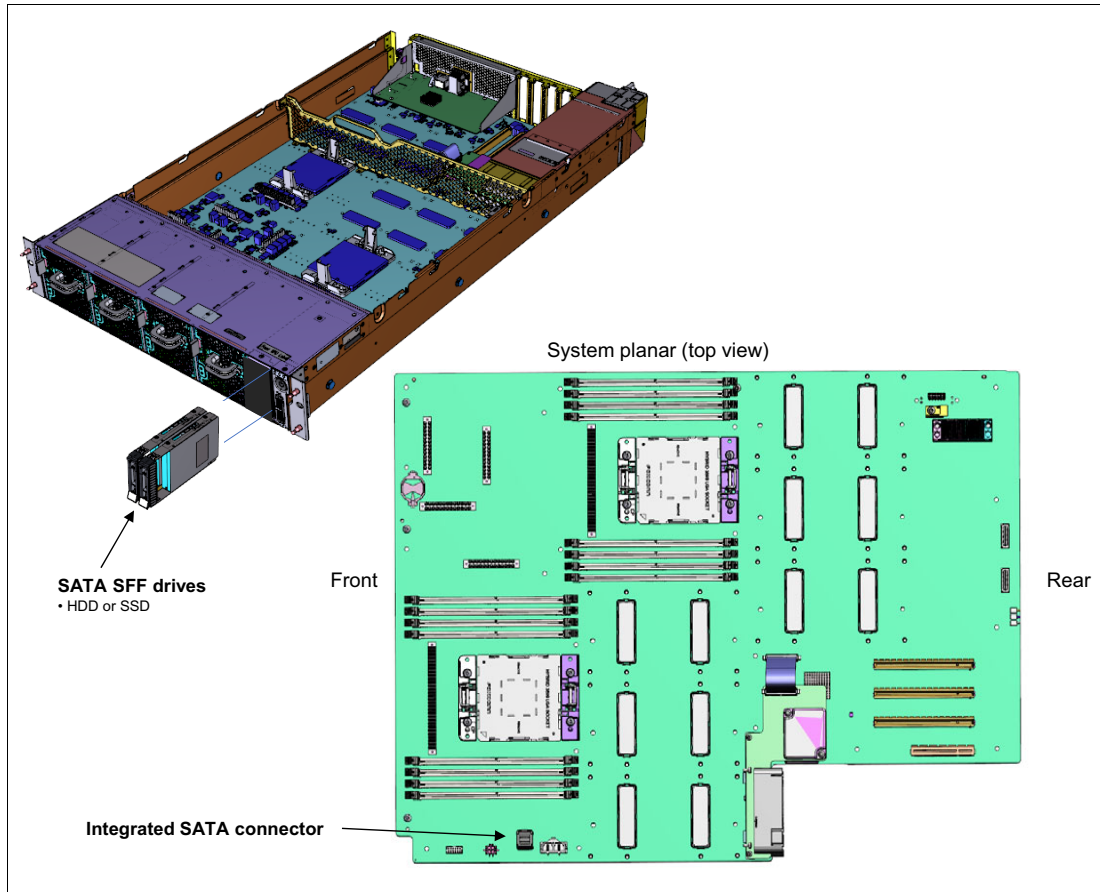


Figure 2-2 Integrated SATA connector

The POWER9 processor brings enhanced memory and I/O connectivity, improved chip to chip communication, and a new bus called NVLink 2.0.

Figure 2-3 shows the external processor connectivity.

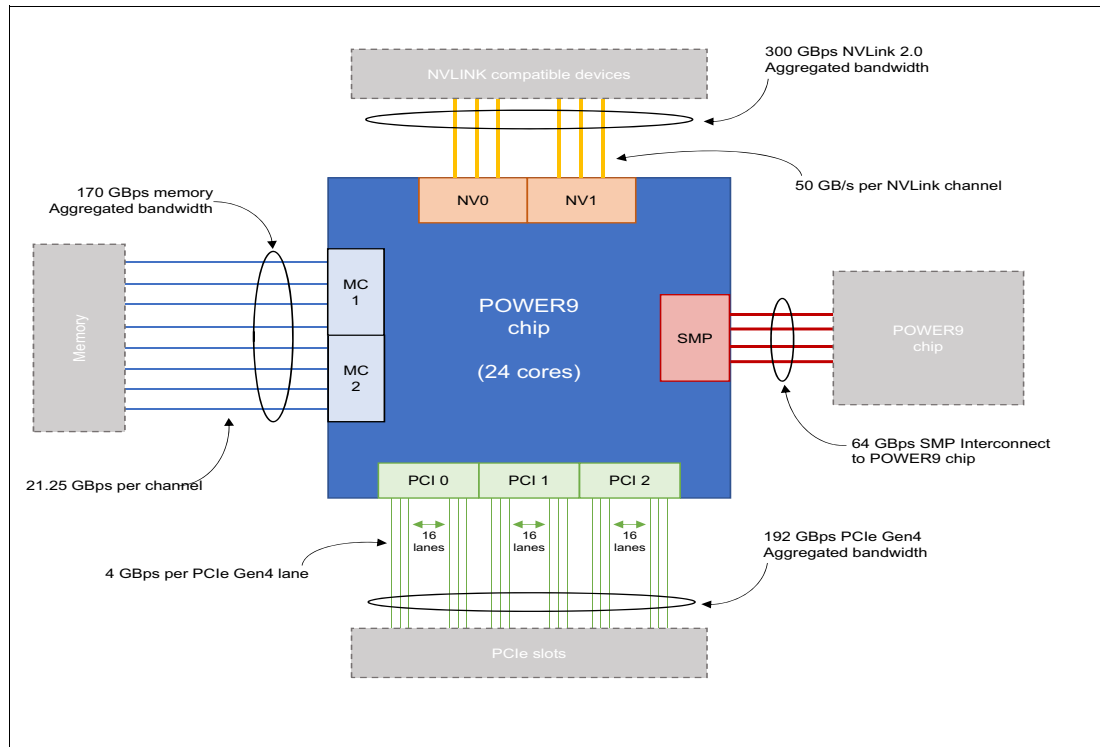


Figure 2-3 POWER9 chip external connectivity

Faster DDR4 memory DIMMs at 2666 MHz are connected to two memory controllers through eight channels with a total peak bandwidth of 170 GBps. Symmetric multiprocessing (SMP) chip-to-chip interconnect is done through a four-channel SMP bus with 64 GBps bidirectional bandwidth.

The current PCIe Gen4 interconnect doubles the channel bandwidth of the previous PCIe Gen3 generation, enabling the 48 PCIe channels with a total of 192 GBps bidirectional bandwidth between the I/O adapters and the POWER9 chip.

The connection between GPUs and between CPUs and GPUs is done through a link that is called NVLink 2.0, which was developed by IBM, NVIDIA, and the OpenPOWER Foundation. This link provides up to 5x more communication bandwidth between CPUs and GPUs (compared to traditional PCIe Gen3) and enables faster data transfer between memory and GPUs and between GPUs. Complex and data-hungry algorithms, such as the ones that are used in machine learning, can benefit from having these enlarged pipelines for data transfer because the amount of data that must be processed is many times larger than the GPU internal memory. For more information about NVLink 2.0, see 2.4.5, “NVLink 2.0” on page 30.

Each POWER9 CPU and each GPU has six NVLink 2.0 channels, called *NVLink Bricks*, with each one delivering up to 50 GBps bidirectional bandwidth. These channels can be aggregated to enable more bandwidth or more peer-to-peer connections.

Figure 2-4 on page 13 compares the POWER9 implementation of NVLink 2.0 with traditional processor chips that use PCIe and NVLink.

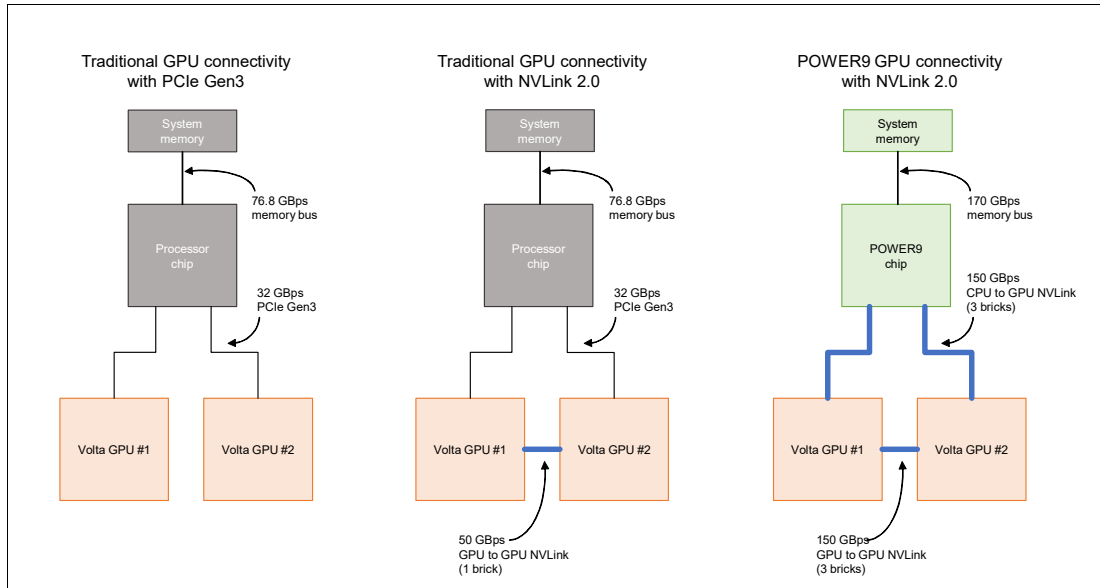


Figure 2-4 NVLink 2.0 POWER9 implementation versus traditional architectures

On traditional processors, communication is done through PCIe Gen3 buses. When the processor must handle all the GPU-to-GPU communication and GPU-to-system memory communication, having more than two GPUs per processor potentially creates a bottleneck on the data flow from system memory to GPUs.

To reduce this impact on the GPU-to-GPU communication, NVLink provides a 50 GBps direct link between GPUs, reducing the dependency on the PCIe bus to exchange data between GPUs, but still depending on PCIe Gen3 to GPU to system memory communications.

The NVLink 2.0 implementation on POWER9 goes beyond the traditional implementation by implementing 1.5x more memory bandwidth and aggregating NVLink Bricks to enable 3x faster communication between GPUs and system memory to GPU, reducing potential bottlenecks throughout the system. The goal is to move data from system memory to the GPU internal memory as fast as possible so that GPU processing does not have to stop and wait for data to be moved to continue processing.

To maximize the bandwidth, NVLink Bricks are combined differently depending on whether the server has four or six GPUs (with two POWER9 processors). There are two logical diagrams, which are based on the number of maximum GPUs that are supported in the system.

Figure 2-5 shows the logical system diagram for the Power AC922 server model GTH with four GPUs, where the six NVLink Bricks are divided into groups of three, which enable 150 GBps buses between GPUs.

NVLink wiring for the two GPU planar uses three NV bricks (x8 lane) from the CPU to the GPU and three NV bricks from each GPU to GPU.

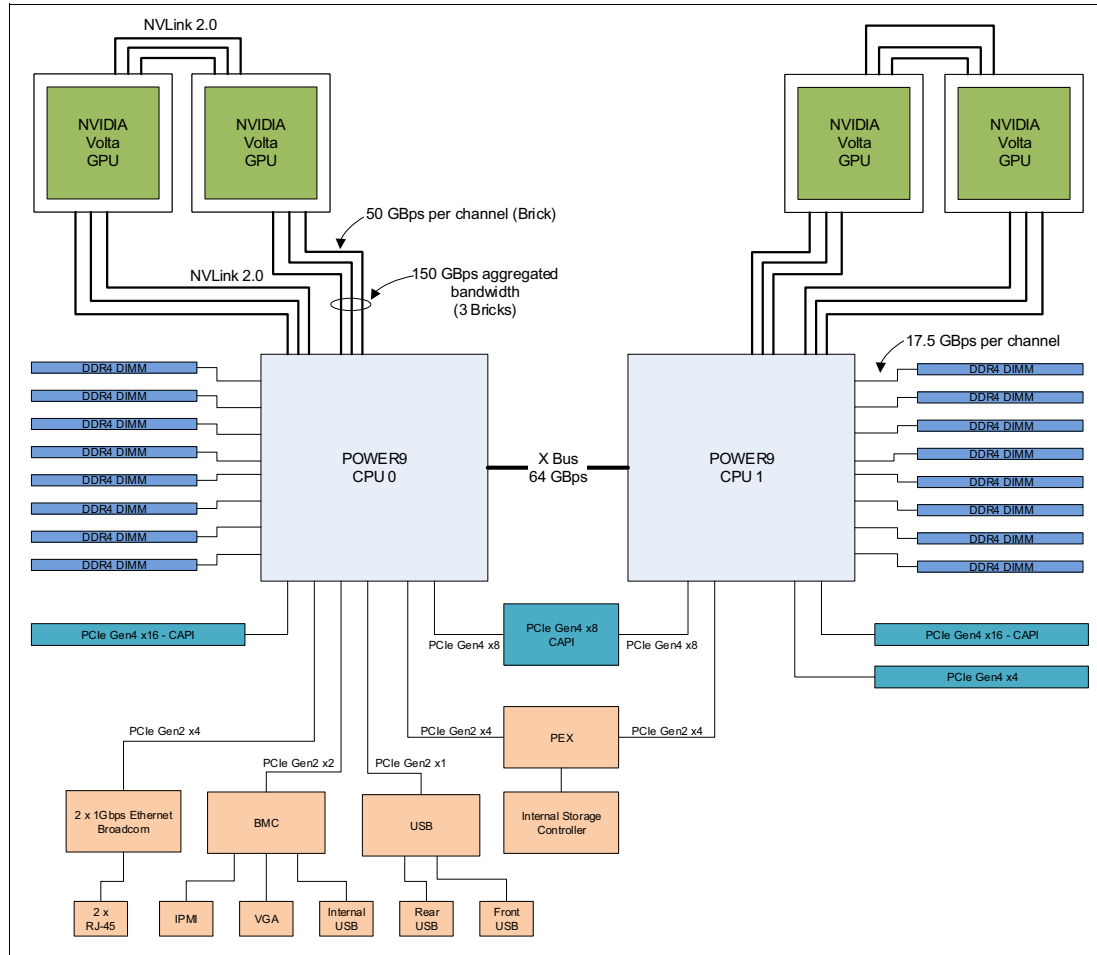


Figure 2-5 The Power AC922 server model GTH logical system diagram

Figure 2-6 shows the logical system diagram for the Power AC922 server model GTX with six connected GPUs, where the six NVLink Bricks are divided into three groups of two NVLink Bricks, enabling 100 GBps buses between GPUs, but enabling more connected GPUs.

NVLink wiring for the three GPU planar has two NV bricks (x8 lane) from the CPU to the GPU and two NV bricks from each GPU to GPU.

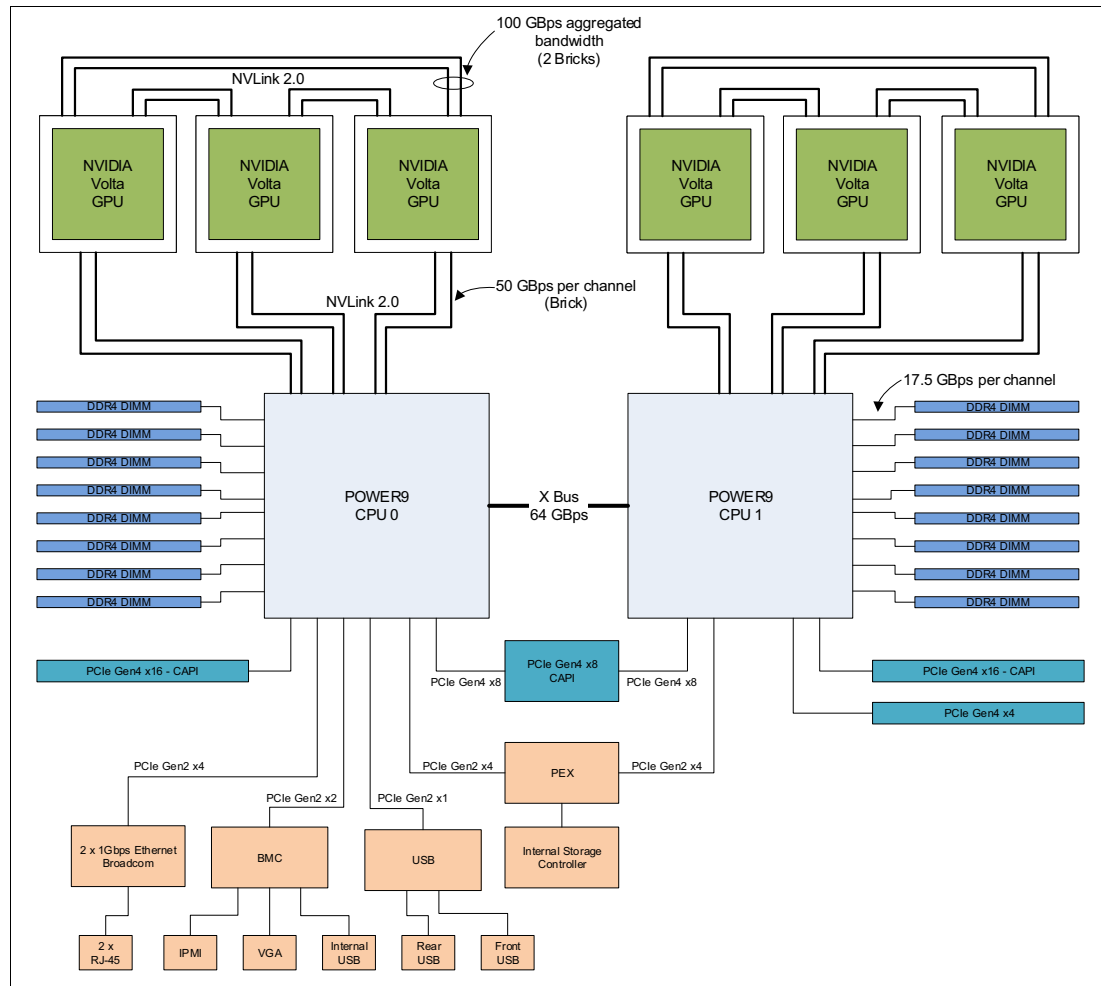


Figure 2-6 The Power AC922 server model GTX logical system diagram

2.2 Processor subsystem

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

The POWER9 processor in the Power AC922 server is the most current generation of the POWER processors family.

2.2.1 POWER9 processor overview

The IBM POWER9 processor is a super-scalar symmetric multiprocessor that is designed for use in servers and large-cluster systems. It also supports a maximum SMP size of two sockets and is targeted for high CPU consuming workloads.

The POWER9 processor offers superior cost and performance benefits. The target market segments are:

- ▶ Technical computing

The POWER9 processor provides superior floating-point performance and high-memory bandwidth to address this market segment. It also supports off-chip floating-point acceleration.

- ▶ Cloud operating environments

The POWER9 processor enables efficient cloud management software, enforces service-level agreements, and provides facilities for charge-back accounting based on the resources that are used.

- ▶ Big data analytics

The POWER9 processor with CAPI-attach, large caches, and on-chip accelerators provide a robust platform for analytic and big data applications.

- ▶ High-performance computing (HPC), high-performance data analytics (HPDA), and artificial intelligence (AI)

The POWER9 processor can be connected with NVIDIA GPUs by using NVLink 2.0 interfaces to increase processing capacities to fit the requirements of HPC, HPDA, and AI.

2.2.2 Processor features

The POWER9 processor is a SCM-based processor that is based on CMOS 14-nm technology with 17 metal layers. It is optimized for cloud and data center applications. Within a 68.5 mm × 68.5 mm footprint, it has eight direct-attached memory channels for scale-out configurations. Each DDR4 channel supports up to 2666 Mbps for one DIMM per channel or 2400 Mbps for two DIMMs per channel. Two processors are tightly coupled through two 4-byte 16 Gbps elastic differential interfaces (EDIs) for SMP. There are 34 lanes of PCIe Gen4 slots at 16 Gbps.

The POWER9 processor consists of the following main components:

- ▶ Twenty-four POWER9 cores that include both L1 instruction and data caches, shared L2 and L3 caches, and a non-cacheable unit (NCU).
- ▶ Each core has up to four threads that use simultaneous multithreading (SMT).
- ▶ On-chip accelerators:
 - CAPI allows a Field Programmable Gate Array (FPGA) or Application-Specific Integrated Circuit (ASIC) to connect coherently to the POWER9 processor SMP interconnect through the PCIe bus.
 - On-chip: Compression, encryption, data move that is initiated by the hypervisor, GZIP engine, or nest memory management unit (MMU) to enable user access to all accelerators.
 - In-core: User invocation encryption (advanced encryption standard (AES) or Secure Hash Algorithm (SHA)).
- ▶ Two memory controllers that support direct-attached DDR4 memory:
 - Supports eight direct-attach memory buses.
 - Supports four and eight 4 - 16 Gb DRAMs and 3D stacked DRAMs.

- ▶ Processor SMP interconnect:
 - Supports one inter-node SMP X-bus link.
 - Maximum two-socket SMP.
- ▶ Six 25 Gb NVLink Brick with support for OpenCAPI 3.0 and NVIDIA NVLink 2.0 interconnect.
- ▶ Three PCIe controllers (PEC) with 34 lanes of PCIe Gen4 I/O:
 - PEC0: One x16 lane.
 - PEC1: Two x1 lanes (bifurcation).
 - PEC2: One x16 lane mode or two x8 lanes (bifurcation), or one x8 lane and two x4 lanes (trifurcation).
 - PEC0 and PEC2 support CAPI 2.0.
- ▶ Power management.
- ▶ Pervasive interface.

From a logical perspective, the POWER9 processor consists of four main components:

- ▶ SMP interconnect (also known as an internal fabric interconnect).
- ▶ Memory subsystem.
- ▶ PCIe I/O subsystem.
- ▶ Accelerator subsystem.

Figure 2-7 shows a POWER9 processor with 24 cores.

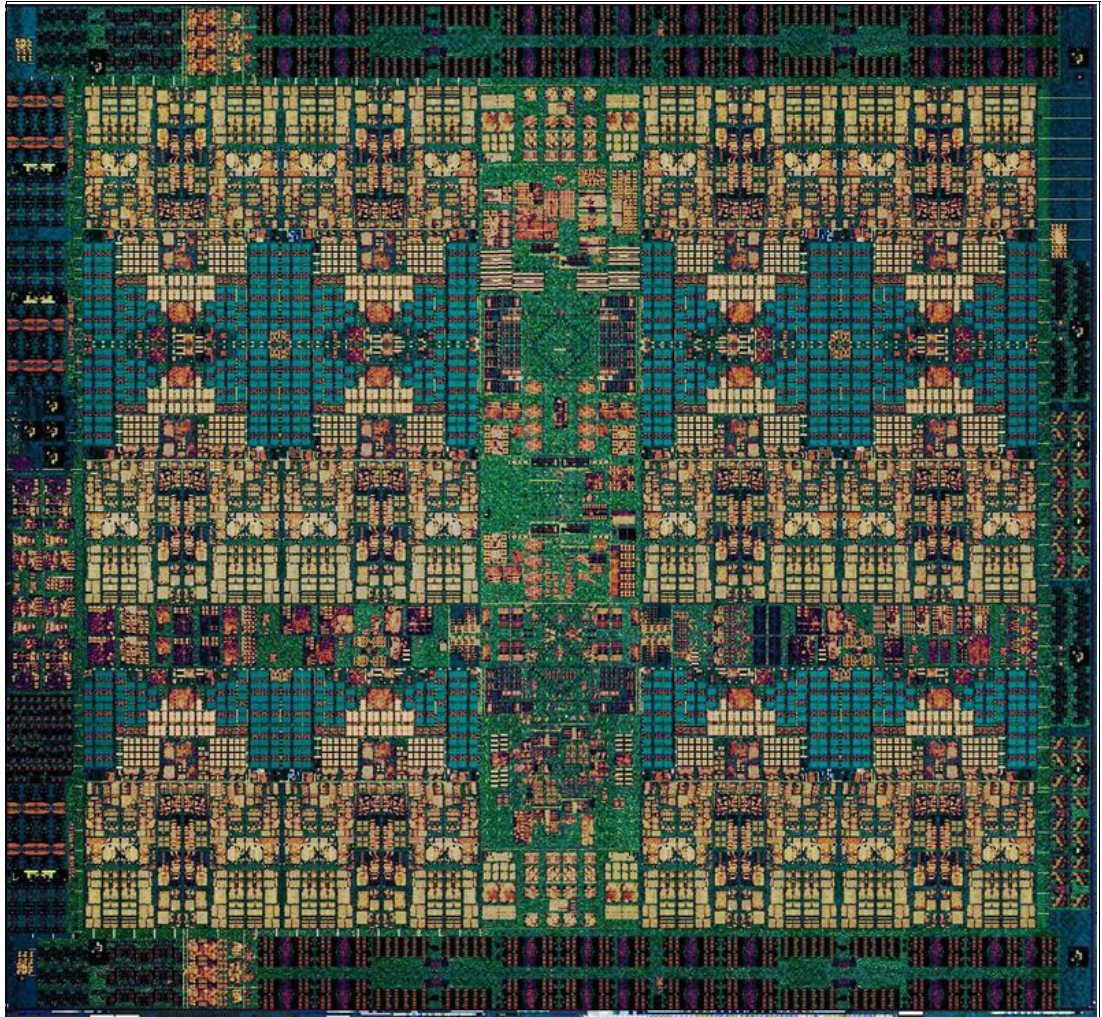


Figure 2-7 POWER9 processor with 24 cores

2.2.3 Supported technologies

The POWER9 processor supports the following technologies:

- ▶ Power Instruction Set Architecture (ISA) Book I, II, and III (Version 3.0)
- ▶ Linux on IBM Power Architecture® Platform Reference
- ▶ IEEE P754-2008 for binary and decimal floating-point compliant
- ▶ Big-endian, little-endian, and strong-ordering support extension
- ▶ 51-bit real address, 68-bit virtual address

2.2.4 Processor feature codes

The Power AC922 model GTH server supports two processor configurations only, as shown in Table 2-1. Processor features must be in quantities of two and cannot be mixed.

