IBM Power Advanced Compute (AC) AC922 Server

The Best Server for Enterprise AI

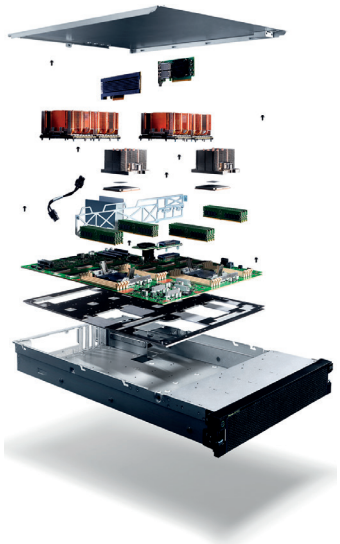
Highlights

IBM® Power Systems™ Accelerated Compute (AC922) server is an acceleration superhighway to enterprise-class AI.

- A Superhighway to and for acceleration—unleashes more accelerated computing potential, in the post CPU-only era
- Designed for the AI Era, architected for the modern analytics and AI workloads that fuel insights
- Delivering enterprise-class AI - cutting-edge AI innovation data scientists desire, with dependability IT requires
- Proven deployments from small clusters to the world's largest supercomputers, with near linear scaling
- Simplest GPU acceleration[3] —focus on what to accelerate not how, with coherence, available with AC922 in addition to larger model sizes enabled

Modern AI, HPC and analytics workloads are driving an ever-growing set of data intensive challenges that can only be met with accelerated infrastructure.

To help meet these demands, IBM Power Systems has designed AC922 for the AI Era, the best server for Enterprise AI. The AC922 leverages IBM's new POWER9™ processor with a myriad of modern connectivity capabilities yielding up to 5.6x[1] the data movement over the antiquated PCIe Gen 3 buses found in x86. IBM Power Systems deliver the only architecture enabling NVLink between CPUs and GPUs, unlocking new potential for accelerated computing.



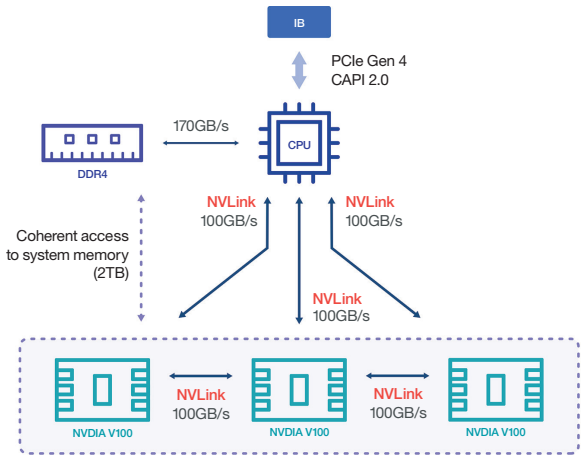
Power Systems AC922 internal components

IBM Power Systems Accelerated Compute (AC922) Servers

IBM Power Systems Accelerated Compute (AC922) server is an acceleration superhighway to enterprise-class AI.

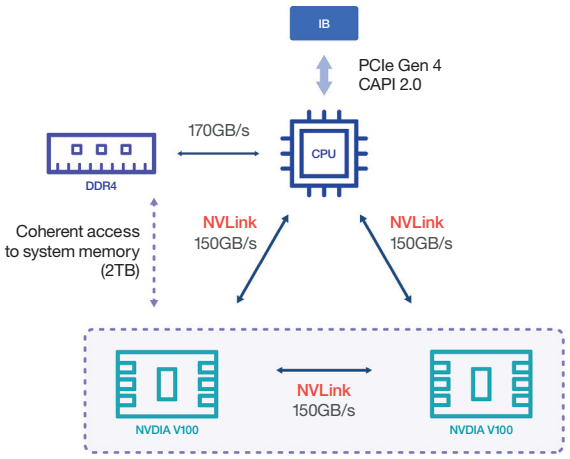
The next generation of IBM Power Systems, with POWER9 technology, is built with innovations that deliver unprecedented speed for the AI Era. The IBM Power System AC922 for high performance computing provides:

- **Faster I/O** -The AC922 includes a variety of next-generation I/O architectures, including: PCIe gen4, CAPI 2.0, OpenCAPI and NVLINK. These interconnects provide up to 5.6 times[1] as much bandwidth for today's data-intensive workloads versus the antiquated PCIe gen3 found in x86 servers.
- **Extraordinary CPUs** - While blazingly fast on their own, POWER9 CPUs truly excel in their ability to unleash and exploit the performance of everything around them. Built for the AI-Era, the POWER9 supports up to 5.6x[1] more I/O and 2x more threads than its x86 contemporaries. The POWER9 is available on configurations with 16, 18, 20 and 22 cores, for up to 44 cores in the AC922 server.
- **Advanced GPUs** -The AC922 pairs what is arguably the best GPUs for Enterprise AI with the best platform f or them. The AC922 pairs POWER9 CPUs and NVIDIA Tesla V100 with NVLink GPUs. Delivers up to 5.6x[1] times the performance for each pairing. This is the only server capable of delivering this I/O performance between CPUs and GPUs. This provides massive throughput capability for HPC, deep learning and AI workloads.
- **1st PCIe generation 4 server** - The AC922 is the industry's first server to feature the next generation of the industry standard PCIe interconnect. PCIe generation 4 delivers approximately 2x the data bandwidth[2] of the PCIe generation 3 interconnect found in x86 servers.



Power Systems AC922 with 6 GPUs and water Cooling

- **Simplest AI architecture**[3] - AI models grow large, easily outgrowing GPU memory capacity in most x86 servers. CPU to GPU coherence in the AC922 addresses these concerns by allowing accelerated applications to leverage System memory as GPU memory. And it simplifies programming by eliminating data movement and locality requirements. And, by leveraging the 5.6x[1] faster NVLink interconnect, sharing memory between CPUs and GPUs doesn't bottleneck down to PCIe 3 speeds, as it would on x86 servers.
- **Enterprise-ready** - Simplifies deep-learning deployment and performance. Unlocks a new, simpler end-to-end toolchain for AI users. Proven AI performance and scalability enable you to start with one node, then scale to a rack or thousands of nodes with near linear scaling efficiency.
- **Built for the world's biggest AI challenges, ideal for yours** - The AC922 is the backbone of the CORAL Summit supercomputer, meeting milestones to deliver 200+ petaflops of HPC and 3 exaflops of AI as a service performance. But with its efficiency and ease of AI deployment, it's also ideally-suited to address your organization's AI aspirations.

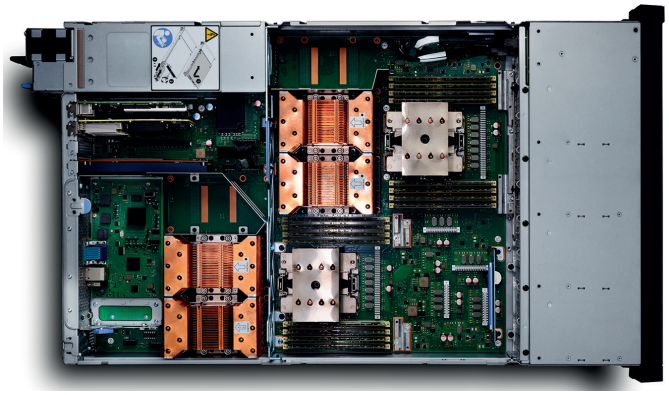


Power Systems AC922 with 4 GPUs and Air Cooling

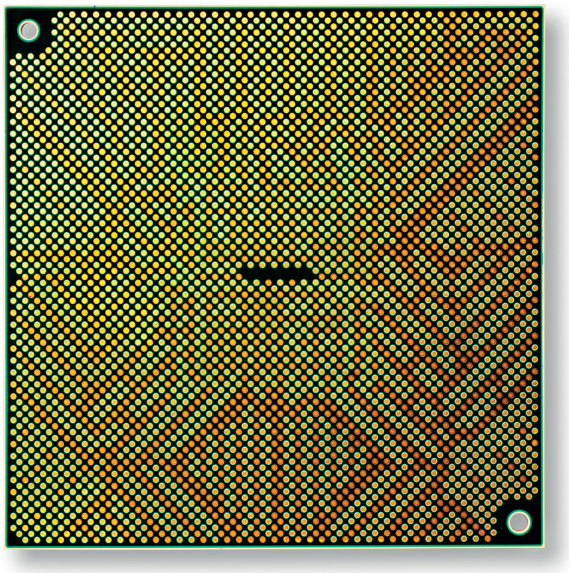
Designed for leaders in AI and Deep Learning, HPC and high-performance analytics, the IBM Power Systems AC922 provides:

- 2x POWER9 with NVLink 2.0 CPUs, with 16 DIMM sockets and up to 2 TB of memory
- A differentiated platform for GPU acceleration
  - POWER9 with NVLink 2.0 Technology: a link with up to 5.6X the performance (100 GB/sec air cooled or 150 GB/sec water cooled) to each NVIDIA V100 with NVLink GPU
  - Incredible CPU to GPU and GPU to GPU communication: up to 5.6X the data flow (100 or 150 GB/sec) between adjacent NVIDIA® Tesla® V100 GPU Accelerators on the same socket of PCI-E Gen3 x16 solutions
- Easy GPU Programming—full coherence and access to systems memory
- Advanced Mellanox ConnectX-5 InfiniBand Fabric, with industry-leading PCIe Gen4 interface
- Optional CAPI-attached NVMe storage for exceptionally fast storage I/O and burst buffer data staging
- Advanced interfaces to all accelerators: NVLink, OpenCAPI™, CAPI 2.0 and PCIe Gen4

Proven AI performance and scalability—start with one node, then scale to a rack much higher with near linear scaling efficiency.



Power Systems AC922 internal view



Power Systems AC922 POWER9 CPU

Power Systems AC922 front view

Power Systems AC922 (8335-GTC, 8335-GTW) at a glance	
System configurations	
Microprocessors	2x POWER9 with NVLink CPUs 16, 20 cores, or 18, 22 cores (with liquid cooling)
Level 2 (L2) cache	512 K
Level 3 (L3) cache	10 MB
RAM (memory)	Up to 2 TB, from 16 DDR4 RDIMM Sockets
Internal disk storage	2x SFF (2.5") drive bays, optional NVMe SSD support in PCIe slots
Processor-to-memory bandwidth	170 GB/s per socket, 340 GB/s per system
L2 to L3 cache bandwidth	7 TB/s on chip bandwidth
Internal SCSI disk bays	n/a
Media bays	n/a
Adapter slots	4 or 6 SXM 2.0 sockets, for NVIDIA Tesla V100 GPU Accelerators with NVLink. 2x PCIe x16 4.0 slots 1x PCIe16x (x8,x8) 4.0 slot (multi-socket host direct supported) 1x PCIe x4 4.0 slot
Standard features	
I/O ports	2x USB 3.0, 2x 1 GB Eth, VGA
Connectivity support (optional)	
POWER Hypervisor™	KVM
Advanced POWER Virtualization[1] (option)	
RAS features	Processor instruction retry Selective dynamic firmware updates Chip kill memory ECC L2 cache, L3 cache Service processor with fault monitoring Hot-swappable disk bays Redundant cooling fans
Operating systems	Red Hat Enterprise Linux, Ubuntu Linux
Power requirements	200 V to 240 V
System dimensions	Width: 441.5 mm (17.4 in.) Depth: 822 mm (32.4 in.) Height: 86 mm (3.4 in.) Weight: 30 kg (65 lbs.)
Warranty	3-year limited warranty, CRU (customer replaceable unit) for all other units (varies by country) next business day 9am to 5pm (excluding holidays), warranty service upgrades and maintenance are available.

Why IBM?

IBM is a trailblazer in AI—From early machine learning system in IBM Research to Watson® on Jeopardy, AI isn't just a buzzword for IBM. And we're applying that innovation to cognitive infrastructure, helping our customers on their journey to AI.

IBM aligns cutting-edge innovation with enterprise dependability—IBM has over 105 years of aligning continuous innovation with our customers' business needs.

IBM is your partner for the AI Era—IBM provides the most flexible and comprehensive range of technology and services needed for your entire journey to AI, whether you're looking to deploy a few nodes in order to investigate AI, or you're looking to deploy exaflops of AI as a Service.

For more information

To learn more about the Power Systems AC922 please contact your IBM representative or IBM Business Partner, or visit the following website:
ibm.com/marketplace/ai-hpc-server-power-ac922

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

IBM Systems
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
December, 2017

[1] **5.6x more I/O bandwidth** – tested results are based on IBM Internal Measurements running the CUDA H2D Bandwidth Test Hardware: Power AC922; 32 cores (2 x 16c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU; Ubuntu 16.04. S822LC for HPC; 20 cores (2 x 10c chips), POWER8 with NVLink; 2.86 GHz, 512 GB memory, Tesla P100 GPU Competitive HW: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04

[2] PCIe Generation 4 provides 2x data throughput vs. PCIe gen 3 (31.5 GB/s vs 15.8 GB/s x16)

[3] Simplest Ai Architecture - Coherence simplifies coding by abstracting data movement and locality for developers

Please Recycle

POD03143-USEN-00