Table 2-1 Power AC922 model GTH supported processor feature codes

Feature code	Description	Minimum/Maximum	OS support
EP0P	16-core 2.7 GHz (3.3 GHz Turbo) POWER9 Processor	2/2	Linux
EP0R	20-core 2.4 GHz (3.0 GHz Turbo) POWER9 Processor	2/2	Linux

The Power AC922 model GTX server supports two processor configurations only, as shown in Table 2-2. Processor features must be in quantities of two and cannot be mixed.

Table 2-2 POWER9 AC922 model GTX supported processor feature codes

Feature code	Description	Minimum/Maximum	OS support
EP0Q	18-core 3.15 GHz (3.45 GHz Turbo) POWER9	2/2	Linux
EP0S	22-core 2.80 GHz (3.1 GHz Turbo) POWER9	2/2	Linux

2.3 Memory subsystem

The Power AC922 server is a two-socket system that supports two POWER9 SCM processor modules. The server supports a maximum of 16 DDR4 RDIMMs slots in the main system board that is directly connected to the POWER9 processors. Each memory slot can hold one DDR4 memory feature. A minimum of 16 memory features is required, filling all the memory slots and providing the following minimum configurations:

- ▶ With 8 GB DIMMs, a minimum of 128 GB of memory is required.
- ▶ With 16 GB DIMMs, a minimum of 256 GB of memory is required.
- ▶ With 32 GB DIMMs, a minimum of 512 GB of memory is required.
- ▶ With 64 GB DIMMs, a minimum of 1,024 GB of memory is required.
- ▶ With 128 GB DIMMs, a minimum of 2,048 GB of memory is required.

Memory features equate to a single memory DIMM. All memory DIMMs must be populated, and mixing of different memory feature codes (FCs) is not supported. The following memory DIMMs are supported:

- ▶ 8 GB DDR4
- ▶ 16 GB DDR4
- ▶ 32 GB DDR4
- ▶ 64 GB DDR4
- ▶ 128 GB DDR4

Plans for future memory growth needs should be accounted for when you decide which memory feature size to use at the time of initial system order because an upgrade requires a full replacement of the installed DIMMs.

2.3.1 Memory feature codes and placement rules

Each FC equates to a single memory DIMM. The supported maximum memory is 2048 GB by installing 16 128 GB memory DIMMs (#EM65). Table 2-3 shows the available memory FCs for ordering.

Table 2-3 Available memory feature codes

Feature code	CCIN	Description	Minimum/Maximum
EM60	324D	8 GB DDR4 Memory	16/16
EM61		16 GB DDR4 Memory	16/16
EM63	324F	32 GB DDR4 Memory	16/16
EM64	325A	64 GB DDR4 Memory	16/16
EM65	324C	128 GB DDR4 Memory	16/16

For the Power AC922 server (models 8335-GTH and 8335-GTX), the following requirements apply:

- ▶ All the memory DIMMs must be populated.
- ▶ Memory features cannot be mixed.
- ▶ The base memory is 128 GB with sixteen 8-GB DDR4 memory modules (#EM60).

Table 2-4 shows the total memory and how it can be achieved by using certain quantities of each memory FC.

Table 2-4 Supported memory feature codes for the Power AC922 server

Memory features	Total installed memory				
	128 GB	256 GB	512 GB	1024 GB	2048 GB
8 GB (#EM60)	16				
16 GB (#EM61)		16			
32 GB (#EM63)			16		
64 GB (#EM64)				16	
128 GB (#EM65)					16

2.3.2 Memory bandwidth

The POWER9 processor has exceptional cache, memory, and interconnect bandwidths. Table 2-5 shows the maximum bandwidth estimates for a single core on the server.

Table 2-5 The Power AC922 server single-core bandwidth estimates

Single core	8335-GTH		8335-GTX	
	3.30 GHz	3.0 GHz	3.45 GHz	3.10 GHz
L1 (data) cache	158.4 GBps	144 GBps	165.6 GBps	148.8 GBps
L2 cache	158.4 GBps	144 GBps	165.6 GBps	148.8 GBps
L3 cache	211.2 GBps	192 GBps	220.8 GBps	198.4 GBps

The bandwidth figures for the caches are calculated as follows:

- ▶ L1 cache: In one clock cycle, two 16-byte load operations and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core, and the formulas are as follows:
 - 3.3 GHz core: $(2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 3.3 \text{ GHz} = 158.4 \text{ GBps}$
 - 3.0 GHz core: $(2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 3.0 \text{ GHz} = 144 \text{ GBps}$
 - 3.45 GHz core: $(2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 3.45 \text{ GHz} = 165.6 \text{ GBps}$
 - 3.1 GHz core: $(2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 3.1 \text{ GHz} = 148.8 \text{ GBps}$
- ▶ L2 cache: In one clock cycle, one 32-byte load operation and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core, and the formulas are as follows:
 - 3.3 GHz core: $(1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 3.3 \text{ GHz} = 158.4 \text{ GBps}$
 - 3.0 GHz core: $(1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 3.0 \text{ GHz} = 144 \text{ GBps}$
 - 3.45 GHz core: $(1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 3.45 \text{ GHz} = 165.6 \text{ GBps}$
 - 3.1 GHz core: $(1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 3.1 \text{ GHz} = 148.8 \text{ GBps}$
- ▶ L3 cache: One 32-byte load operation and one 32-byte store operation can be accomplished at half-clock speed, and the formulas are as follows:
 - 3.3 GHz core: $(1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 3.3 \text{ GHz} = 211.2 \text{ GBps}$
 - 3.0 GHz core: $(1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 3.0 \text{ GHz} = 192 \text{ GBps}$
 - 3.45 GHz core: $(1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 3.45 \text{ GHz} = 220.8 \text{ GBps}$
 - 3.1 GHz core: $(1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 3.1 \text{ GHz} = 198.4 \text{ GBps}$

Table 2-6 shows the overall bandwidths for the entire Power AC922 server that is populated with the two processor modules.

Table 2-6 Overall bandwidths for a Power AC922 server that is populated with two processor modules

Total bandwidth	8335-GTH		8335-GTX	
	16 cores @ 3.3 GHz	20 cores @ 3.0 GHz	18 cores @ 3.45 GHz	22 cores @ 3.1 GHz
L1 (data) Cache	2534.4 GBps	2880 GBps	2980.8 GBps	3273.6 GBps
L2 Cache	2534.4 GBps	2880 GBps	2980.8 GBps	3273.6 GBps

Total bandwidth	8335-GTH		8335-GTX	
Core	16 cores @ 3.3 GHz	20 cores @ 3.0 GHz	18 cores @ 3.45 GHz	22 cores @ 3.1 GHz
L3 Cache	3379.2 GBps	3840 GBps	3974.4 GBps	4364.8 GBps
Total memory	280 GBps	280 GBps	280 GBps	280 GBps
SMP interconnect	64 GBps	64 GBps	64 GBps	64 GBps
PCIe interconnect	272 GBps	272 GBps	272 GBps	272 GBps

Where:

- ▶ Total memory peak bandwidth: Each POWER9 processor has eight memory channels running at 21.25 GBps. The bandwidth formula is calculated as follows:
Eight channels x 21.25 GBps = 170 GBps per processor module
- ▶ SMP interconnect: The POWER9 processors are connected by using an X-bus. The bandwidth formula is calculated as follows:
 $1 \text{ X bus} * 4 \text{ bytes} * 16 \text{ GHz} = 64 \text{ GBps}$
- ▶ PCIe interconnect: Each POWER9 processor has 34 PCIe lanes running at 16 Gbps full-duplex. The bandwidth formula is calculated as follows:
 $34 \text{ lanes} * 2 \text{ processors} * 16 \text{ Gbps} * 2 = 272 \text{ GBps}$

2.4 I/O subsystem

The key components of the I/O subsystem are described in this section.

2.4.1 PCI Express Controller

The PEC bridges between the internal processor bus and the high-speed serial (HSS) links that drive the PCIe I/O. The PEC acts as a processor bus master on behalf of the PCIe port, converting inbound memory read and write packets into processor bus direct memory access (DMA) traffic. The PEC also acts as a processor bus subordinate, transferring processor load and store commands to the PCIe devices that are attached to the port.

PCIe uses a serial interface and enables 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 these configurations, the connection is labeled as x1, x2, x8, x12, x16, or x32, where the number is effectively the number of lanes.

The Power AC922 server supports the new PCIe Gen4 adapter, which is capable of 32 GBps simplex (64 GBps duplex) on a single x16 interface. PCIe Gen4 slots also support previous generation (Gen3 and Gen2) 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 lower 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).

Note: PCIe x4, x8, and x16 adapters use different types of slots. If you attempt to force an adapter into the wrong type of slot, you might damage the adapter or the slot.

The POWER9 processor-based AC922 servers support PCIe low-profile (LP) cards because of the restricted height of the server.

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 existing update prerequisites to install. To obtain this information, see [Power Systems Prerequisites](#).

The following sections describe other I/O technologies that enhance or replace the PCIe interface.

2.4.2 IBM CAPI2

IBM CAPI2 is the evolution of CAPI that 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) that 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 2-8 shows a comparison of the traditional model, where the accelerator must go through the processor to access memory with CAPI.

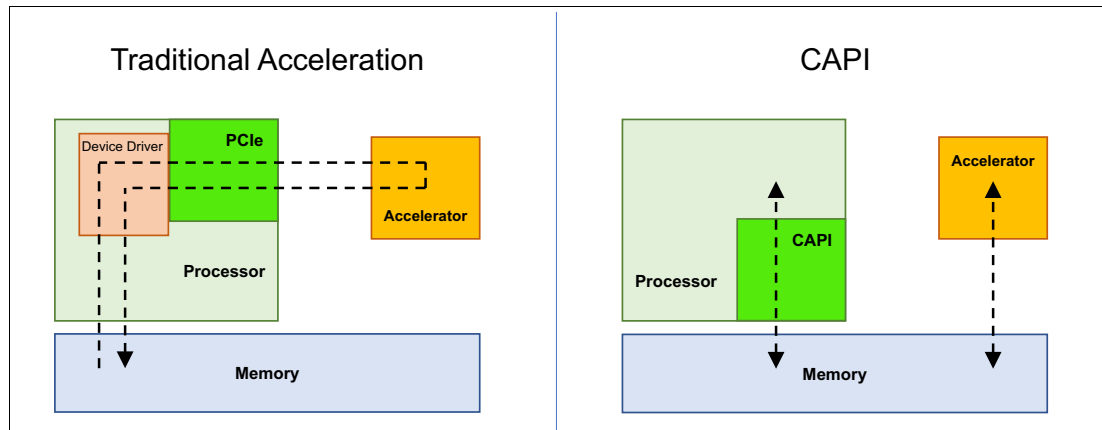


Figure 2-8 CAPI accelerator that is attached to the POWER9 processor

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.

CAPI 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 DMA 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.

For a list of supported CAPI adapters, see 2.5.4, “CAPI-enabled adapters” on page 34.

2.4.3 OpenCAPI

Although CAPI is a technology that is present in IBM POWER processors and depends on IBM intellectual property (the PSL), several industry solutions might benefit from having a mechanism of connecting different devices to the processor with low latency, including memory attachment. The PCIe standard is pervasive to every processor technology, but its design characteristics and latency do not enable the attachment of memory for load/store operations.

Therefore, the OpenCAPI Consortium was created, with the goal of defining a device attachment interface to open the CAPI interface to other hardware developers and extending its capabilities. OpenCAPI aims to enable memory, accelerators, network, storage, and other devices to connect to the processor through a high-bandwidth, low-latency interface, becoming the interface of choice for connecting high-performance devices.

By providing a high-bandwidth, low-latency connection to devices, OpenCAPI enables several applications to improve networking, use FPGA accelerators, use expanded memory beyond server internal capacity, and reduce latency to storage devices. Some of these use cases and examples are shown in Figure 2-9.

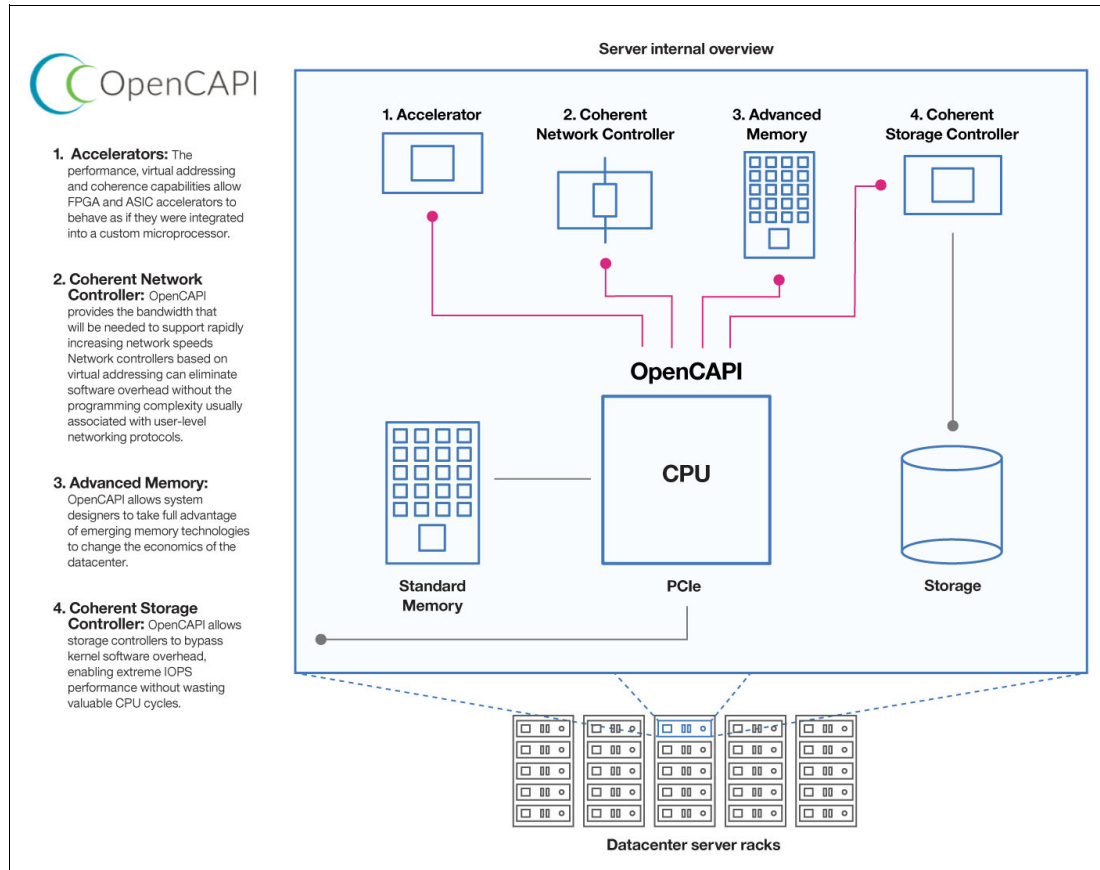


Figure 2-9 OpenCAPI use cases

The design of OpenCAPI enables low latency in accessing attached devices (in the same range of DDR memory access, that is, 10 ns), which enables memory to be connected through OpenCAPI and serve as main memory for load/store operations. In contrast, PCIe latency is 10 times larger (around 100 ns). Therefore, OpenCAPI has a significant enhancement compared to traditional PCIe interconnects.

OpenCAPI is neutral regarding processor architecture, so the electrical interface is not defined by the OpenCAPI consortium or any of its workgroups. On the POWER9 processor, the electrical interface is based on the design from the 25G workgroup within the OpenPower Foundation, which encompasses a 25 Gbps signaling and protocol that is built to enable a low-latency interface on CPU and attached devices.

The current design for the adapter is based on a PCIe card that draws power from the PCIe slot while connecting to the OpenCAPI port on the system board through a 25-GBps cable, as shown in Figure 2-10.

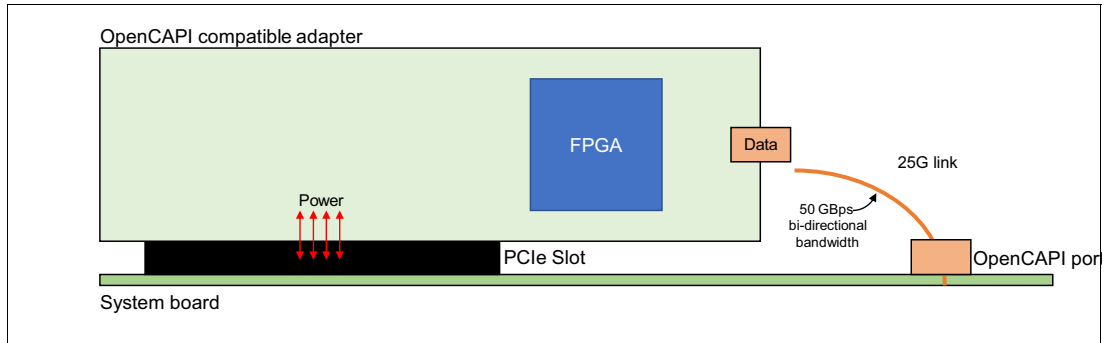


Figure 2-10 OpenCAPI compatible adapter and 25G link

The OpenCAPI interface uses the same electrical interconnect as NVLink 2.0. Systems can be designed to have an NVLink-attached GPU, an OpenCAPI-attached device, or both. The use of OpenCAPI adapters limits the number of NVLink ports that are available for GPU communication. Each POWER9 chip has six NVLink ports, four of which can be used for OpenCAPI adapters, as shown in Figure 2-11.

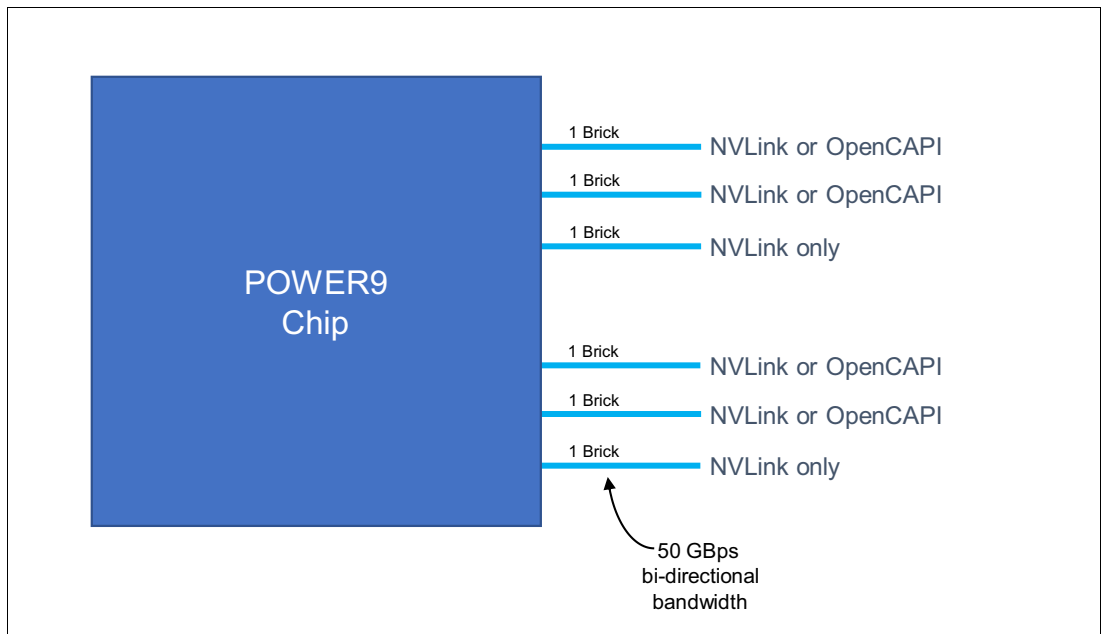


Figure 2-11 OpenCAPI and NVLink shared ports on the POWER9 chip

2.4.4 The NVIDIA Tesla V100

The new NVIDIA Tesla V100 accelerator, code name Volta, takes GPU computing to the next level. This section describes the Tesla V100 accelerator, which is shown in Figure 2-12.

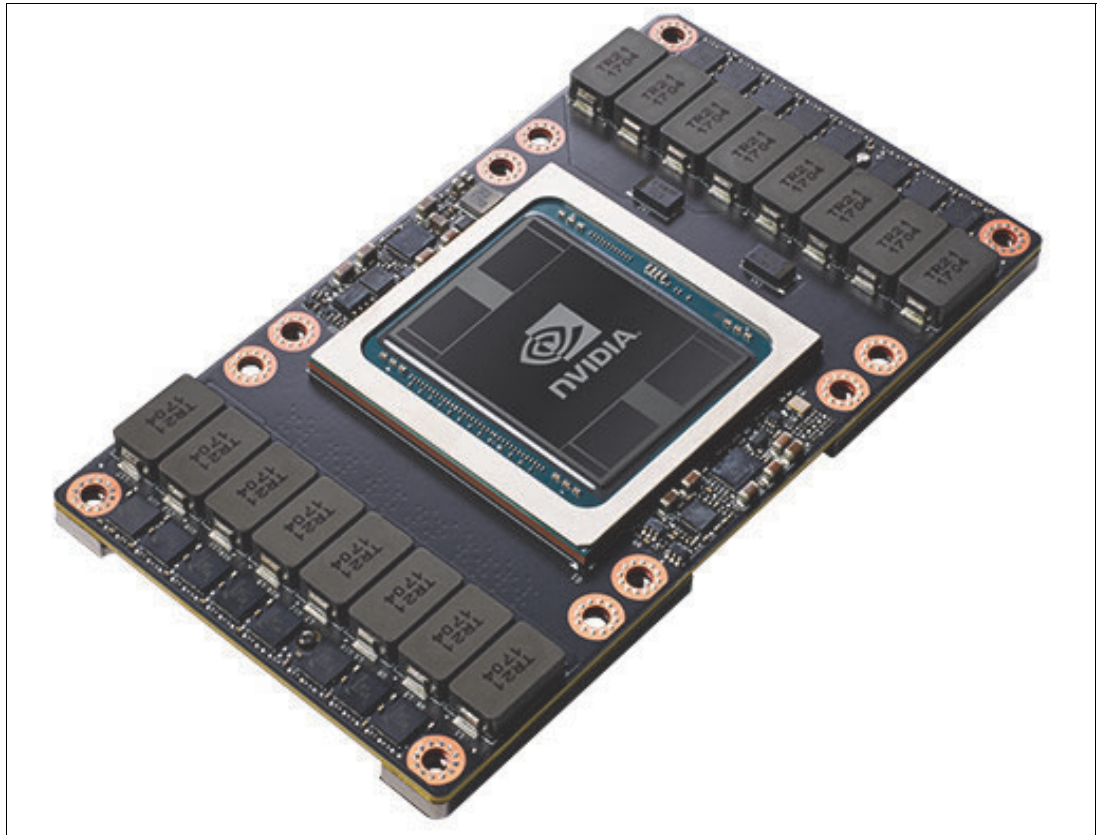


Figure 2-12 NVIDIA Tesla V100 for NVLink accelerator

NVIDIA Tesla V100 is the most advanced data center GPU built to accelerate AI, HPC, and graphics. Powered by NVIDIA Volta, the current GPU architecture, Tesla V100 offers the performance of 100 CPUs in a single GPU, which enables data scientists, researchers, and engineers to tackle challenges that were once impossible.

The Tesla V100 includes the following key features:

- ▶ Volta architecture

By pairing CUDA cores and Tensor cores within a unified architecture, a single server with Tesla V100 GPUs can replace hundreds of commodity CPU servers for traditional HPC and deep learning (DL).

- ▶ Tensor core

Equipped with 640 Tensor cores, Tesla V100 delivers 125 teraflops (TFLOPS) of DL performance, that is, 12x Tensor TFLOPS for DL training, and 6x Tensor TFLOPS for DL inference compared to NVIDIA Pascal GPUs.

- ▶ Next generation NVLink

NVIDIA NVLink in Tesla V100 delivers 2x higher throughput compared to the previous generation. Up to eight Tesla V100 accelerators can be interconnected at up to 300 GBps to unleash the highest application performance possible on a single server.

- ▶ **Maximum efficiency mode**

The new maximum efficiency mode enables data centers to achieve up to 40% higher compute capacity per rack within the existing power budget. In this mode, Tesla V100 runs at peak processing efficiency, providing up to 80% of the performance at half the power consumption.

- ▶ **HBM2**

With a combination of improved raw bandwidth of 900 GBps and higher DRAM usage efficiency at 95%, Tesla V100 delivers 1.5x higher memory bandwidth over Pascal GPUs, as measured on STREAM.

- ▶ **Programmability**

The Tesla V100 is designed to simplify programmability. Its new independent thread scheduling enables finer-grain synchronization and improves GPU usage by sharing resources among small jobs.

The Tesla V100 delivers exceptional performance for the most demanding compute applications. It delivers the following performance benefits:

- ▶ 7.8 TFLOPS of double-precision floating point (FP64) performance
- ▶ 15.7 TFLOPS of single-precision (FP32) performance
- ▶ 125 Tensor TFLOPs of mixed-precision

With 640 Tensor cores, Tesla V100 is the first GPU to break the 100 TFLOPS barrier of DL performance. The next generation of NVIDIA NVLink connects multiple V100 GPUs at up to 300 GBps to create the most powerful computing servers. AI models that would use weeks of computing resources on previous systems can now be trained in a few days. With this dramatic reduction in training time, many problems are now solvable with AI.

Multiple GPUs are common in workstations, as are the nodes of HPC clusters and DL training systems. A powerful interconnect is valuable in multiprocessing systems. The NVIDIA Tesla V100 does not rely on traditional PCIe for data transfers, but instead uses the new NVLink 2.0 bus that creates an interconnect for GPUs that offer higher bandwidth than PCIe Gen3. The GPUs are compatible with the GPU ISA to support shared memory multiprocessing workloads.

Once PCIe buses are not used for data transfer, the GPU cards do not need to comply with the traditional PCIe card format. To improve density in the Power AC922 server, the GPUs have a different form factor called *SXM2*. This form factor enables the GPU to be connected directly on the system board.

Figure 2-13 shows the SXM2 GPU module top and bottom views and the connectors that are used for the GPU modules.

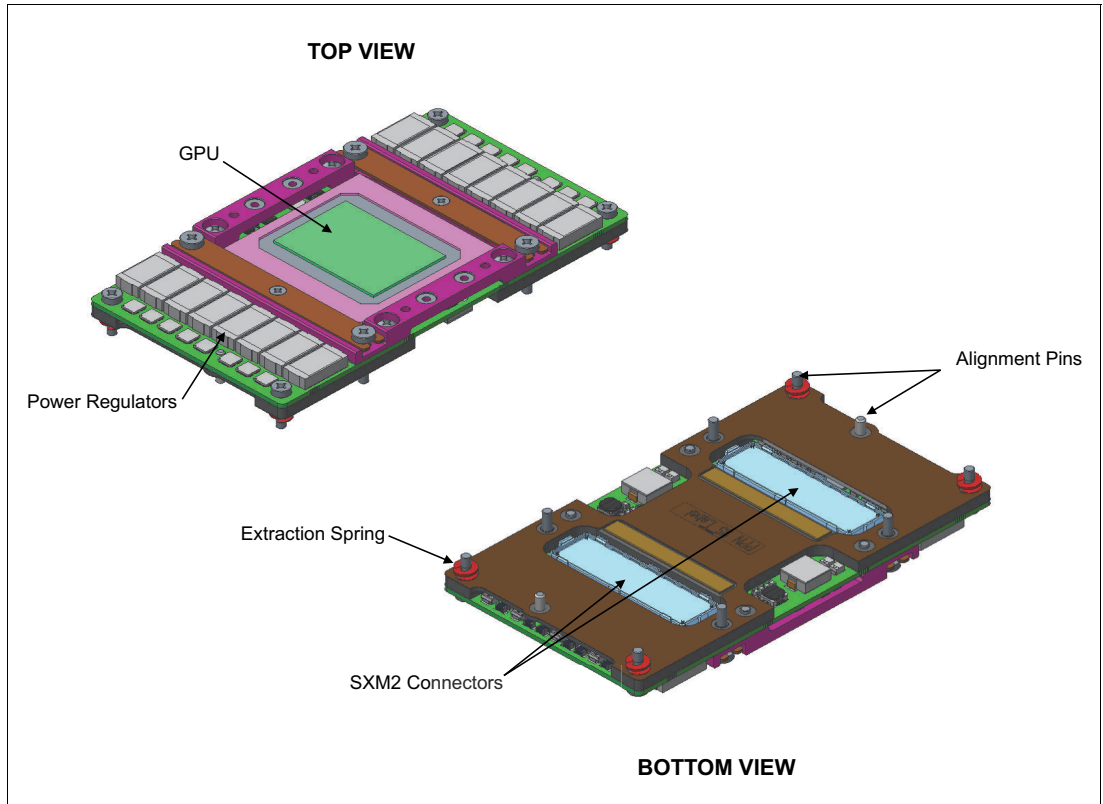


Figure 2-13 SXM2 GPU module views

Figure 2-14 shows the location of the GPUs on the Power AC922 system board.

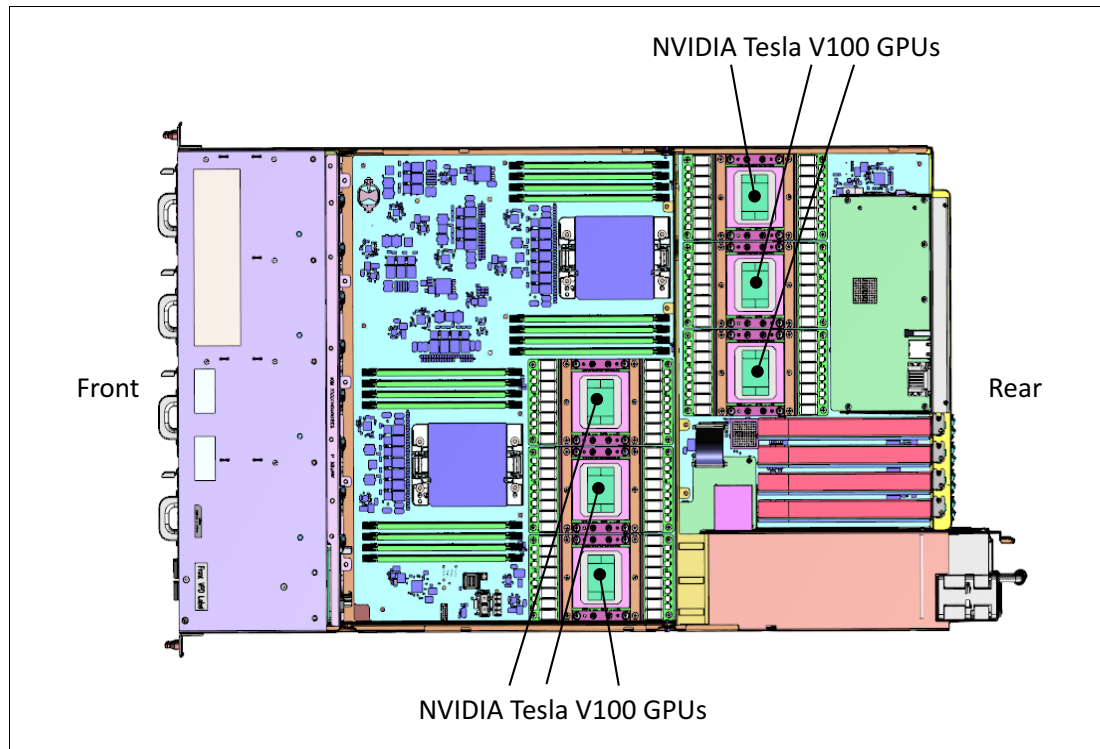


Figure 2-14 GPU location in a six-GPU configuration of the model GTX

Cooling for four-GPU configurations on Power AC922 model GTH is done by air cooling, and cooling for six-GPU configurations on Power AC922 model GTX is done by water cooling. For more information about server water cooling, see Chapter 3, “Physical infrastructure” on page 45.

For more information about the Tesla V100, see [Inside Volta Parallel for All](#).

2.4.5 NVLink 2.0

NVLink 2.0 is the NVIDIA new generation high-speed interconnect technology for GPU-accelerated computing. Supported on SXM2-based Tesla V100 accelerator system boards, NVLink increases performance for both GPU-to-GPU communications and for GPU access to system memory.

Support for the GPU ISA enables programs running on NVLink-connected GPUs to run directly on data in the memory of another GPU and on local memory. GPUs can also perform atomic memory operations on remote GPU memory addresses, enabling much tighter data sharing and improved application scaling.

NVLink 2.0 uses the NVIDIA High-Speed Signaling (NVHS) interconnect. NVHS transmits data over a link that is called NVLink Brick that connects two processors (GPU-to-GPU or GPU-to-CPU). A single NVLink Brick supports up to 50 Gbps of bidirectional bandwidth between the endpoints. Multiple links can be combined to form *Gangs* for even higher-bandwidth connectivity between processors. The NVLink implementation in Tesla V100 supports up to six links, enabling a Gang with an aggregate maximum theoretical bandwidth of 300 GBps bidirectional bandwidth.

Although traditional NVLink implementation primarily focuses on interconnecting multiple NVIDIA Tesla V100 GPUs together, under POWER9 it also connects Tesla V100 GPUs with IBM POWER9 CPUs, enabling direct system memory access and providing GPUs with an extended memory orders of magnitude larger than the internal 16 GB memory.

On a Power Systems implementation, NVLink Bricks are always combined to provide the highest bandwidth possible.

Figure 2-15 compares the bandwidth of the POWER9 processor that is connected with two GPUs and three GPUs.

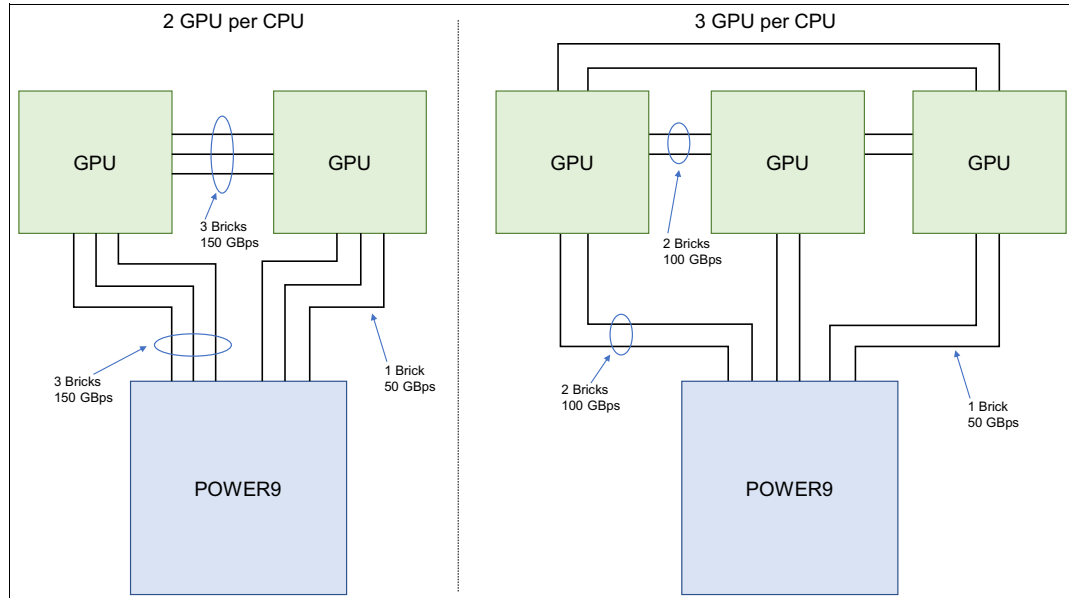


Figure 2-15 CPU to GPU and GPU to GPU interconnect that uses NVLink 2.0

All the initialization of the GPU is through the PCIe interface. The PCIe interface also contains the side-band communication for status, power management, and so on. After the GPU is running, all data communication uses the NVLink.

2.5 PCI adapters

This section describes the types and functions of the PCI adapters that are supported by the Power AC922 server.

The Power AC922 server uses the current PCIe Gen4 technology, enabling 32 GBps unidirectional and 64 GBps bidirectional bandwidth.

Note: PCIe adapters on the Power AC922 server are not hot-pluggable.

2.5.1 Slot configuration

The Power AC922 server has four PCIe Gen4 slots.

Figure 2-16 shows a rear-view diagram of the Power AC922 server with its PCIe slots.

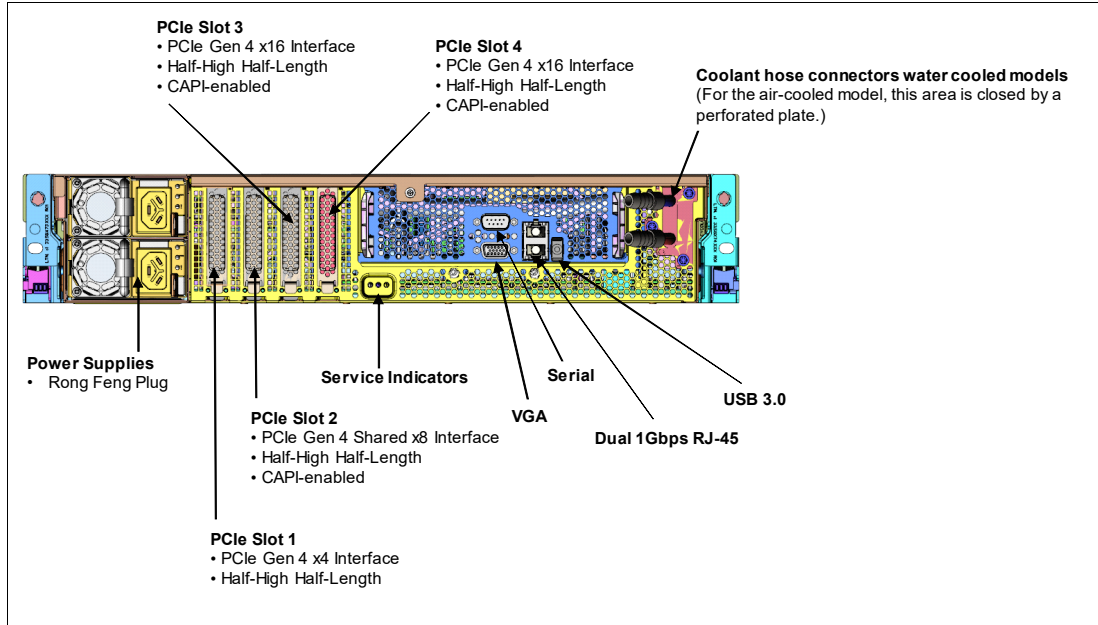


Figure 2-16 Rear-view PCIe slots and main components

Table 2-7 shows the PCIe Gen4 slot properties.

Table 2-7 The Power AC922 server PCIe Gen4 slot properties

Slot	Description	Card size	CAPI-capable	Location
Slot 1	PCIe Gen4 x4	Half-height, half-length	No	P1C5
Slot 2	PCIe Gen4 x8 Shared	Half-height, half-length	Yes	P1C4
Slot 3	PCIe Gen4 x16	Half-height, half-length	Yes	P1C3
Slot 4	PCIe Gen4 x16	Half-height, half-length	Yes	P1C2

Slot 2 has a shared connection between the two POWER9 CPUs. When you use a dual-channel Mellanox InfiniBand ConnectX5 (EDR) Network Interface Card (NIC) (#EC64), it enables each CPU to have direct access to the InfiniBand card. If the #EC64 card is not installed, the shared slot operates as a single x8 PCIe Gen4 slot that is attached to processor 0.

PCIe adapters can be placed only in PCIe slots. Place x4, x8, and x16 speed adapters in the same connector size 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 (for example, a x16 adapter cannot go into an x8 PCIe slot connector).

There are exceptions to these rules, as explained below:

- ▶ Slot P1C4 runs at x8 speeds but has a x16 connector interface. Slot P1C5 runs at x4 speeds, but has a x8 connector interface.

- ▶ PCI-X cards are not offered/supported on a GTX system.

Note: PCIe x8 adapters use a different type of slot than PCI x16 adapters. If you attempt to force an adapter into the wrong type of slot, you might damage the adapter or the slot.

Figure 2-17 shows the logical diagram of the slot 2 that is connected to the two POWER9 processors.

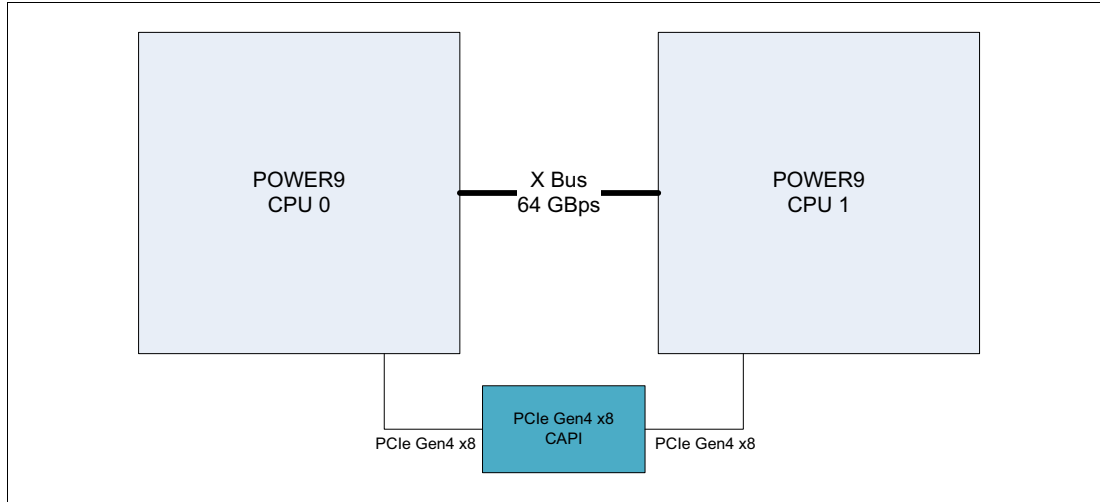


Figure 2-17 Shared PCIe slot 2 logical diagram

Only LP adapters can be placed in LP slots. A x8 adapter can be placed in a x16 slot, but a x16 adapter cannot be placed in a x8 slot.

Note: PCIe x4, x8, and x16 adapters use different types of slots. If you attempt to force an adapter into the wrong type of slot, you might damage the adapter or the slot.

2.5.2 Local area network adapters

To connect the Power AC922 server to a local area network (LAN), you can use the LAN adapters that are supported in the PCIe slots of the system unit. Table 2-8 lists the supported LAN adapters for the server.

Table 2-8 Supported LAN adapters for model GTH and GTX

Feature code	CCIN	Description	Maximum	OS support
EC2R	58FA	PCIe3 LP 2-port 10 Gb NIC & ROCE SR/Cu Adapter	3	Linux
EC2T	58FB	PCIe3 LP 2-port 25/10 Gb NIC & ROCE SR/Cu Adapter	3	Linux
EC3L	2CEC	PCIe3 LP 2-port 100 GbE (NIC & RoCE) QSFP28 Adapter x16	2	Linux
EC62	2CF1	PCIe4 LP 1-port 100 Gb EDR InfiniBand CAPI adapter	3	Linux
EC64	2CF2	PCIe4 LP 2-port 100 Gb EDR InfiniBand CAPI adapter	3	Linux
EL3Z	2CC4	PCIe2 LP 2-port 10/1 GbE BaseT RJ45 Adapter	4	Linux
EL4M	576F	PCIe2 LP 4-port 1 GbE Adapter	4	Linux

Feature code	CCIN	Description	Maximum	OS support
EN0T	2CC3	PCIe2 LP 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter	4	Linux
EN0V	2CC3	PCIe2 LP 4-port (10 Gb+1 GbE) Copper SFP+RJ45 Adapter	4	Linux

2.5.3 Fibre Channel adapters

The Power AC922 server supports direct or SAN connection to devices that use Fibre Channel adapters. Table 2-9 summarizes the available Fibre Channel adapters, which all have LC connectors.

If you are attaching a device or switch with an SC-type fiber connector, an LC-SC 50-micron fiber converter cable (#2456) or an LC-SC 62.5-micron fiber converter cable (#2459) is required.

Table 2-9 Available Fibre Channel adapters for the GTH and GTX models

Feature code	CCIN	Description	Maximum	OS support
EL43	577F	PCIe3 LP 2-port 16 Gb Fibre Channel Adapter	3	Linux
EL5V	578F	PCIe3 LP 2-port 32 Gb Fibre Channel Adapter	3	Linux

2.5.4 CAPI-enabled adapters

CAPI defines a coherent accelerator interface structure for attaching special processing devices to the POWER9 processor bus. Table 2-10 shows the available CAPI adapters.

Table 2-10 Available CAPI adapters for the GTH and GTX models

Feature code	CCIN	Description	Maximum	OS support
EC3L	2CEC	PCIe3 LP 2-port 100 GbE (NIC & RoCE) QSFP28 Adapter x16	2	Linux
EC62	2CF1	PCIe4 LP 1-port 100 Gb EDR InfiniBand CAPI adapter	3	Linux
EC64	2CF2	PCIe4 LP 2-port 100 Gb EDR InfiniBand CAPI adapter	3	Linux

2.5.5 Compute-intensive accelerators

Compute-intensive accelerators are GPUs that are developed by NVIDIA to off load processor-intensive operations to a GPU accelerator and boost performance. The Power AC922 server can be configured with GPUs that are air-cooled and water-cooled based on the model.

Table 2-11 lists the available compute-intensive accelerators that are supported in the Power AC922 model GTH server.

Table 2-11 Supported graphics processing units adapters for the Power AC922 model GTH server

Feature code	Description	Maximum	OS support
EC4J	NVIDIA Tesla V100 GPU with NVLink Air-Cooled (16 GB)	4	Linux
EC4L	NVIDIA Tesla V100 GPU with NVLink Air-Cooled (32 GB)	4	Linux

Table 2-12 lists the available compute-intensive accelerators that are supported in the Power AC922 model GTX server.

Table 2-12 Supported graphics processing units adapters for Power AC922 model GTX server

Feature code	Description	Maximum	OS support
EC4H	NVIDIA Tesla V100 GPU with NVLink Water-Cooled (16 GB)	6	Linux
EC4K	NVIDIA Tesla V100 GPU with NVLink Water-Cooled (32 GB)	6	Linux

Note: The Power AC922 model GTX server is a water-cooled server and cannot be configured without GPUs. The minimum order is four GPUs.

2.5.6 Flash storage adapters

The available flash storage adapters are shown in Table 2-13.

Table 2-13 Supported flash storage adapter for the GTH and GTX models

Feature code	CCIN	Description	Maximum	OS support
EC5C	58FD	PCIe3 LP 3.2 TB SSD NVMe adapter	3	Linux
EC5E	58FE	PCIe3 LP 6.4 TB SSD NVMe adapter	3	Linux
EC5G	58FC	PCIe3 LP 1.6 TB SSD NVMe Adapter	3	Linux

2.6 System ports

The system board has two 1 Gbps Ethernet ports, one Intelligent Platform Management Interface (IPMI) port, one rear USB 3.0 port (the models 8335-GTH/8335-GTX also have one front USB 3.0 port), and a VGA port, as shown in Figure 2-18.

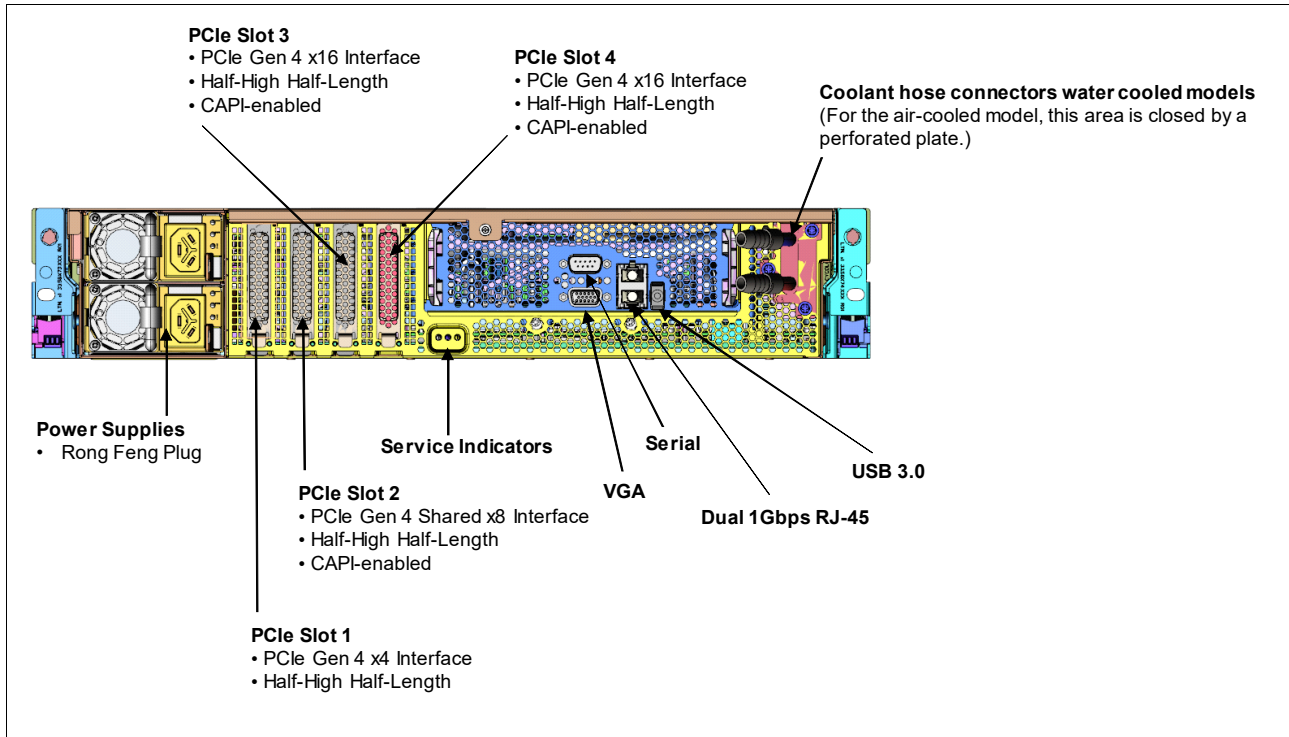


Figure 2-18 Power SAC922 system ports

The integrated system ports are supported for modem and asynchronous terminal connections with Linux. Any other application that uses serial ports requires a serial port adapter to be installed in a PCI slot. The integrated system ports do not support IBM PowerHA® configurations. The VGA port does not support cable lengths that exceed 3 meters.

2.7 Internal disks

The internal storage on the Power AC922 server contains the following features:

- ▶ A storage backplane for two 2.5-inch small form factor (SFF) Gen4 SATA HDDs or SDDs.

Limitation: The disks use an SFF-4 carrier. Disks that are used in other Power Systems servers usually have an SFF-3 or SFF-2 carrier and are not compatible with this system.

- ▶ One integrated SATA disk controller (non-RAID).
- ▶ The storage split backplane feature is not supported.

Table 2-14 presents a summarized view of these features.

Table 2-14 Summary of features for the integrated SATA disk controller

Option	Integrated SATA disk controller
Supported RAID types	None - JBOD
Disk bays	Two SFF Gen4 (HDDs/SDDs)
SATA controllers	Single
IBM Easy Tier® capable controllers	No
External SAS ports	No
Split backplane	No

The 2.5-inch or SFF SAS bays can contain SATA drives (HDDs or SSDs) that are mounted on a Gen4 tray or carrier (also known as SFF-4). SFF-2 or SFF-3 drives do not fit in an SFF-4 bay. All SFF-4 bays support concurrent maintenance or hot-plug capability.

2.7.1 Disk and media features

The server supports the attachment of up to two SATA storage devices.

Table 2-15 lists the supported devices that can be installed. Disk features cannot be mixed.

Table 2-15 Supported disk and media features for the Power AC922 server

Feature code	CCIN	Description	Maximum
ELU4		960 GB 2.5-inch SATA/SSD Disk Drive	2
ELD0		1 TB 7.2 K RPM 5xx SATA SFF-4 Disk Drive	2
ELU5		1.92 TB 2.5-inch SATA/SSD Disk Drive	2
ES6A	5B22	2 TB 7.2 K RPM 5xx SATA SFF-4 Disk Drive	2
ELU6		3.84 TB 2.5-inch SATA/SSD Disk Drive	2

The Power AC922 server is designed for network or USB media installation. It does not support an internal DVD drive.

A stand-alone USB DVD drive can be configured (#EUA5).

2.8 External I/O subsystems

The Power AC922 server does not support external PCIe Gen3 I/O expansion drawers or the IBM EXP24S, EXP12X, and EXP24SX storage drawers.

2.9 Location codes

Figure 2-19 shows the location for the Power AC922 server main components with key features.

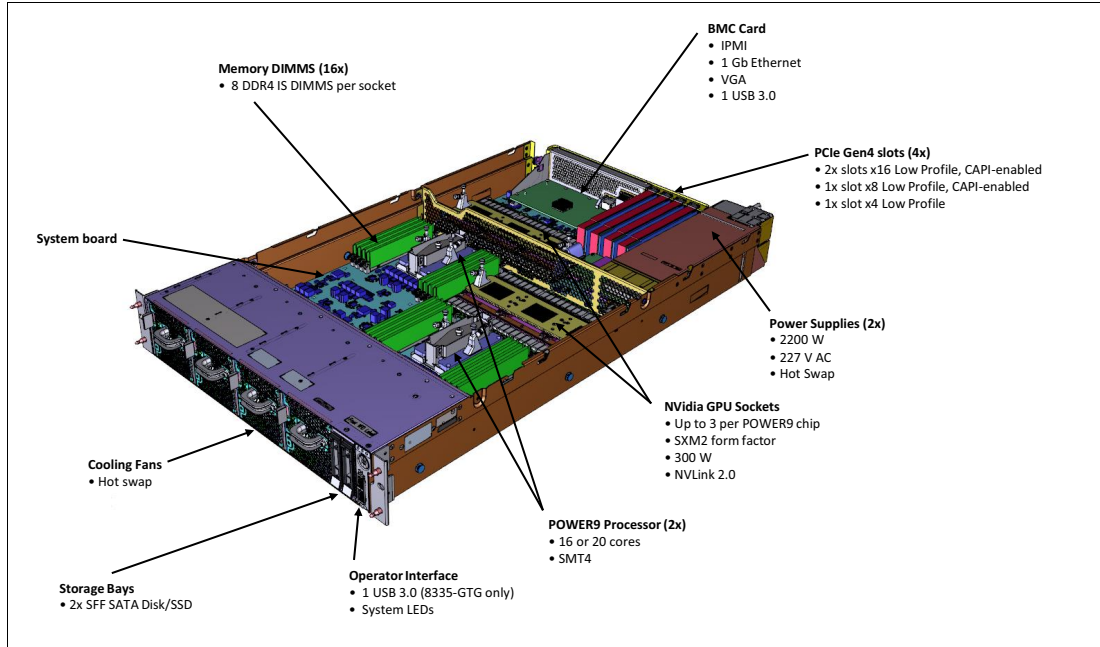


Figure 2-19 Power AC922 main components

Figure 2-20 shows the location codes and detailed information for the Power AC922 server main components.

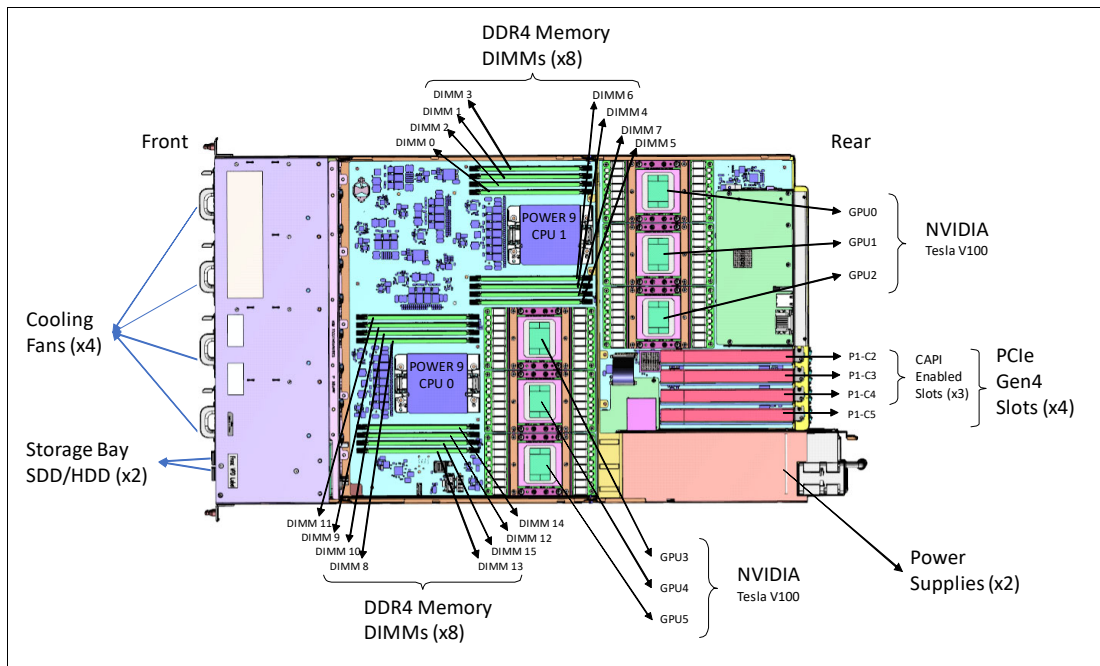


Figure 2-20 POWER AC922 main components with location codes

2.10 IBM System Storage

The IBM System Storage™ disk systems products and offerings provide compelling storage solutions with superior value for all levels of business, from entry-level to high-end storage systems. For more information about the various offerings, see [IBM System Storage](#).

The following section highlights a few of the offerings.

2.10.1 IBM Flash Storage

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

2.10.2 Software-defined storage

Software-defined storage (SDS) manages data growth and enables multi-cloud flexibility by providing an agile, scalable, and operations-friendly infrastructure. For more information, see [IBM Storwize](#).

2.10.3 Hybrid storage

Optimize your mix of storage media, including flash storage, to achieve the best balance of performance and economics. For more information, see [IBM Flash Storage](#).

2.10.4 Storage area network

IBM offers a comprehensive portfolio of Fibre Channel storage area network (SAN) switches to support your virtualization, cloud, and big data requirements. For more information, see [IBM XIV Storage System](#).

2.11 Operating system support

The Power AC922 server supports Linux, which provides a UNIX-like implementation across many computer architectures.

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

2.11.1 Ubuntu

Ubuntu Server 18.04 LTS for IBM POWER9 is supported on the server with support for later distributions as they become available.

For more information about Ubuntu Server for Ubuntu for POWER9, see [Ubuntu Server for IBM POWER](#).

2.11.2 Red Hat Enterprise Linux

Red Hat Enterprise Linux (ppc64le) Version 7.4 is supported on the POWER9 server with support for later distributions as they become available.

For more questions about this release and supported Power Systems servers, consult the [Red Hat Hardware Catalog](#).

2.12 Java

When running Java applications on the POWER9 processor, the pre-packaged Java that is part of a Linux distribution is designed to meet the most common requirements. If you require a different level of Java, there are several resources that are available:

- ▶ Current information about IBM Java and tested Linux distributions are available at [Java JDK](#).
- ▶ More information about the OpenJDK port for Linux on PPC64 LE and some pre-generated builds can be found at [OpenJDK PowerPC Port](#).
- ▶ Launchpad.net has resources for Ubuntu builds. You can find out about them at the following websites:
 - [The openjdk-9 package in Ubuntu](#)
 - [The openjdk-8 package in Ubuntu](#)
 - [The openjdk-7 package in Ubuntu](#)

2.13 Reliability, availability, and serviceability

The Power AC922 server brings POWER9 processor and memory reliability, availability, and serviceability (RAS) functions into a cloud data center, with open source Linux technology supplying the operating system (OS) and virtualization. The OPAL firmware provides a hypervisor and OS-independent layer that uses the error detection and self-healing functions that are built into the POWER9 processor.

The processor address paths and data paths, the control logic, state machines, and computational units are protected with parity or error-correcting code (ECC). The processor core soft errors or intermittent errors are recovered with processor instruction retry. Unrecoverable errors are reported as machine check (MC) errors, and errors that affect the integrity of data lead to a system checkstop.

The Level 1 (L1) data and instruction caches in each processor core are parity-protected, and data is stored to L2 cache immediately. L1 caches have a retry capability for intermittent errors and a cache set delete mechanism for handling solid failures. The L2 and L3 caches in the POWER9 processor are protected with double-bit detect, single-bit correct ECC.

In addition, a threshold of correctable errors that are detected on cache lines can result in the data in the cache lines being purged and the cache lines removed from further access without requiring a restart. An uncorrectable error that is detected in these caches can also trigger a purge and delete of cache lines. This purge and delete does not impact the current operation if the cache lines contained data that is unmodified from what was stored in the system memory.

The memory subsystem has proactive memory scrubbing to help prevent the accumulation of multiple single-bit errors. The ECC scheme can correct the complete failure of any one memory module within an ECC word. After marking the module as unusable, the ECC logic can still correct single symbol (two adjacent bit) errors. An uncorrectable error of data of any layer of cache up to the main memory is marked to prevent usage of fault data. The processor's memory controller has retry capabilities for certain fetch and store faults.

2.13.1 Error handling

This section describes various types of error handling.

Special Uncorrectable Error handling

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

Thermal management and current/voltage monitoring

The On Chip Controller (OCC) monitors various temperature sensors in the processor module, memory modules, and environmental temperature sensors, and directs the throttling of processor cores and memory channels if the temperature rises over thresholds that are defined by the design. The power supplies have their own independent thermal sensors and monitoring. Power supplies and voltage regulator modules monitor over-voltage, under-voltage, and over-current conditions. They report into a power good tree that is monitored by the Service Processor.

PCI enhanced error handling

PCI enhanced error handling (EEH)-enabled adapters respond to a special data packet that is generated from the affected PCI slot hardware by calling system firmware, which examines the affected bus, allows the device driver to reset it, and continues without a system restart. For Linux, EEH support extends to many devices, although some third-party PCI devices might not provide native EEH support.

Graphics processing unit acceleration

GPUs are attached with second-generation NVLink to the system POWER processors and provide cache coherence capabilities. The GPUs should be run in compute mode, which enables Single Error Correction and Double Error Detection (SECDEC) ECC for the GPU memory, SM register file, L1 cache, and L2 cache to improve data integrity for GPU-accelerated workloads.

Input power loss and auto restart after system-check-stop

The boot parameter “chassis policy” controls whether the server returns to the previous state or powers up axiomatically after an input power loss. The system automatically restarts after a system checkstop, and it is up to the system management software to decide whether to use the server with potentially fewer resources.

2.13.2 Serviceability

The server is designed for system installation and setup, feature installation and removal, proactive maintenance, and corrective repair that is performed by the client. Warranty Service Upgrades are offered for an OnSite Repair (OSR) by an IBM Support Services Representative (IBM SSR) or an authorized warranty service provider.

IBM Knowledge Center provides up-to-date documentation to effectively service the system:

- ▶ *Quick Install Guide*
- ▶ *User's Guide*
- ▶ *Troubleshooting Guide*
- ▶ *Boot Configuration Guide*

The documentation can be downloaded in PDF format or used online with an internet connection. For more information, see [IBM Knowledge Center](#).

Service Processor

The Service Processor supports the IPMI 2.0 and Data Center Management Interface (DCMI) V1.5 and Simple Network Management Protocol (SNMP) V2 and V3 for system monitoring and management. The Service Processor provides platform system functions such as power on/off, power sequencing, power fault monitoring, power reporting, fan/thermal control, fault monitoring, vital product data (VPD) inventory collection, Serial over LAN (SOL), Service Indicator LED management, code update, and event reporting through system event logs (SELS).

All SELs can be retrieved either directly from the Service Processor or from the host OS (Linux). The Service Processor monitors the operation of the firmware during the boot process and also monitors the hypervisor for termination. The firmware code update is supported through the Service Processor and IPMI interface. The firmware image can be updated or flashed regardless of its current state.

Service interface

The service interface enables the client and the support personnel to communicate with the service support applications in a server by connecting directly or remotely through a web browser or command-line interface (CLI). It provides access to various service applications and available actions. The service interface enables client and support personnel to manage system resources, inventory, and service information in an efficient and effective way.

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

- ▶ Service processor: Ethernet Service Network with IPMI Version 2.0, systems management GUI through web browser
- ▶ Service indicator LEDs: System attention and system identification (front and back)
- ▶ Host OS: CLI

The primary service applications are:

- ▶ SELs
- ▶ OS event logs
- ▶ Sensor status GUI LEDs for Problem Determination (PD) when next to the system, locally

Concurrent maintenance

The following components can be replaced without powering off the server:

- ▶ HDDs
- ▶ Fans

Error handling and reporting

If there is a system hardware failure or environmentally induced failure, the system error capture capability systematically analyzes the hardware error signature to help determine the cause of failure. The processor and memory recoverable errors are handled through Processor Runtime Diagnostics (PRD) in the OPAL layer and generate a SEL. An extended SEL (eSEL) is associated with each SEL. It contains extra first failure data capture (FFDC) information for use by the support structure.

For system checkstop errors, the OCC collects Failure Information Register (FIR) data and saves it in nonvolatile memory. PRD analyzes the data upon restart and creates a SEL and eSEL. The host Linux OS can monitor the SELs on the Service Processor through the IPMI tool. Hardware and firmware failures are recorded in the SELs and can be retrieved through the IPMI interface. The system can report errors that are associated with PCIe adapters/devices through the host OS.



Physical infrastructure

The objective of this section is to summarize all the physical infrastructure requirements regarding the IBM Power System AC922 servers.

For more information, see [IBM Knowledge Center for the Power AC922 server](#).

3.1 Operating environment

Table 3-1 provides the operating environment specifications for the Power AC922 server model GTH with zero, two, or four graphics processing units (GPUs) installed with 32 Gb or 16 Gb GPUs.

Table 3-1 Operating environment for Power AC922 server model GTH

Server operating environment			
Description	Recommended operating	Allowable operating	Non-operating
Temperature ^{ab}	18 - 27 °C (64 - 80.6°F)	5 - 40°C (41 - 104°F)	1 - 60°C (34 - 140°F)
Humidity range	5.5°C (42°F) dew point (DP) to 60% relative humidity (RH) and 15°C (59°F) dew point	-12°C *19.4°F) DP and 8% - 80% RH	8 - 80% RH
Maximum dew point		24°C (75° F)	27°C (80°F)
Maximum operating altitude		3050 m (10,000 ft.)	
Operating voltage		200 - 240 V AC	
Operating frequency		50 - 60 Hz +/- 3 Hz	
Power consumption		2300 watts maximum	
Power source loading		2.6 kVA maximum	
Thermal output		8872 BTU/hr maximum	
Noise level and sound power		7.6/6.7 bels operating/idling	

- a. Derate maximum allowable dry-bulb temperature 1°C (1.8°F) per 175 m above 950 m. IBM recommends a temperature of 18°C - 27°C (64°F - 80.6°F).
- b. For model GTH, heavy workloads might see performance degradation if internal temperatures result in a central processing unit (CPU) or graphics processing unit (GPU) clock reduction.

Table 3-2 provide the operating environment specifications for the Power AC922 server model GTX with four or six GPUs installed with 32 Gb or 16 Gb GPUs.

Table 3-2 Operating environment for Power AC922 server model GTX

Server operating environment			
Description	Recommended operating	Allowable operating	Non-operating
Temperature ^a	18 - 27 °C (64 - 80.6°F)	5 - 40°C (41 - 104°F)	1 - 60°C (34 - 140°F)
Humidity range	5.5°C (42°F) dew point (DP) to 60% relative humidity (RH) and 15°C (59°F) dew point	-12°C *19.4°F) DP and 8% - 80% RH	8 - 80% RH
Maximum dew point		24°C (75° F)	27°C (80°F)
Maximum operating altitude		3050 m (10,000 ft.)	

Server operating environment			
Description	Recommended operating	Allowable operating	Non-operating
Operating voltage		200 - 240 V AC	
Operating frequency		50 - 60 Hz +/- 3 Hz	
Power consumption		2300 watts maximum	
Power source loading		2.6 kVA maximum	
Thermal output		8872 BTU/hr maximum	
Noise level and sound power		7.6/6.7 bels operating/idling	

a. Derate maximum allowable dry-bulb temperature 1°C (1.8°F) per 175 m above 950 m. IBM recommends a temperature of 18°C - 27°C (64°F - 80.6°F).

Figure 3-1 shows the flow rate of water that is required based on the inlet temperature of the water to the rack for a single system.

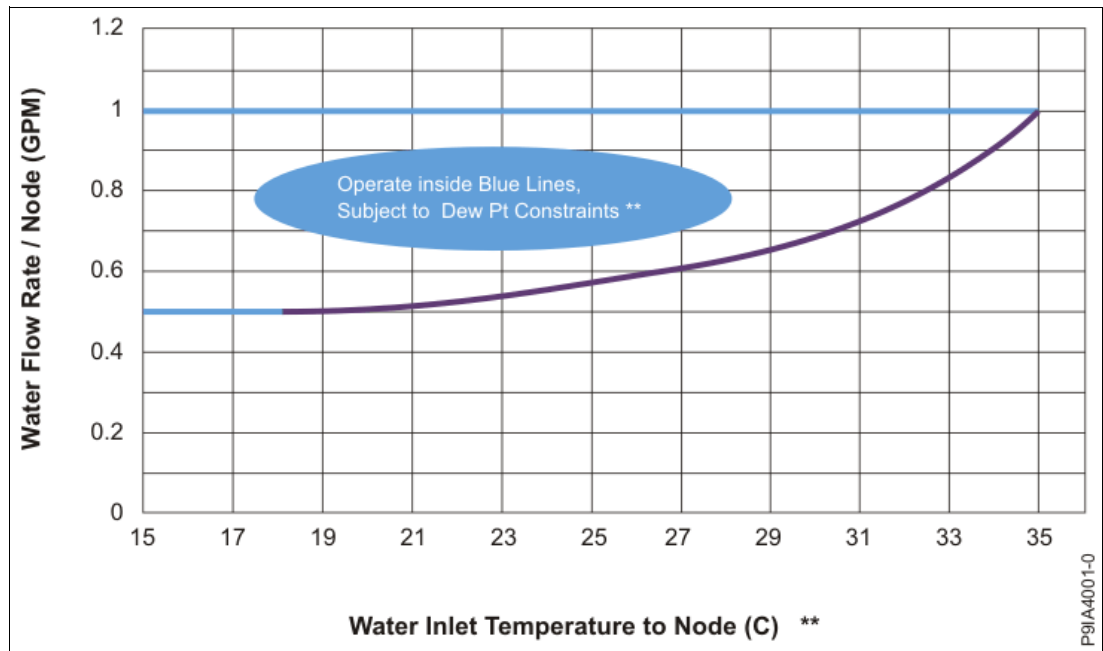


Figure 3-1 Water flow rate versus temperature

Note: The customer must constantly monitor the dew point and adjust the water temperature if necessary. The water temperature must *always* be above the dew point so that there is never a chance of condensation.

Figure 3-2 provides data about the water flow versus pressure drop as a function of the number of systems in a rack. The facility rack-level pressure drop includes the following pressure drops:

- ▶ Supply-side Eaton ball valve quick-connect pair
- ▶ Supply-side 1-in. ID, 6-ft. long hose going to the supply manifold
- ▶ Supply-side manifold
- ▶ 8335-GTX nodes
- ▶ Return-side manifold
- ▶ Return-side 1-inch ID, 6-ft. long hose leaving the return manifold
- ▶ Return-side Eaton ball valve quick-connect pair

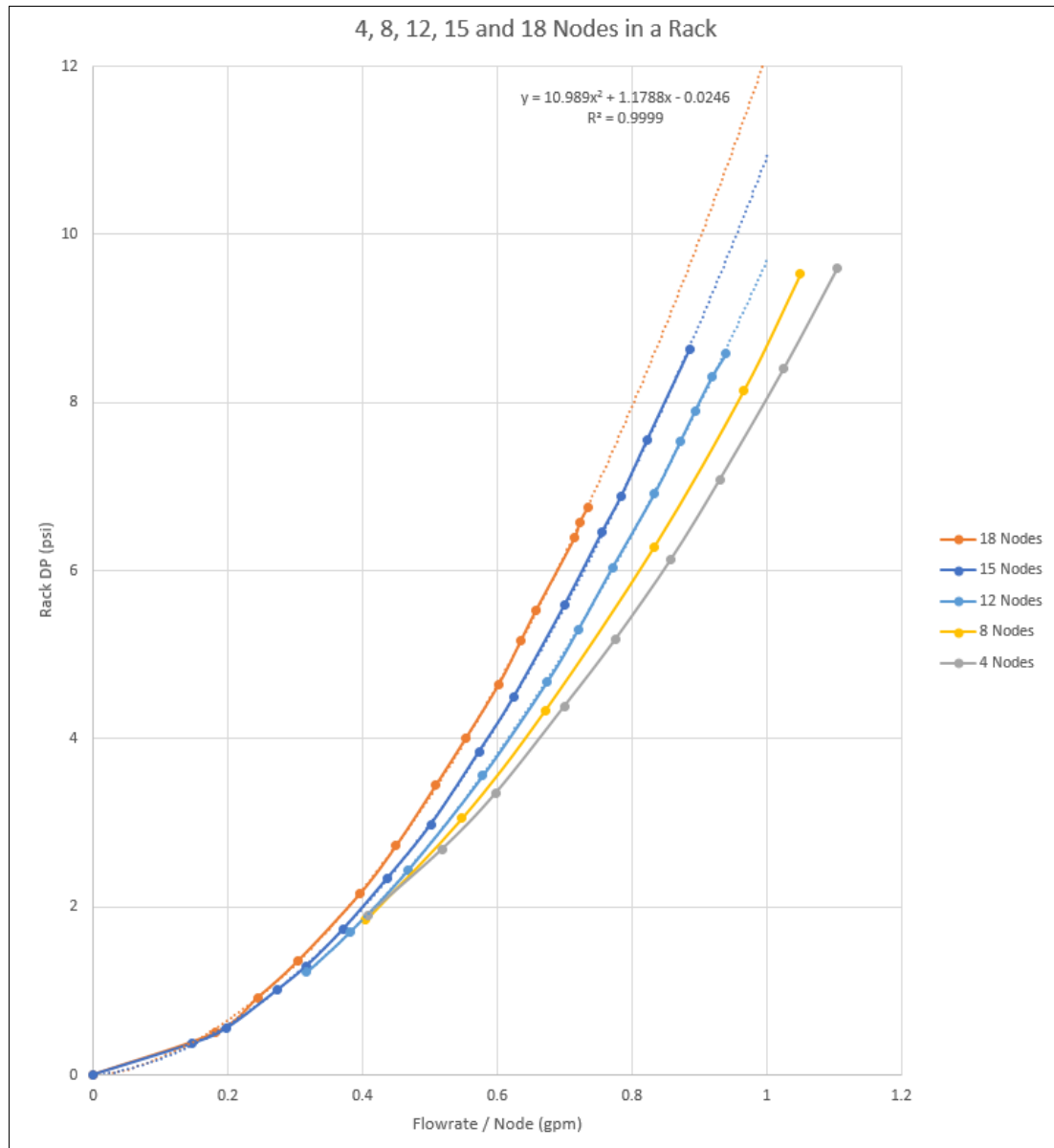


Figure 3-2 Water flow rate versus pressure drop

3.1.1 Leak detection

No leak detection is included in the system. You should have facility-level leak sensors or detectors as a preventive measure.

3.1.2 Water pressure

The rack manifold cannot exceed 40 PSI water pressure upon entrance to the rack during normal operating conditions. In a worst case, single-fault condition, the maximum pressure cannot exceed 55 PSI.

3.2 Physical package

Table 3-3 shows the physical dimensions of the chassis. The server is available only in a rack-mounted form factor and requires 2U (2 EIA units) of rack space.

Table 3-3 Physical dimensions for the Power AC922 model GTH and GTX server

Dimension	Power AC922 server models 8335-GTH and 8335-GTX
Width	441.5 mm (17.4 in.)
Depth	845.0 mm (33.3 in.)
Height	86.0 mm (3.4 in.)
Weight (maximum configuration)	30 kg (65 lbs.)

3.3 System power

The Power AC922 server is powered by two 2200 W power supplies at the rear of the unit.

The power supplies provide redundancy if a power supply failure occurs. GPUs are the most power-consuming devices in the server. Depending on the configuration and utilization, if a power supply failure occurs then throttling might happen when six GPUs are installed. In this case, the system remains operational, but might experience reduced performance until the power supply is replaced.

The power supplies on the server use a Rong Feng 203P-HP connector. A new power cable to connect the power supplies to the power distribution units (PDUs) in the rack is required, rendering the reuse of existing power cables not viable. The PDU connector type (IEC C20 or IEC C19) depends on the selected rack PDU.

Figure 3-3 shows an example of the power cable with its connectors.

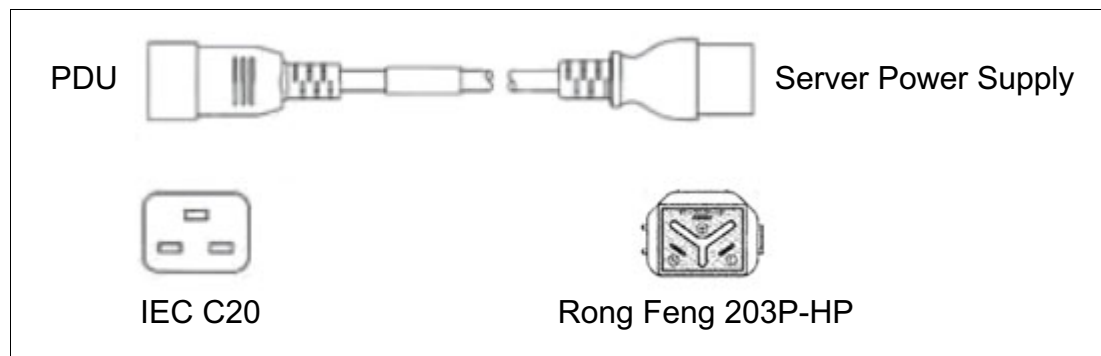


Figure 3-3 Power AC922 power cables with the Rong Feng connector

Both 1-phase and 3-phase PDUs are supported. For more information, see 3.5.2, “AC power distribution units” on page 59.

When opting for 3-phase 60A PDUs, a total of 4 PDUs are required to support a full rack with 18 Power AC922 servers that are configured with four GPUs. If 1-phase PDUs are selected, a minimum of five PDUs are required to support a full rack of 18 Power AC922 servers with a four-GPU configuration. If the 1-phase PDUs are limited to 48A, no more than four Power AC922 servers can be connected to a single PDU.

3.4 System cooling

Air or water cooling depends on the model of the server and the GPUs feature codes (FCs) that are selected. For a list of available GPUs, see 2.5.5, “Compute-intensive accelerators” on page 34.

Rack requirement: The IBM 7965-S42 rack with #ECR3 or #ECR4 installed supports the water-cooling option for the Power AC922 server (see “Optional water cooling” on page 56).

When using water-cooled systems, the customer is responsible for providing the system that supplies the chilled and conditioned water to the rack. Water condensation can occur with certain combinations of temperature and relative humidity, which define the dew point. The system that supplies the cooling water must be able to measure the room dew point and automatically adjust the water temperature several degrees above dew point. Otherwise, the water temperature must be above the maximum dew point for that data center installation. Typical primary chilled water is too cold for use in this application because building-chilled water can be as cold as 4°C - 6°C (39°F - 43°F).

In air-cooled systems (8335-GTH), all components are air-cooled, including processors and GPUs that use heat sinks.

Figure 3-4 shows the internal view of the server with two processors and the four GPUs heat sinks that are installed.

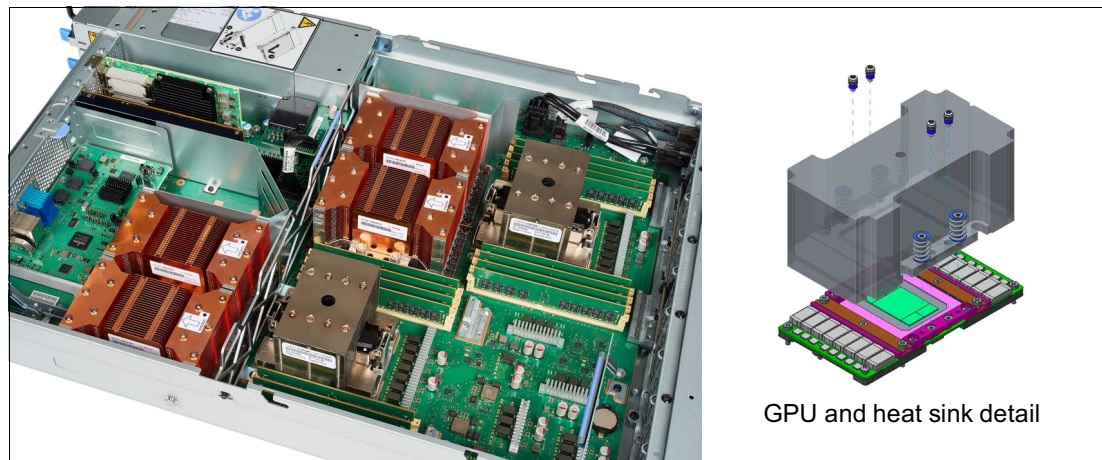


Figure 3-4 Power AC922 air-cooled model internal view

In the water-cooled model GTX, the processors and GPUs are cooled with a water system that moves water between the CPUs and GPUs coldplates. Other components, such as memories, Peripheral Component Interconnect Express (PCIe) adapters, and power supplies, are cooled by using traditional air-cooling systems. Coldplates to cool two processor modules and up to six GPUs are included. Water lines carrying cool water in and warm water out are also included. This feature is installed in the system unit when the server is manufactured and is not installed in the field.

When ordering the Power AC922 model GTX, a cooling kit is required. It contains the pipes, coldplates, and splitters that are required to cool the system. #EJ31 provides the internal cooling system of the server for a 4-GPU configuration, and #EJ34 is for a 6-GPU configuration, as shown in Figure 3-5.

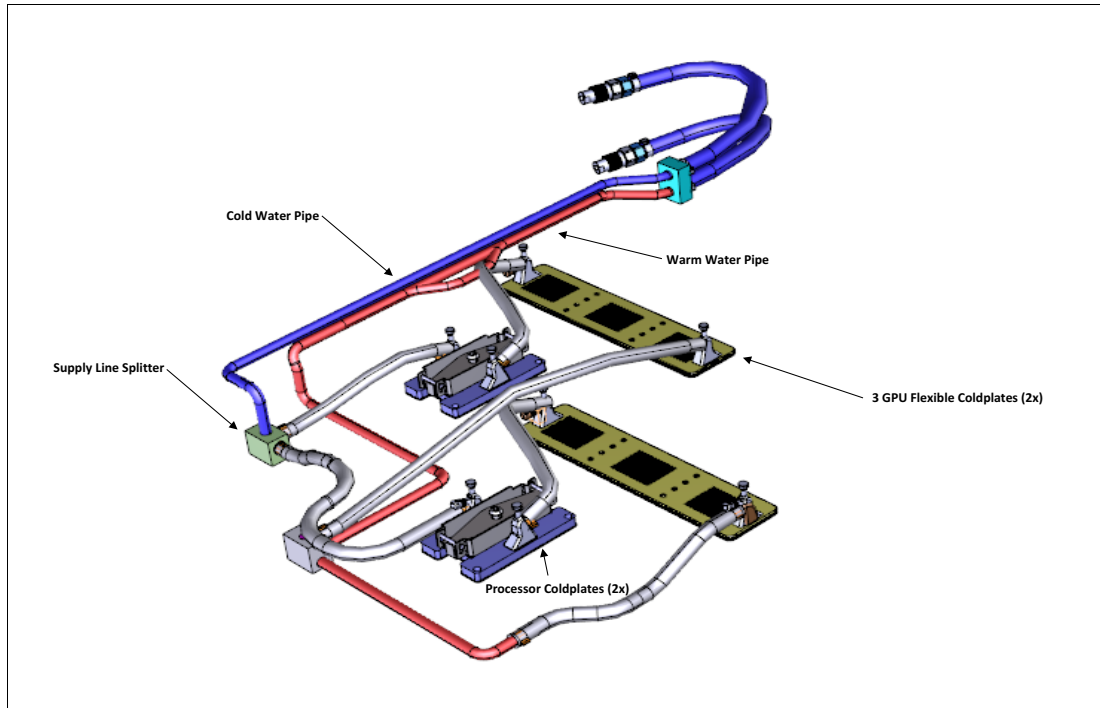


Figure 3-5 Internal cooling components for the 8335-GTX server

Figure 3-6 shows a view of the cooling system that is installed in the server.

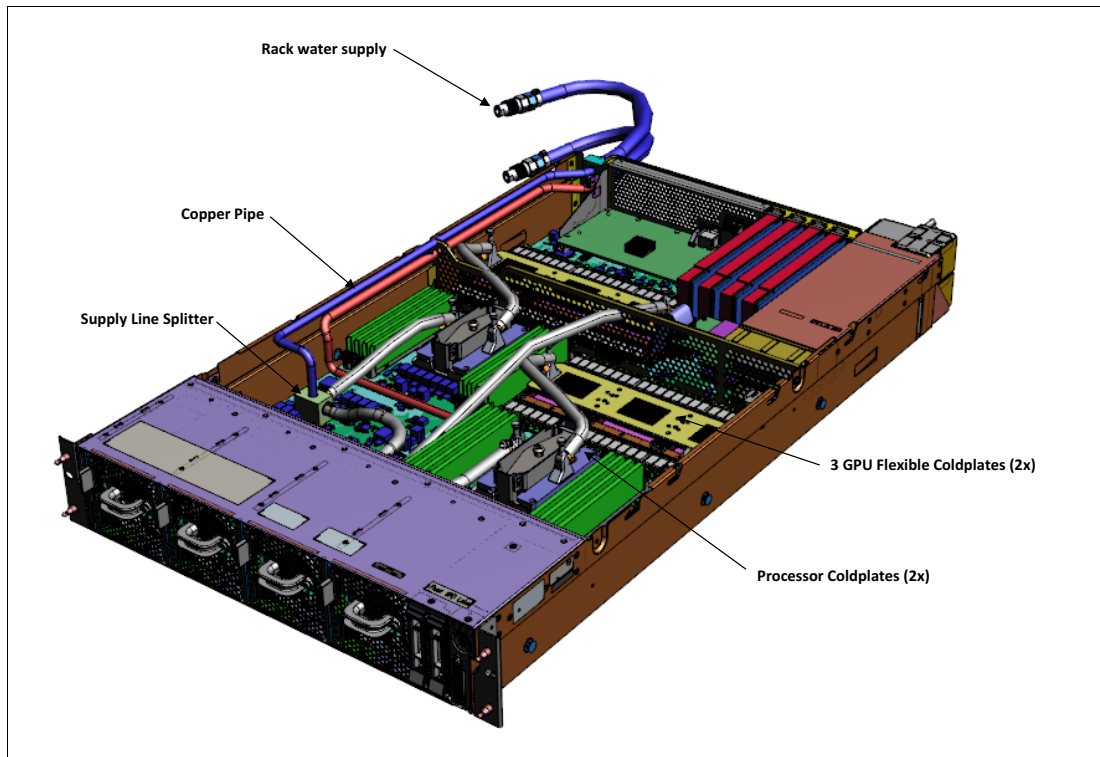


Figure 3-6 Internal cooling that is installed in an 8335-GTX model server

Figure 3-7 shows a detailed view of a processor and three-GPU cooling system.

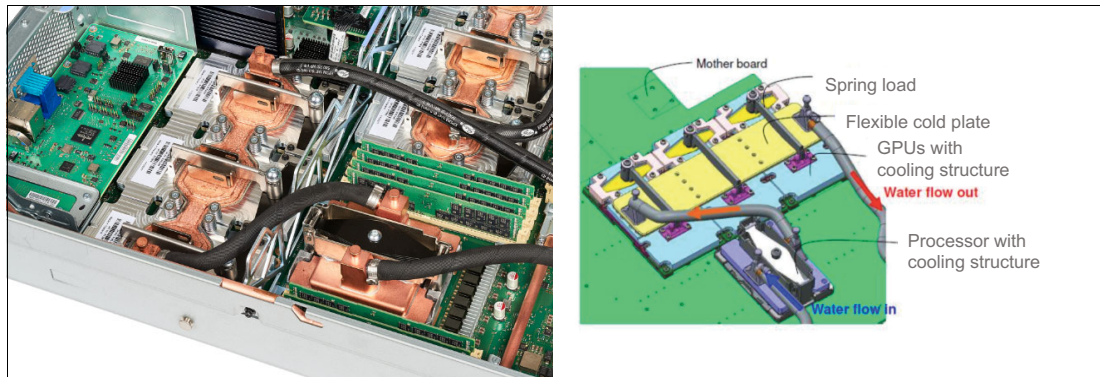


Figure 3-7 Processor and GPU water-cooling details

Water enters the system and passes through a splitter block, where the water goes to two different flow paths. In each flow path, the water flows first through the CPU coldplate and then through the GPU coldplate. Then, the warm water enters a return line splitter block and goes out of the server.

Figure 3-8 shows the cold water in blue and warm water in red.

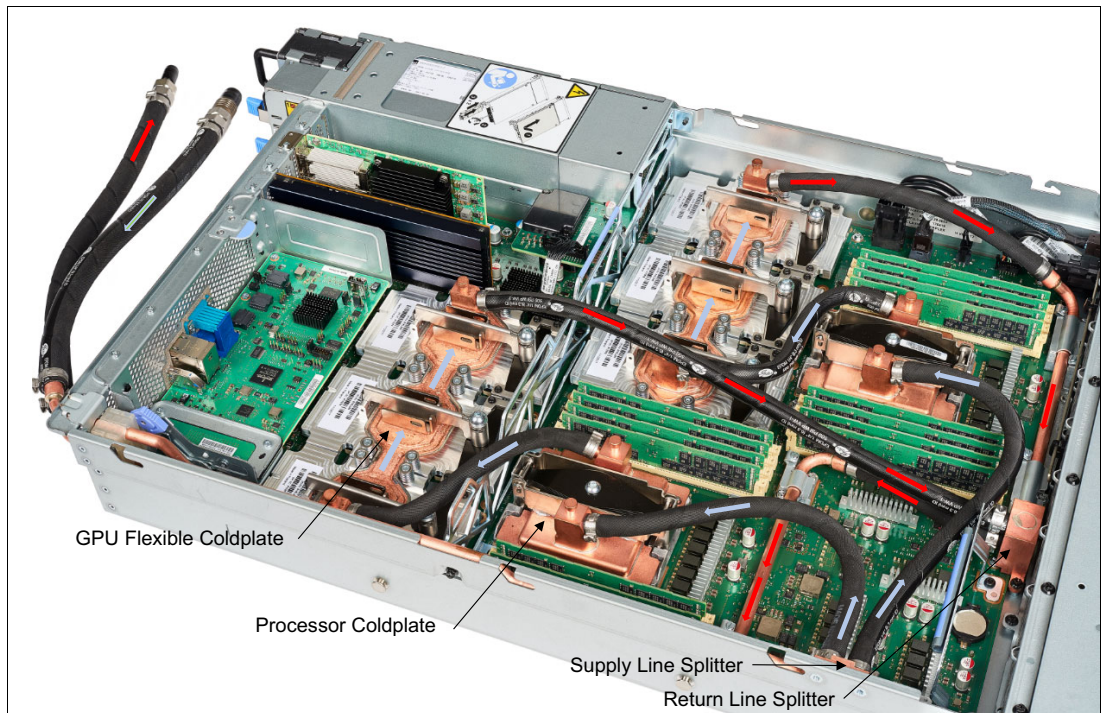


Figure 3-8 Cold and warm water flow through the Power AC922 system

When shipped from IBM, an air-cooled server cannot be changed into a water-cooled server, and a water-cooled server cannot be changed into an air-cooled server.

Important: You must provide a 1-inch ID barb fitting to attach your facility to the hose kit for each hose. Only clean, filtered, and chemically treated water must be used, *not generic building water*.

The GPU air-cooled and water-cooled servers have the following ordering differences:

- ▶ Air-cooled server model GTH can be ordered with zero, two, or four GPUs.
- ▶ Water-cooled server model GTX can be ordered with four or six GPUs.

Note: The Power AC922 model GTX server offers only the fixed rail kit option. Ordering this model with slide rails is not supported. Maintenance of components other than power supplies and fans must be done on a bench with the server unplugged from the cooling system.

For more information about the water-cooling option, see [IBM Knowledge Center](#).

3.5 Rack specifications

Depending on the selected model for the Power AC922 server, there are different supported racks that are available. Although the air-cooled model GTH supports many racks, water-cooled options have just one rack that is supported, as shown in Table 3-4.

Table 3-4 Supported racks by model

Type/Model	Description	Supported by 8335-GTH	Supported by 8335-GTX
7014-T00	IBM 7014 Rack Model T00	Yes	No
7014-T42	IBM 7014 Rack Model T42	Yes	No
7014-S25	IBM Entry Rack Cabinet Model S25	Yes	No
7965-94Y	IBM 42U Slim Rack ^a	Yes	No
7965-S42	IBM Enterprise Slim Rack	Yes	Yes
	Original equipment manufacturer (OEM) 19-inch rack	See 3.5.4, "Original equipment manufacturer racks" on page 62.	No

a. At the time of writing IBM plans to withdraw the 7965-94Y rack on January 17, 2020.

Note: Because of the water-cooling system, the server model GTX mounts only in the IBM 42U Slim Rack (7965-94Y) or the 42U IBM Enterprise Slim Rack (7965-S42).

These racks are built to the 19-inch EIA 310D standard.

Here are the requirements to integrate the Power AC922 server into the 19-inch rack:

- ▶ Rack integration is supported only on Rack MTM 7965-S42 (Constellation). Rack integration is not supported on rack MTM 7965-94Y (Railhawk) or MTM 7014-T00/T42 or Entry Rack S25.
- ▶ Left and right shipping brackets, P/N 02CL806 and 02CL807, must be installed at the rear of the Power AC922 drawer before shipment.
- ▶ The completed rack assembly must ship on the IBM shock pallet, P/N 00RP170.
- ▶ Normal drawer depopulation rules apply. Rack drawers should not be installed past rack EIA position 32 unless approved by IBM Safety.

If a system is installed in a rack or cabinet that is not an IBM rack, ensure that the rack meets the requirements that are described in 3.5.4, "Original equipment manufacturer racks" on page 62.

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.

3.5.1 IBM Enterprise Slim Rack 7965-S42

The new 2.0-meter (79-inch) Model 7965-S42 is compatible with past and present Power Systems servers and provides an excellent 19-inch rack enclosure for your data center.

This is a 19-inch rack cabinet that provides 42U of rack space for use with rack-mounted, non-blade servers, and I/O drawers. Its 600 mm (23.6 in.) width combined with its 1070 mm (42.1 in.) depth plus its 42 EIA enclosure capacity provides great footprint efficiency for your systems, and enables it to be easily placed on standard 24-inch floor tiles, enabling better thermal and cable management capabilities.

Another difference between the 7965-S42 model rack and the 7014-T42 model rack is that the “top hat” is on the 40U and 41U boundary instead of the 36U and 37U boundary in the 7014-T42 model.

The IBM PDUs are mounted vertically in four side bays, two on each side. After the side bays are filled, PDUs can be mounted horizontally at the rear of the rack. For more information about IBM PDUs, see 3.5.2, “AC power distribution units” on page 59.

To enable maximum airflow through the data center and the rack cabinets, filler panels are mounted at the front of the rack in empty EIA locations, and the rack offers perforated front and rear door designs.

Figure 3-9 shows the front view of the 7965-S42 rack.

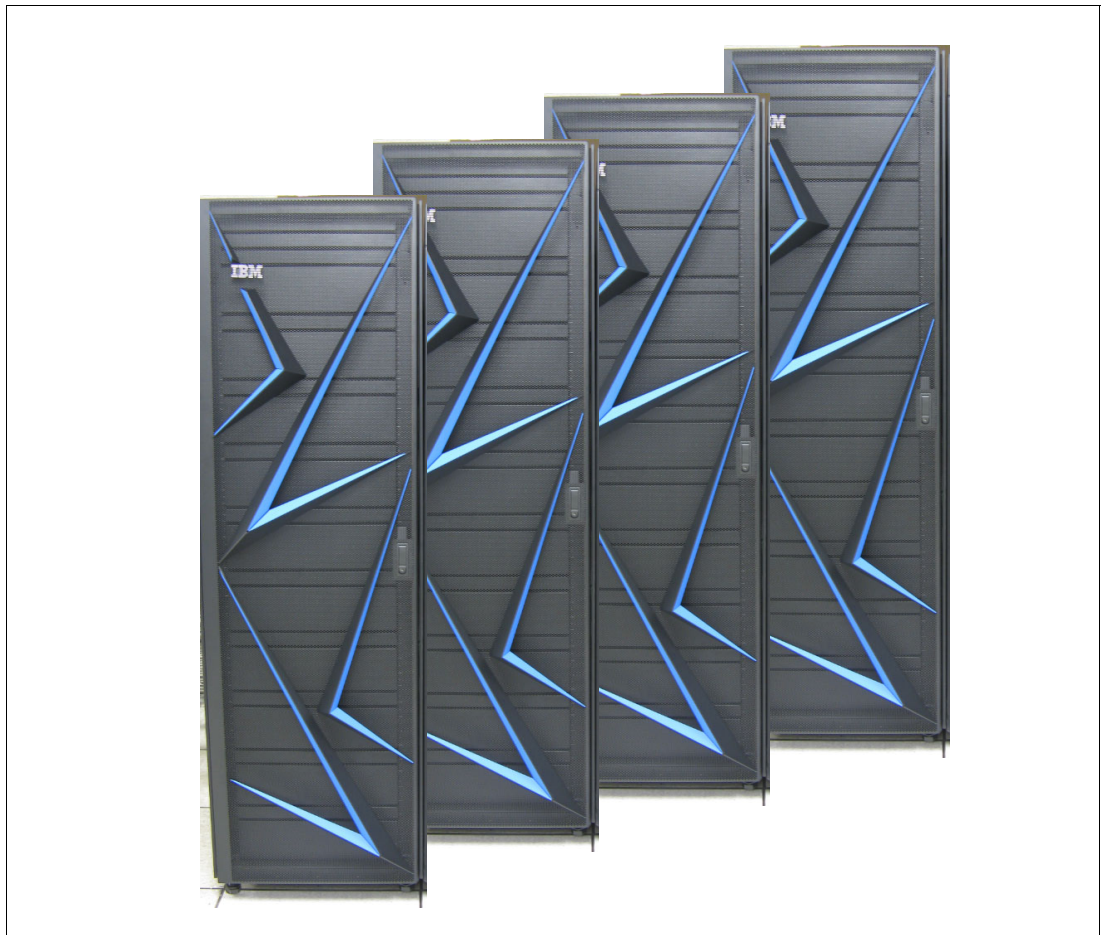


Figure 3-9 IBM 7965-S42 racks front view

Ballasts for more stability are available, so the 7965-S42 racks should not require the depopulate rules above the 32 EIA location, as required with 7014-T42 rack models.

Optional water cooling

When opting for model 8335-GTX of the Power AC922 server, water cooling is mandatory, so the 7965-S42 rack must be ordered with the water-cooling option (#ECR3 or #ECR4). There is no miscellaneous equipment specification (MES) for these features in the field.

These features represent a manifold for water cooling, and provide a water supply and water return for 1 - 20 servers that are mounted in a 7965-S42 Enterprise Slim Rack.

#ECR3 indicates the manifold with water input and output at the top of the rack.
Use #ECR4 to order the manifold with water input and output at the bottom of the rack.
Because the hose exits might require some space inside the rack, leave a 2U space vacant on the top or bottom of the rack, depending on the location of the hoses that you choose.

Figure 3-10 shows both options of the water input and output.

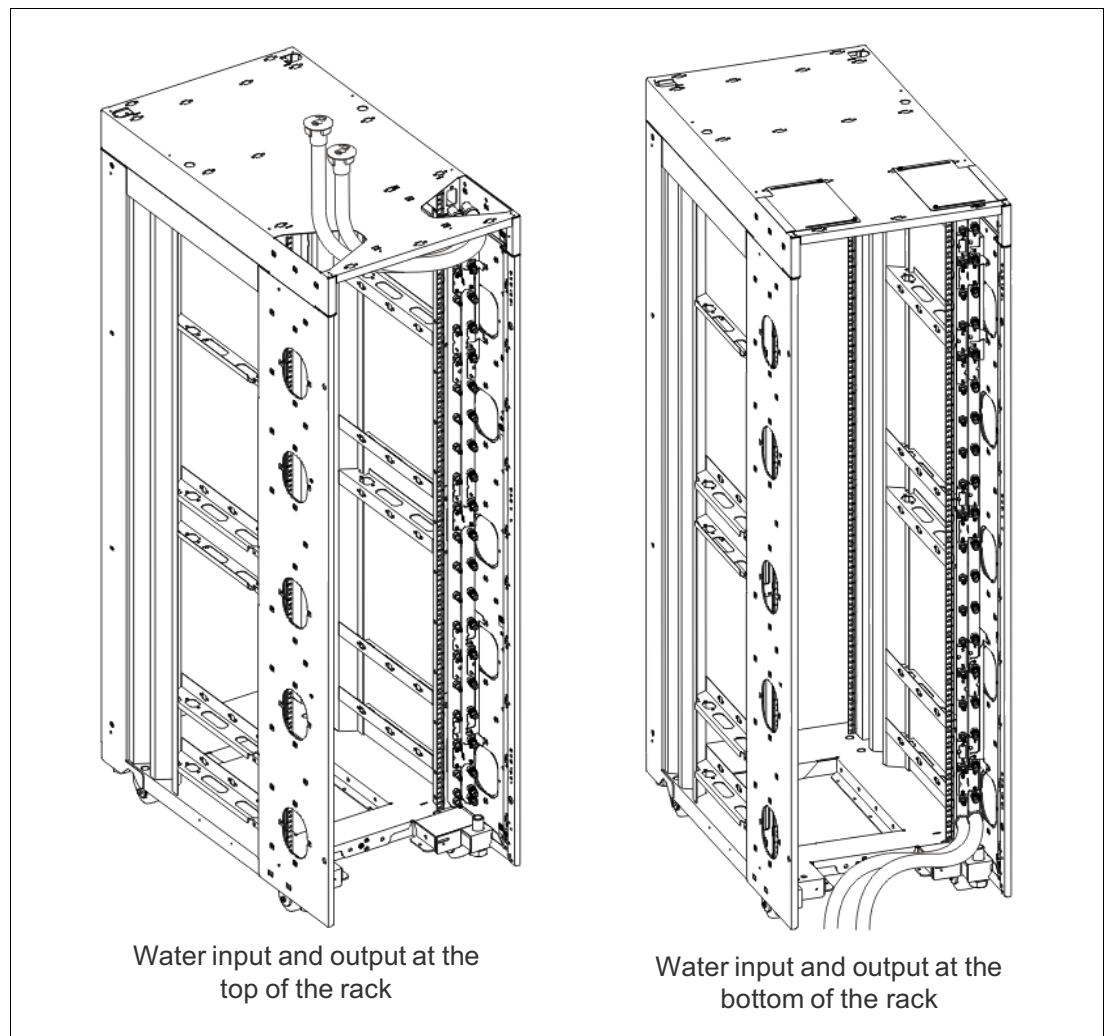


Figure 3-10 Top and bottom water input and output for the 7965-S42 rack

Figure 3-11 shows a data center rack row with the 7965-S42 racks with water input and output at the top of the rack.

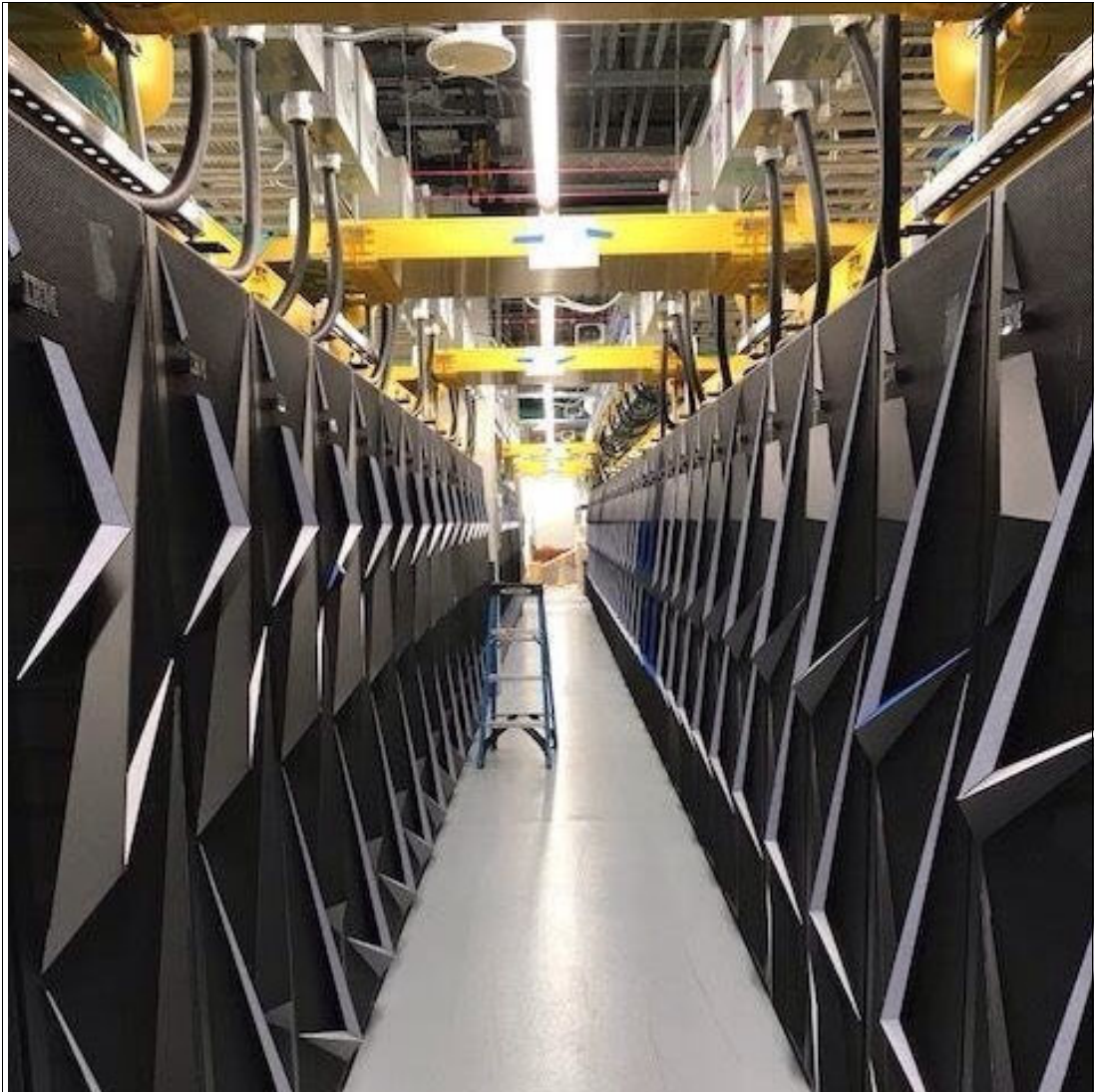


Figure 3-11 Data center rack row with water input and output at the top of racks

The manifold is mounted on the right side of the rack as viewed from the rear and extends for 40U. The manifold does not interfere with the placement of servers or other I/O drawers. Quick connect fittings are located every 2U on the manifold for water supply and return, which provides 20 pairs of fittings.

Figure 3-12 shows a manifold for the 7965-S42 rack.

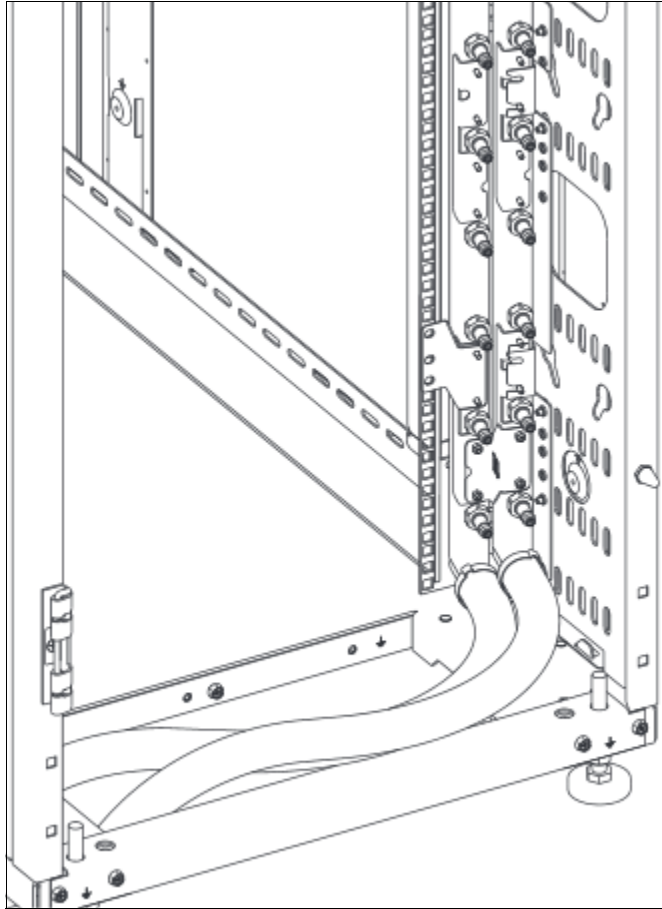


Figure 3-12 Manifold for the 7965-S42 rack

The servers are connected to the manifold through quick-connects. Supply and return hoses from the manifold to the server are provided as part the server cooling feature.

The manifold has one cold water inlet that leads to the rack and one warm water outlet. Two 4.25 m (14-foot) hose kits are provided with the manifold to connect water supply and return. The outer diameter of the hoses is approximately 34.5 mm (1.36 in.).

You must provide a 1-inch ID barb fitting to attach your facility to the hose kit for each hose. Only clean, filtered, and chemically treated water must be used, *not generic building water*.

For more information, see the IBM Knowledge Center at:

<https://www.ibm.com/support/knowledgecenter/POWER9/p9hdx/POWER9welcome.htm>

Important: Avoid vertically mounted PDUs on the right side as viewed from the rear of the rack. The manifold makes access to PDU impossible. Use either horizontally mounted PDUs, or use vertically mounted PDUs on the left side of the rack.

3.5.2 AC power distribution units

AC power distribution is fulfilled by PDUs, which include the AC single phase PDU #EPTG and the AC Intelligent single phase PDU+ #EPTJ. The Intelligent PDU+ is identical to #EPTG PDUs, but it is equipped with one Ethernet port, one console serial port for power monitoring. The AC three phase intelligent PDUs are #EPTK and #EPTL.

The PDUs have 9 client usable IEC 320-C19 outlets. This is an intelligent, switched 200-240 volt AC Power Distribution Unit (PDU). The PDU is mounted on the rear of the rack making the nine C19 receptacles easily accessible. Each receptacle has a 20 amp circuit breaker. Four PDUs can be mounted vertically in the back of the T00 and T42 racks.

Figure 3-13 shows the #EPTJ AC Intelligent Single-Phase PDU.



Figure 3-13 AC Intelligent Single-Phase PDU

In addition, the following high-function PDUs were announced in October 2019 and are supported on the model GTH server:

- ▶ High Function 9xC19 PDU plus (#ECJJ):

This is an intelligent, switched 200-240 volt AC Power Distribution Unit (PDU) plus with nine C19 receptacles on the front of the PDU. The PDU is mounted on the rear of the rack making the nine C19 receptacles easily accessible. For comparison, this is most similar to the earlier generation #EPTJ PDU.

- ▶ High Function 9xC19 PDU plus 3-Phase (#ECJL):

This is an intelligent, switched 208 volt 3-phase AC Power Distribution Unit (PDU) plus with nine C19 receptacles on the front of the PDU. The PDU is mounted on the rear of the rack making the nine C19 receptacles easily accessible. For comparison, this is most

Figure 3-14 on page 60 shows the placement of the four vertically mounted PDUs.

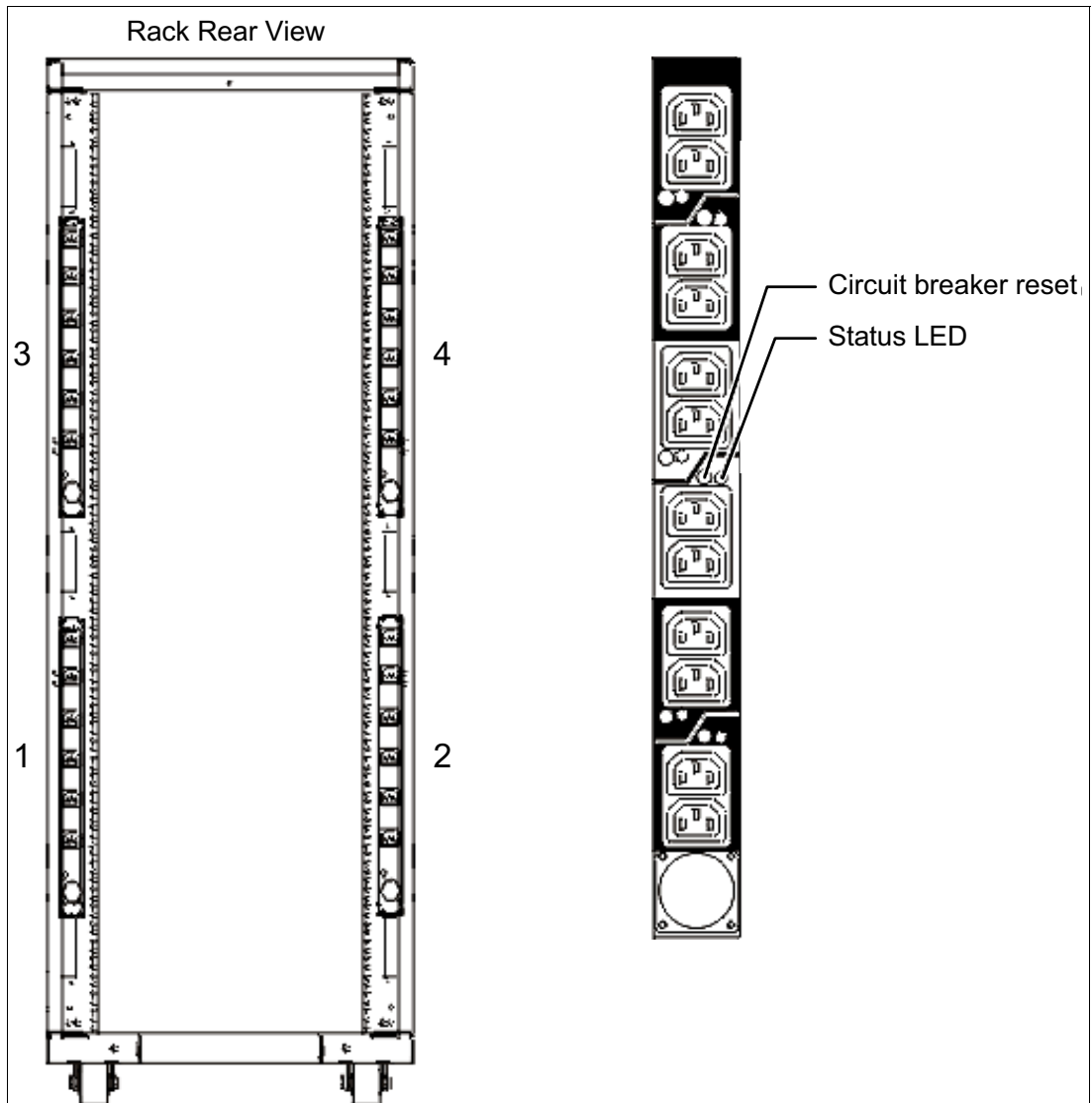


Figure 3-14 PDU placement and PDU view

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 that is available for other racked components. When mounting PDUs horizontally, the preferred practice is to use fillers in the EIA units that are occupied by these PDUs to facilitate the correct airflow and ventilation in the rack.

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

Table 3-5 shows the available wall power cord options for the PDU and iPDU features, which must be ordered separately.

Table 3-5 Wall power cord options for the PDU and iPDU features

Feature code	Wall plug	Rated voltage (V AC)	Phase	Rated amperage	Geography
6653	IEC 309, 3P+N+G, 16 A	230	3	16 amps/phase	Internationally available
6489	IEC309, 3P+N+G, 32 A	230	3	32 amps/phase	EMEA
6654	NEMA L6-30	200 - 208, 240	1	24 amps	US, Canada, LA, and Japan
6655	RS 3750DP (watertight)	200 - 208, 240	1	24 amps	US, Canada, LA, and Japan
6656	IEC 309, P+N+G, 32 A	230	1	24 amps	EMEA
6657	PDL	230 - 240	1	32 amps	Australia and 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 correct 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 of the drawers that are plugged into the PDU must not exceed the power cord limitation.

The Universal PDUs are compatible with previous models. To better enable electrical redundancy, each server has 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 [IBM Knowledge Center](#).

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

3.5.3 Rack-mounting rules

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

- ▶ The system can 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.

3.5.4 Original equipment manufacturer racks

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

IBM Knowledge Center mentions the following key points:

- ▶ The front rack opening must be 450 mm wide ± 0.75 mm (17.72 in. ± 0.03 in.).

Figure 3-15 is a top view that shows the specification dimensions.

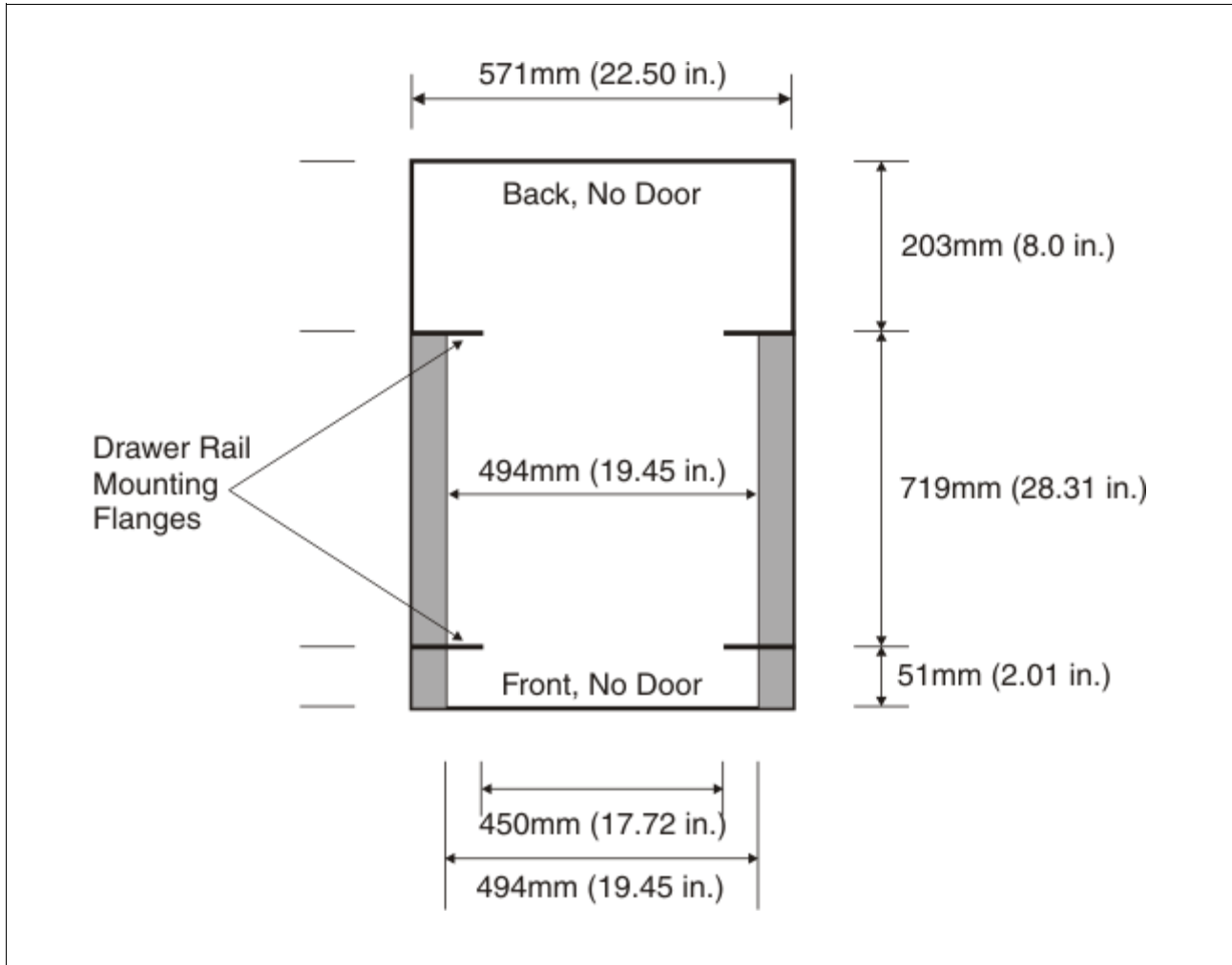


Figure 3-15 Top view of rack specification dimensions (not specific to IBM)

- ▶ The rail-mounting holes must be $465 \text{ mm} \pm 0.8 \text{ mm}$ ($18.3 \text{ in.} \pm 0.03 \text{ in.}$) apart on-center (horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges), as shown in Figure 3-16.

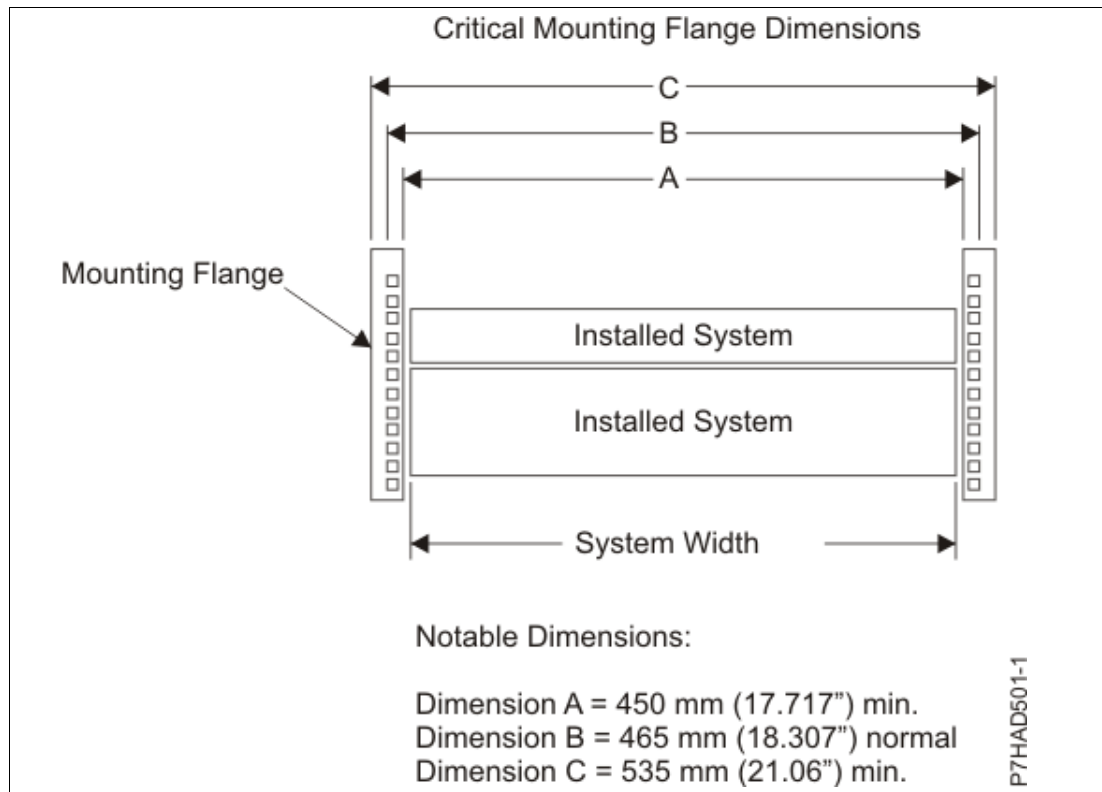


Figure 3-16 Mounting flange dimensions

- ▶ 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, which makes 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 3-17 shows the top front specification dimensions.

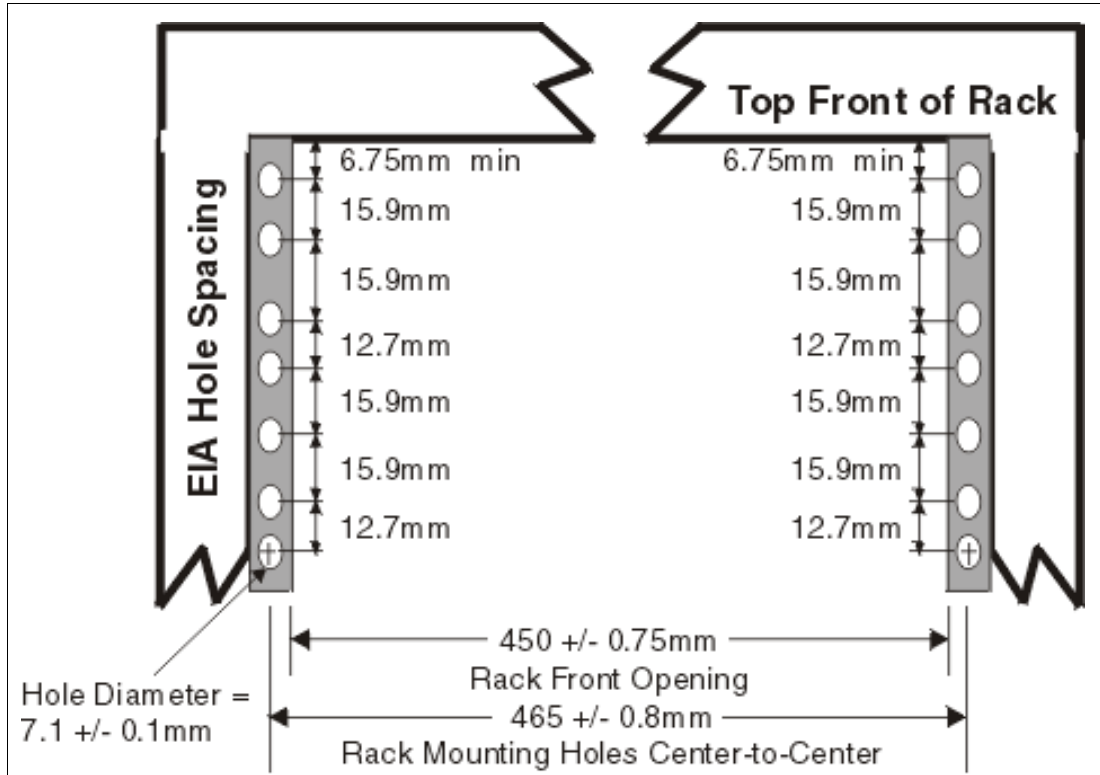


Figure 3-17 Rack specification dimensions top front view

- ▶ A minimum rack opening width of 500 mm (19.69 in.) for a depth of 330 mm (12.99 in.) is needed behind the installed system for maintenance, service, and cable management. The recommended depth is at least 254 mm (10 in.) within the rack from the rear rack mount flange to the frame line, as shown in Figure 3-18.

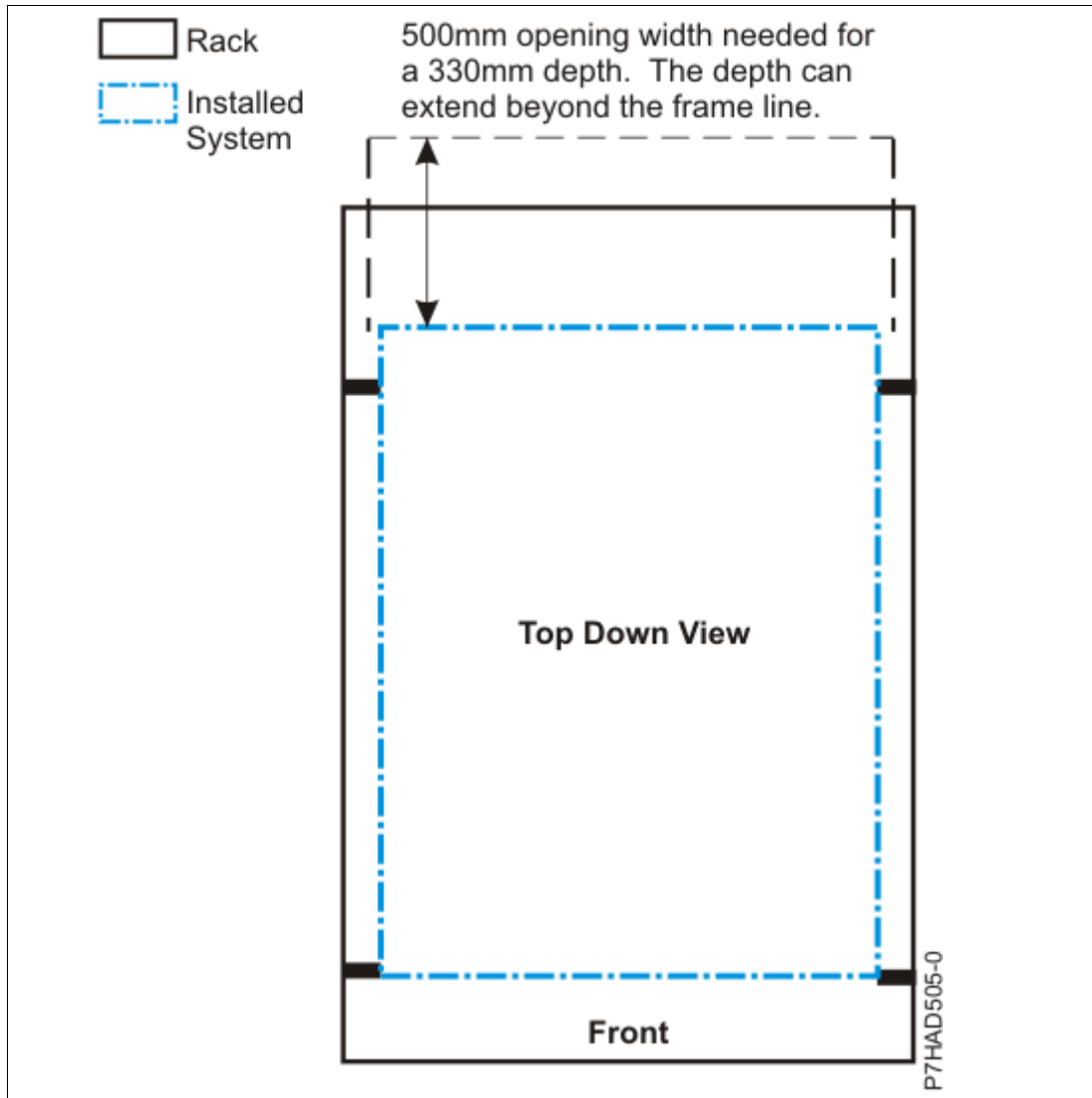


Figure 3-18 OEM rack opening depth



IBM PowerAI

Data, in all forms, is expanding as a resource to be used. In many industries, the data explosion is outstripping the human capacity to understand the meaning that is hidden within that data.

Cognitive computing brings value to your organization's data by using an entirely new approach to problems: Using deep learning (DL), machine learning, and artificial intelligence (AI) to reason and act upon data that could not be used until now, such as structured and unstructured data, images, videos, and sounds.

DL consists of algorithms that permit software to train itself by exposing multilayered neural networks to vast amounts of data. It is most frequently used to perform tasks such as speech and image recognition, but theoretically can be used on any data.

The intelligence in the process sits within the DL software frameworks themselves, which develop that neural model of understanding by building weights and connections between many, many data points, often millions in a training data set.

To ease the process of installation, configuration, and adoption of DL, IBM Cognitive Systems created an offering that is called IBM PowerAI. PowerAI brings a suite of capabilities from the open source community and combines them into a single enterprise distribution of software. Incorporating complete lifecycle management from installation and configuration, data ingest and preparation, building, optimizing, and training the model, to inference, testing, and moving the model into production. Taking advantage of a distributed architecture, PowerAI can help enable your teams to quickly iterate through the training cycle on more data to help continuously improve the model over time.

PowerAI offers many optimizations that can ease installation and management, and accelerate performance:

- ▶ Ready-to-use DL frameworks (Tensorflow and IBM Caffe).
- ▶ Distributed as easy to install binary files.
- ▶ Includes all dependencies and libraries.
- ▶ Easy updates: Code updates arrive through a repository.
- ▶ Validated DL platform with each release.

- ▶ Dedicated support teams for DL.
- ▶ Designed for enterprise scale with multisystem cluster performance and large memory support.
- ▶ Supported by IBM Power Systems servers with NVLink attached CPUs and NVIDIA graphics processing units (GPUs).

PowerAI brings a set of unique and innovative tools to allow for companies to adopt cognitive computing, such as PowerAI Vision, IBM PowerAI Distributed Deep Learning (DDL), and large model support.

Deep learning frameworks requirements

IBM PowerAI V5.1 has the following requirements for the operating system (OS):

- ▶ RHEL 7.5 (architecture: ppc64le).
- ▶ NVIDIA driver Version 396.26 or higher is required.
- ▶ NVIDIA CUDA Version 9.2 or higher is required.
- ▶ NVIDIA cuDNN 7.1 is required.
- ▶ Anaconda 5.1.0 is required

Note: For more information about these requirements, see [Deep Learning and IBM PowerAI](#).

IBM PowerAI Vision

Most of enterprises are facing challenges adopting AI techniques to understand and act upon their data. The most common challenges are:

- ▶ No experience on the development team on computer vision applications
- ▶ Lack of skills for deep neural networks (DNN) design and development
- ▶ Poor understanding about how to build a platform to support enterprise scale DL, including data preparation, training, and inference

IBM PowerAI Vision can solve these issues by training models to classify images and deploy models to infer real-time images, allowing users with little experience in DL to perform complex image analysis with virtually no coding necessary.

PowerAI Vision focus on optimizing the workflow that is required for DL, as shown in Figure A-1.

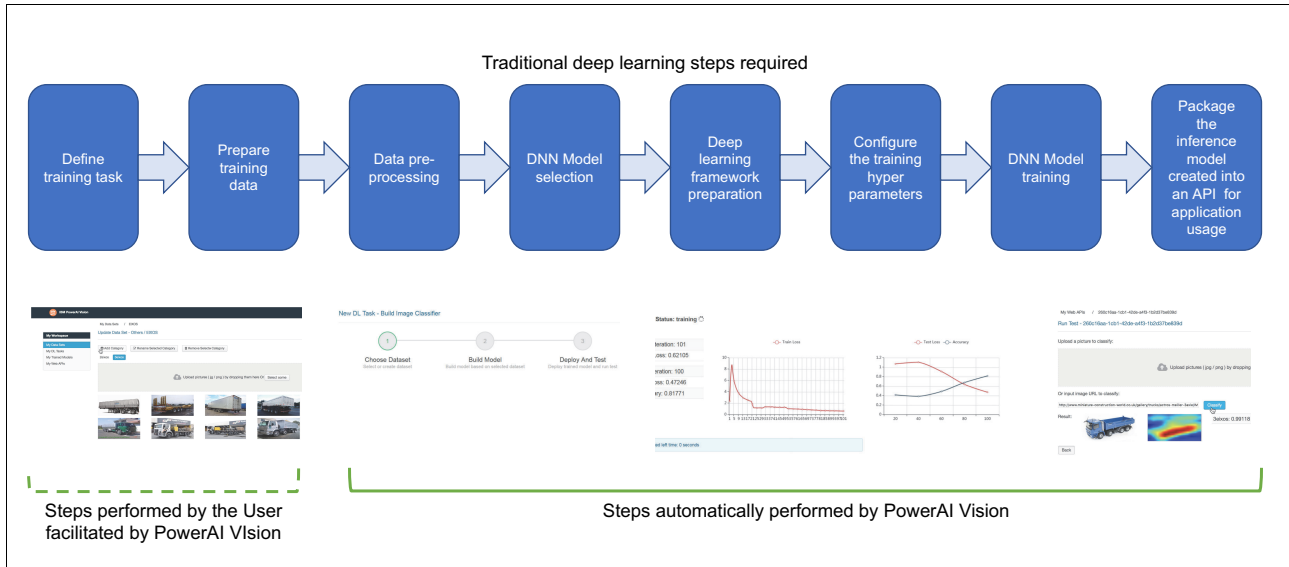


Figure A-1 Traditional deep learning steps that are required by PowerAI Vision

Usually, most of the time in DL is spent in preparing the data to be ingested by the system. Individual image analysis, categorization, and data transformation are some of the steps that are involved in this phase.

PowerAI Vision allows the user to prepare the data for DL training by using the following pre-processing features:

- ▶ Copy-bundled sample data sets to begin learning of PowerAI Vision.
- ▶ Create categories and load data sets for training by dragging images from local disks.
- ▶ Create labels, mark objects on images, and assign labels. Train models with the labeled objects.
- ▶ Optionally, import data sets for training in compressed formats.
- ▶ Preprocess images, including rotating, resizing, and cropping.

After the model is trained, PowerAI Vision allows the use of a simple GUI interface to trigger DL inference through APIs, deploying and engaging trained models to infer categories and detect occurrences of trained objects in test and real-time data sets.

Distributed Deep Learning

Many times, the amount of data that must be processed exceeds the capacity processing of a single server. In other cases, the training time might potentially be reduced by splitting the tasks among a cluster of servers.

To accelerate the time that is dedicated to training a model, the PowerAI stack uses new technologies to deliver exceptional training performance by distributing a single training job across a cluster of servers.

PowerAI DDL provides intelligence about the structure and layout of the underlying cluster (topology), which includes information about the location of the cluster’s different compute resources, such as GPUs, CPUs, and the data on each node.

PowerAI is unique in that this capability is incorporated into the DL frameworks as an integrated binary file, reducing the complexity for clients when they use high-performance clusters. As a result of this capability, PowerAI DDL can scale jobs across many cluster resources with little loss in communication.

Large model support

One of the challenges that customers face in the DL space is that they are limited by the size of memory that is available within GPUs.

Today, when data scientists develop a DL workload, that is, the structure of the matrixes in the neural model and the data, elements that train the model (in a batch) must sit within the memory on GPU. This situation is a problem because current GPUs memory is limited to 32 GB, which leads to even more time on preparation and modeling. Not all data can be analyzed at once, forcing data scientists to split the problem being solved into hundreds of smaller problems, as shown in Figure A-2.

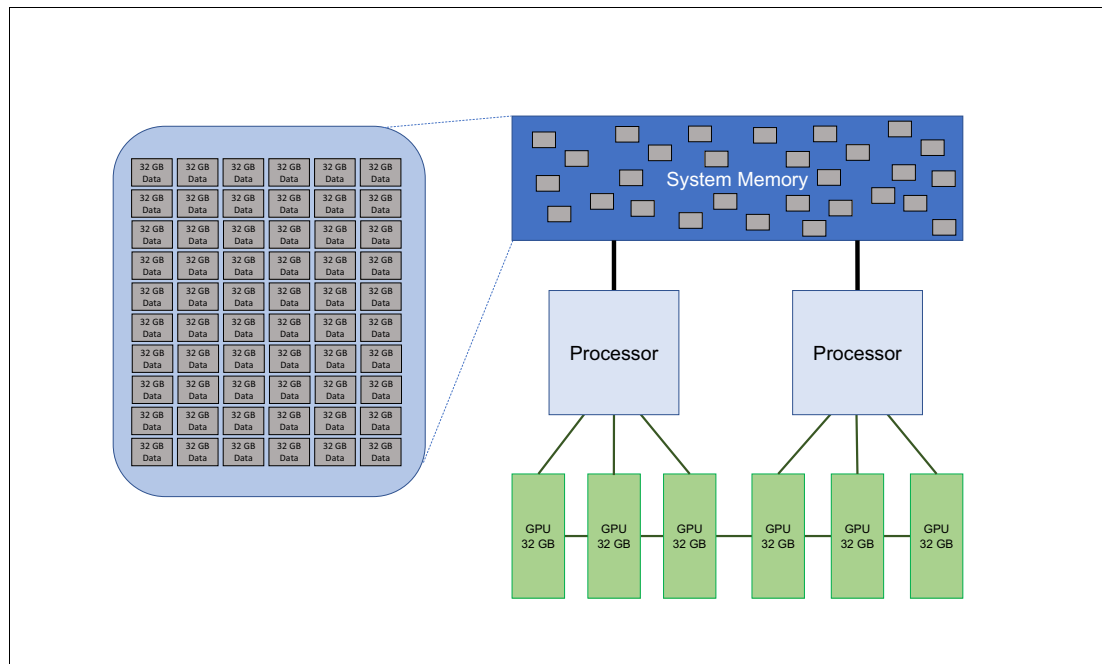


Figure A-2 Traditional approach of splitting the problem into 32 GB chunks

As models grow in complexity (DNNs that contain more layers and larger matrixes) and data sets increase in size (high definition video versus web scale images), data scientists are forced to make tradeoffs to stay within the 32 GB memory limits of each GPU.

With Large Model Support, which is enabled by the PowerAI unique NVLink connection between CPU (memory) and GPU, the entire model and data set may be loaded into system memory and cached down to the GPU for processing.

Users now may increase model sizes, data elements, and batch or set sizes significantly, which leads to processing far larger models and expanding to nearly 2 TB of system memory across all GPUs, as shown in Figure A-3.

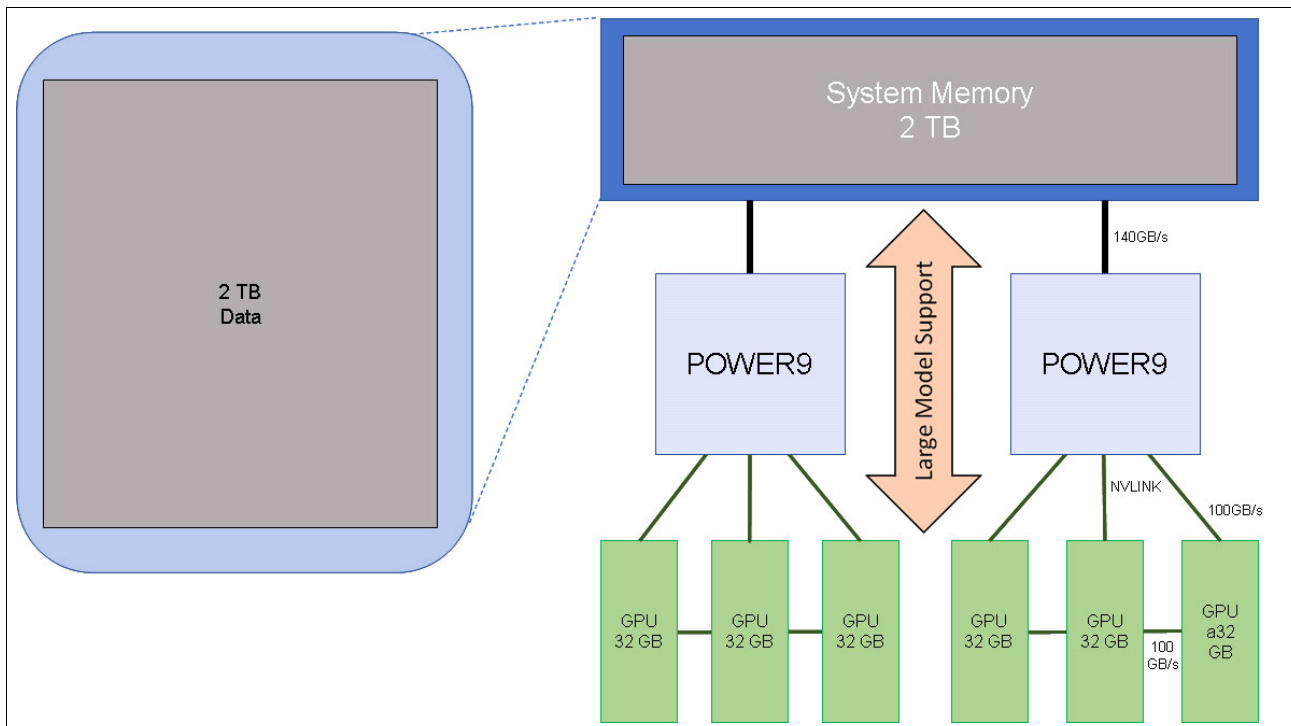


Figure A-3 Large Model Support approach with larger than 32 GB data being consumed by GPUs

This capability is unique to PowerAI, and provides the opportunity to address larger challenges and accomplish more work within a PowerAI single server, increasing organizational efficiency.

Software download

If you have a Power AC922 and want to test and use IBM PowerAI, you can download PowerAI (distributed as a binary file for RHEL 7.5) from [IBM PowerAI Enterprise](#).

For installation guides and other pertinent information, see [IBM PowerAI developer portal](#).

IBM PowerAI Enterprise

PowerAI Enterprise V1.1 provides robust, end-to-end workflow support for DL application logic, including the complete lifecycle management of installation and configuration, data ingest and preparation, building, optimizing, and training the model, and testing and moving the model into production. Taking advantage of a distributed architecture, PowerAI Enterprise enables your teams to quickly iterate through the training cycle on more data to continuously improve the model over time.

PowerAI Enterprise combines popular open source DL frameworks such as TensorFlow and IBM Caffe with unique tools that increase data scientist and cognitive developer productivity.

PowerAI Enterprise includes platform technical support, from the servers to the entire PowerAI Enterprise software stack, helping organizations confidently bring DL and machine learning into production.

PowerAI Enterprise features

PowerAI Enterprise provides many features that accelerate performance, improve resource utilization, and greatly reduce installation, configuration, and management complexities:

- ▶ Distributed DL architecture: Simplifying the process of training DL models across a cluster for faster time to results.
- ▶ Large model support: Increasing the amount of memory that is available for DL models up to 32 GB per network layer, enabling more complex models with larger, more high-resolution data inputs.
- ▶ Enhanced data ingest, preparation, and transformation tools by using Apache Spark for data management.
- ▶ IBM Power Systems servers are designed for AI applications, incorporating high-bandwidth and low-latency NVLink connections between GPU accelerators for peer-to-peer communications, and directly connecting GPU accelerators to system CPUs (and system memory).
- ▶ Powerful model development tools, including real-time training visualization and runtime monitoring of accuracy and hyper-parameter search and optimization for faster model development.
- ▶ Ready-to-use DL frameworks (TensorFlow and IBM Caffe) are provided with the PowerAI base package.
- ▶ The multitenant architecture is designed to run DL, high-performance analytics, and other long-running services and frameworks on shared resources.

PowerAI Enterprise provides a highly available and resilient multitenant framework, incorporating accelerated Spark, DL model lifecycle support, centralized management and monitoring, end-to-end security, and global technical support for the software and open source software that is provided by IBM.

IBM Spectrum Conductor® Deep Learning Impact is built on IBM Spectrum Conductor with Spark, which is a highly available and resilient multitenant distributed framework that provides Apache Spark and DL application lifecycle support, centralized management and monitoring, end-to-end security, and support from IBM. It is supported by IBM Power Systems servers with NVLink and NVIDIA GPUs.

Distributed training with IBM Fabric technology and Elastic Deep Learning

Many of the recent DL applications involve training models with a large amount of training data, which is a time-consuming process. With the development of new hardware technology, GPU devices can accelerate the training process. IBM Spectrum Conductor Deep Learning Impact leverages the power of GPUs and distribution training algorithms to provide a distributed training engine that supports distributed DL across multiple GPUs and multiple hosts.

Monitoring and optimization with deep learning insights

Training a DNN is complex and time-consuming. With IBM Spectrum Conductor Deep Learning Impact, you can monitor the current progress and status of your training job. You also have the capabilities to:

- ▶ Monitor DL training jobs by capturing logs from the underlying DL frameworks.
- ▶ Visualize training progress of your job by viewing monitoring data charts and summary information.
- ▶ Get advice about how to optimize the training job.

DL insights provide the ability to visualize run times, iteration, loss, accuracy and histograms of weights, activations, and gradients of the neural network. From these charts, you can see whether a training job is running smoothly and get advice about how to improve the training run. Improvements can be made based on the recommendations and the training can be run again.

Hyperparameter tuning

Hyperparameters are parameters whose values are set before stating the model training process. DL models such as convolutional neural network (CNN) models and recurrent neural network (RNN) models might have a few to several hundred hyperparameters. These parameters affect the model training process and the final model performance. Hyperparameter optimization is one of the direct impediments to tuning DL models. Hyperparameter optimization in IBM Spectrum Conductor Deep Learning Impact uses three major algorithms to atomically optimize hyperparameters:

- ▶ Random Search
- ▶ Tree-structured Parzen Estimator Approach (TPE)
- ▶ Bayesian Optimization based on Gaussian Process

IBM Spectrum Conductor Deep Learning Impact

Designed to get you set up and running as quickly as possible, IBM Spectrum Conductor Deep Learning Impact V1.1, when coupled with PowerAI, is delivered as a set of software packages that can deploy a fully functional DL environment within hours, and usually less than 1 hour. The software distributions contain a precompiled set of the software that is needed to build and manage a distributed environment, including the DL frameworks and any supporting software components that they require to run.

An Enterprise Class DL solution in a block diagram is shown in Figure A-4.

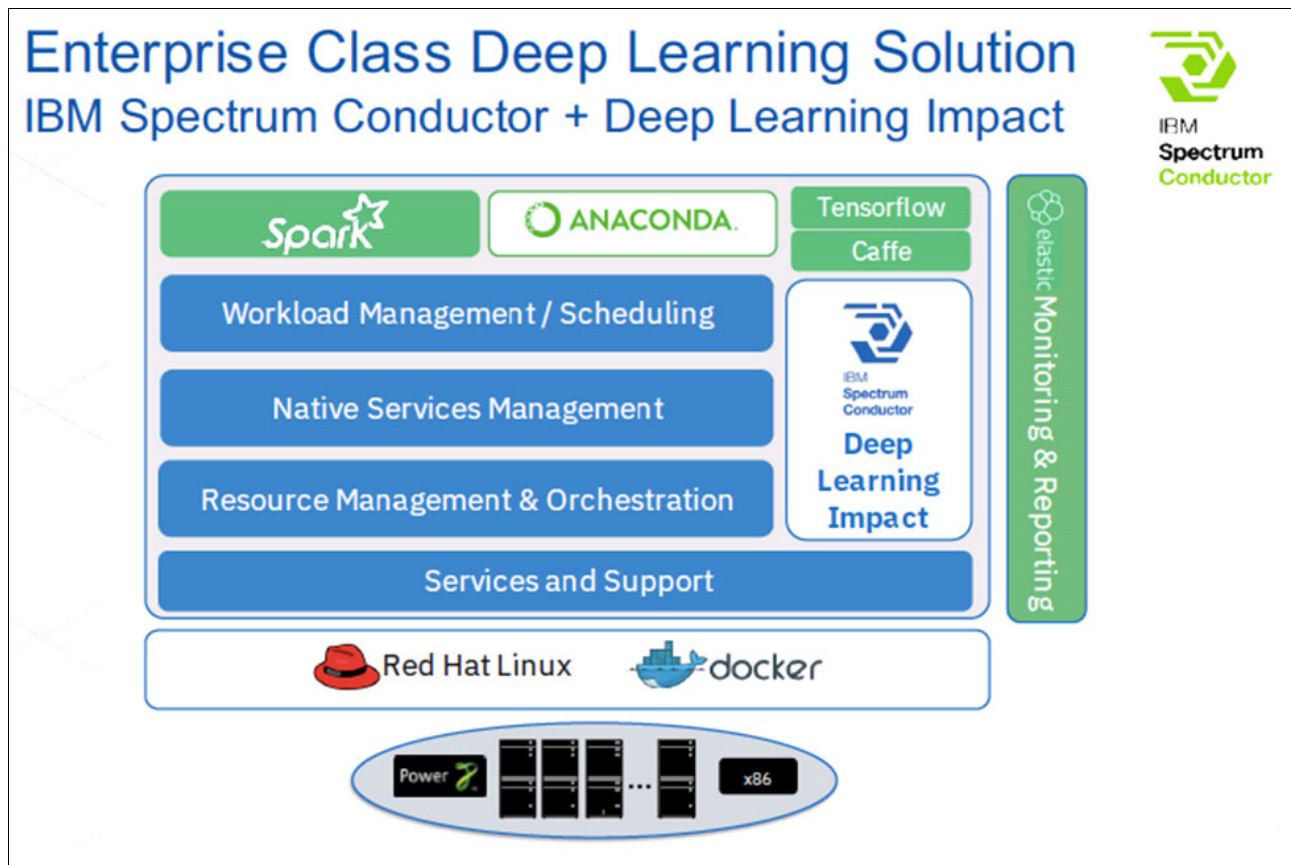


Figure A-4 Enterprise Class Deep learning Solution: Block diagram

Importing, preparing, and transforming data faster

Most of a data scientist's time is spent importing, transforming, and preparing data for training. IBM Spectrum Conductor Deep Learning Impact is designed to reduce that time with a rich set tools, automation, and workflow, enabling the data scientist to spend more time training and optimizing their models. A distributed Spark implementation can reduce the amount of compute time that is needed to import and run transformations by running multiple parallel import and transformation tasks simultaneously.

IBM Spectrum Conductor Deep Learning Impact increases the productivity of data scientist, as shown in Figure A-5.

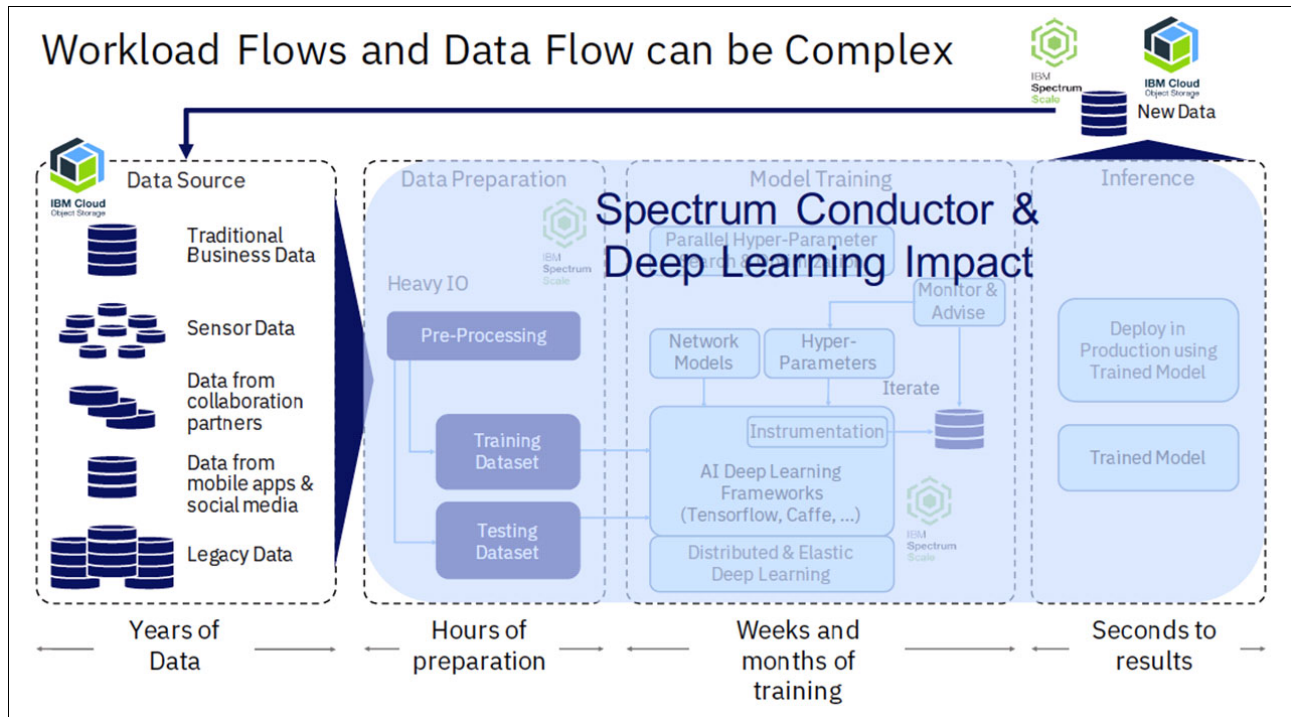


Figure A-5 PowerAI Enterprise: IBM Spectrum Conductor Deep Learning Impact

Automation to help with optimization and training

There is a scarcity of skills for model optimization. Even when the skills are present, the many possible optimizations can make choosing the best optimization a daunting task. IBM Spectrum Conductor Deep Learning Impact provides the following optimizations:

- ▶ Hyper-parameter search and optimization that uses multiple iterations and parallel processes that run on a distributed Spark architecture help you to find a better model and hyper-parameters for improved accuracy.
- ▶ Training visualization and runtime monitoring provide visualization of the training accuracy while in mid-process. These functions allow data scientists to interact with the model and judge accuracy during run time. For example, if the accuracy levels are clearly low after a short period of run time, the modeling can be stopped, adjusted, and restarted. This process saves a tremendous amount of time when users are stuck waiting for models to finish before having any insight into their accuracy and usability.

Dramatically reduce training times

Model optimization and training are compute-heavy and iterative. IBM Spectrum Conductor Deep Learning Impact can dramatically reduce the time that is needed to train models, enabling you to go into production faster or train by using larger data sets with the following features:

- ▶ Software and frameworks that are optimized to take full advantage of IBM Power Systems servers with NVLink CPUs and NVIDIA GPUs.
- ▶ Distributed training enables jobs to be broken down and distributed over a cluster of servers so that they can be processed in parallel.

- ▶ Elastic resource management makes it possible for compute resources to be automatically added to a training job as the demand increases, and then removed as the workload slows.

Deploying, inferring, scoring, capturing organizational value, and reiterating

After training is complete, IBM Spectrum Conductor Deep Learning Impact provides tools to help package and deploy models for testing and then put them into production. The models can be run in the distributed Spark environment in parallel with other DL projects, or deployed as an API or into any target device. To ensure that the model remains up-to-date and gets smarter as more data is collected, more iterations are taken through the DL workflow that is supported by IBM Spectrum Conductor Deep Learning Impact. The more times that you iterate, the more times you capture new data, and the more powerful the model becomes. Faster iterations mean higher fidelity and higher-quality models.

High-performance computing

A high-performance computing (HPC) system includes multiple servers that can run parallel programs. In this context, a parallel program is a piece of software that is specifically designed to run simultaneously on multiple servers. HPC is used mainly in the research and education segments. AI and HPC combine into one data lake to yield potentially dramatic enhanced value, and might be the catalyst to facilitate changes in the following situations:

- ▶ Automotive companies do research in computer-aided engineering (CAE) applications and run various independent software vendor (ISV) application for their machine improvements.
- ▶ Pharmaceutical companies do research in the simulation of molecular dynamics to engage various protein folding and check their reactions to particular cells.
- ▶ Meteorologists forecast various parameters, such as temperature, rainfall, wind speed, wind direction, storm direction, and sensitivity.
- ▶ Oceanographers simulate a cyclone's intensity and direction.
- ▶ Oil and gas companies run reservoir simulations to find oil in particular geographical areas.
- ▶ Running a computational fluid dynamics (CFD) application to simulate various fluid and flow processes and their effects, such as fluids or gas flowing through pipes, aircraft wing simulation, and rocket engines fuel burning simulation.
- ▶ The financial sector runs big data analytics, forecasts the future, and invests, for example, a Monte Carlo Simulation.
- ▶ Quantum mechanics researches simulate the law of quantum principle.
- ▶ Nuclear fusion researchers do simulations before doing it practically.

The basic principle of HPC is to improve the quality life of people, innovate, and develop new products. For HPC, a minor delay in an application, for example, a microsecond on one server, has a cascade effect, and that delay can increase the simulation time to minutes, hours, days, or months. Therefore, the whole infrastructure must be designed so that it does not impact on the application. Each component of HPC system, such as the server, network adapter, network switch, cables lengths, and software, must be optimized to provide the best results.

In each case above, the subject matter expert has the input data and the application, and wants to perform simulations efficiently to develop a new product or get the greatest benefits from the result. An HPC system can range between two servers to thousands of servers, and management software from IBM makes such HPC solutions easy to manage.

Figure A-6 shows a block diagram of an HPC solution.

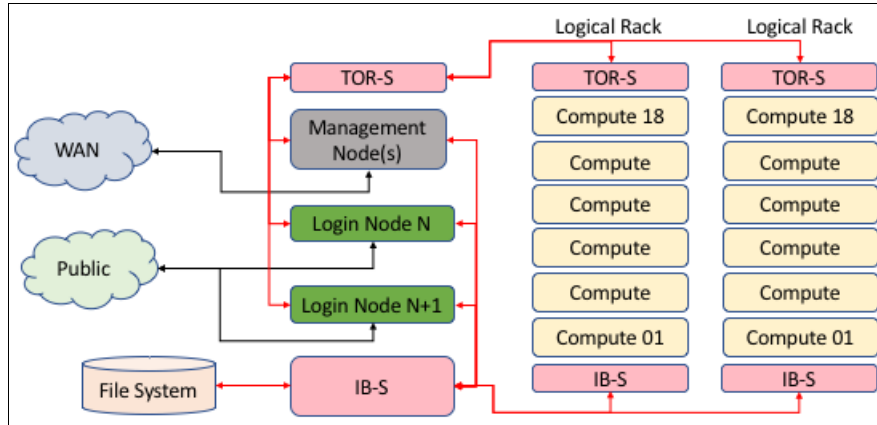


Figure A-6 Block diagram of an HPC system

A base HPC Power System server typically consists of the following server components:

- ▶ Login nodes (Power AC922 server with or without GPUs)
- ▶ Management nodes (Power AC922 or IBM Power System LC922 server)
- ▶ Compute nodes (Power AC922 server with GPUs), 2 or higher
- ▶ A storage system with a parallel file system (Elastic Storage Server or IBM Spectrum Scale with IBM Storwize®)
- ▶ Tape library for archives

The server components of an HPC system are coupled to the following networks:

- ▶ OS-level network
- ▶ Node and hardware management network
- ▶ High-performance interconnect network (Mellanox 100G InfiniBand or 10G network)

High-speed network switching between the servers is required for the Message Passing Interface (MPI), and there are various network topologies that are available:

- ▶ Fat tree topology (the most widely used)
- ▶ Mesh Topology or All-to-all
- ▶ 2D-tours topology
- ▶ 3D-tours topology
- ▶ 5D-tours topology (Use by the Bluegene/Q system)
- ▶ Dragonfly
- ▶ Hypercube

Note: Any network topology can be used, but it might take time and resources to check the performance of the application, so it is a preferred practice to use the fat tree topology.

A base HPC Power System server has the following software components:

- ▶ The current firmware for the server and network components.
- ▶ Mellanox OFED drivers.

- ▶ Unified Fabric Manager (UFM) from Mellanox is used for large HPC environments, and to manage the Mellanox network.
- ▶ OS (Red Hat Enterprise Linux).
- ▶ IBM Spectrum Cluster Foundation (cluster management).
- ▶ IBM Spectrum LSF® (job scheduler):
 - IBM LSF Application Center
 - IBM LSF License Scheduler
 - IBM LSF Process Manager
- ▶ IBM Spectrum MPI and Open MPI, or OpenMPI.
- ▶ IBM XL C / IBM XL Fortran (IBM XLF) (compilers).
- ▶ Advance Toolchain set (GNU compilers).
- ▶ IBM Scientific Libraries (Engineering and Scientific Subroutine Library (ESSL) and Parallel Engineering and Scientific Subroutine Library (PESSL)).
- ▶ IBM Parallel Performance Toolkit Programming Environment.
- ▶ NVIDIA CUDA driver for GPUs.
- ▶ NVIDIA CUDA compiler.
- ▶ NVIDIA PGI Compilers for OpenACC (XL C also contains an OpenACC framework).
- ▶ IBM Spectrum Scale.
- ▶ IBM Spectrum Archive.

The IBM software infrastructure for HPC Power Systems servers is GUI-based.

Here are some capabilities of the IBM Software solution:

- ▶ Provisioning and GPU-enabled workload management for demanding, distributed, and mission-critical science and engineering computing environments.
- ▶ Monitoring, alerting, and log analytics for optimal operational management.
- ▶ Workflow, license management, and intelligent policy-driven scheduling, all designed to work together to address HPC needs.
- ▶ Application-centric workload submission and management to improve user productivity.
- ▶ IBM POWER9 optimized development and runtime environments that use the MPI for its parallel version.
- ▶ Access data in IBM Spectrum Scale environments.
- ▶ HPC and AI in one package.

Note: Validate the applications' compatibility with the IBM Power System server so that you can take advantage of the following features:

- ▶ NVIDIA GPUs for floating point operations
- ▶ POWER9 processor and its memory bandwidth for integer operations
- ▶ Coherent Accelerator Processor Interface (CAPI) technology (Field Programmable Gate Array (FPGA) build application)

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-5472
- ▶ *IBM Power System L922 Introduction and Technical Overview*, REDP-5496
- ▶ *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 Power Systems E950 Introduction and Technical Overview*, REDP-5509
- ▶ *IBM Power Systems E980 Introduction and Technical Overview*, REDP-5510
- ▶ *IBM PowerVM Best Practices*, SG24-8062
- ▶ *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

You can search for, view, download, or order these documents and other Redbooks publications, Redpapers, web docs, drafts, and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- ▶ IBM Fix Central website
<http://www.ibm.com/support/fixcentral/>
- ▶ IBM Knowledge Center
<http://www.ibm.com/support/knowledgecenter/>
- ▶ IBM Knowledge Center: IBM Power Systems Hardware
<https://www.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm>

- ▶ IBM Knowledge Center: Migration combinations of processor compatibility modes for active Partition Mobility

<https://www.ibm.com/support/knowledgecenter/POWER7/p7hc3/iphc3pcmcombosact.htm>

- ▶ IBM Portal for OpenPOWER - POWER9 Monza Module
https://www.ibm.com/systems/power/openpower/tgcmDocumentRepository.xhtml?aliasId=POWER9_Monza
- ▶ IBM Power Systems website
<http://www.ibm.com/systems/power/>
- ▶ IBM Storage website
<http://www.ibm.com/systems/storage/>
- ▶ IBM System Planning Tool website
<http://www.ibm.com/systems/support/tools/systemplanningtool/>
- ▶ IBM Systems Energy Estimator
<http://www-912.ibm.com/see/EnergyEstimator/>
- ▶ NVIDIA Tesla V100
<https://www.nvidia.com/en-us/data-center/tesla-v100/>
- ▶ NVIDIA Tesla V100 Performance Guide
<http://images.nvidia.com/content/pdf/volta-marketing-v100-performance-guide-us-r6-web.pdf>
- ▶ OpenCAPI
<http://opencapi.org/technical/use-cases/>
- ▶ OpenPOWER Foundation
<https://openpowerfoundation.org/>
- ▶ Power Systems Capacity on Demand website
<http://www.ibm.com/systems/power/hardware/cod/>
- ▶ Support for IBM Systems website
<http://www.ibm.com/support/entry/portal/Overview?brandind=Hardware~Systems~Power>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



REDP-5494-00

ISBN 0738457027

Printed in U.S.A.

Get connected

