IBM Power AC922 Server

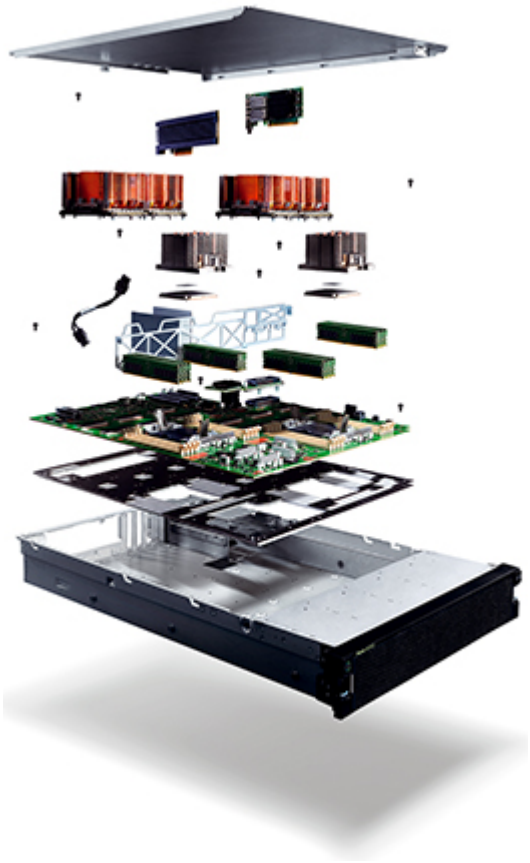
The Best Server for Enterprise AI

Highlights

- More accuracy - GPUs access system RAM for larger models
- Faster insights - significant deep learning speedups
- Rapid deployment - integrated software & hardware stack
- Enterprise-ready- cutting edge AI IBM Power reliability & IBM support
- Proven at scale - from one node to the world's fastest supercomputers

As AI initiatives shift from backroom experiments to boardroom imperatives, many organizations find their existing infrastructure ill-suited to help them make the journey. IT leaders are now under pressure to implement untested tools, circumvent proven processes, and shrink the timeframe from beta to production. Organizations that do not evaluate their existing IT infrastructure, ensuring it is appropriate for Enterprise AI, will risk getting caught in an endless loop of failed experiments.

The IBM Power System AC922 supports Enterprise AI initiatives throughout their lifecycle. With an optimized hardware and software stack, the AC922 delivers earlier prototypes and yields faster ROI. The world's only server enabling NVIDIA NVLink between CPUs and GPUs, the Power AC922 delivers 5.6x[1]the data movement of x86-based servers. This I/O enables GPU-based algorithms to leverage system memory, enabling models up to 60x[4] larger than x86 based systems. The Power AC922 delivers faster AI insights, providing up to 3.8x[5] speedups for deep learning workloads. Proven as the backbone of the world's largest supercomputers, the Power AC922 is capable of meeting an enterprise's loftiest AI aspirations.



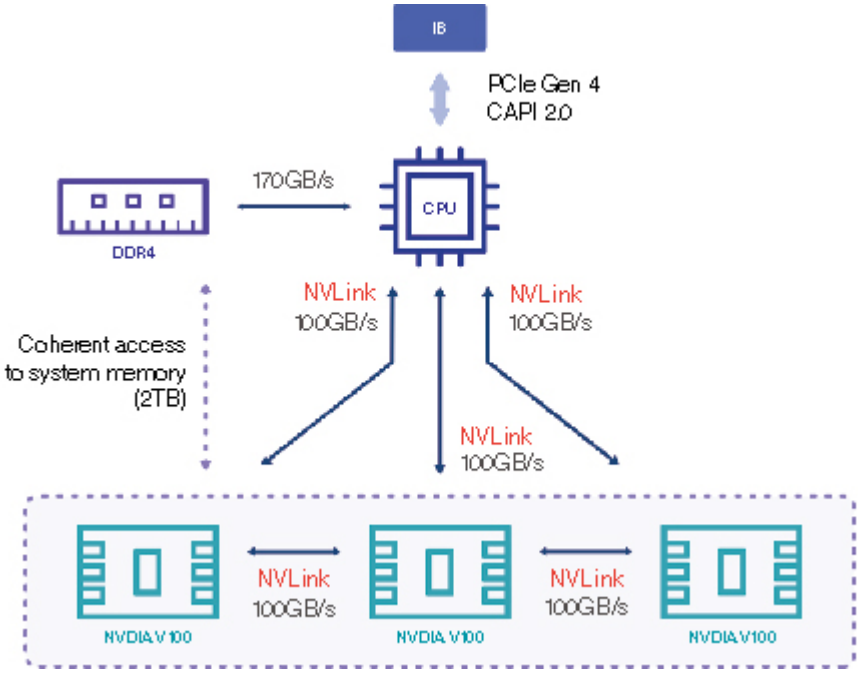
Power System AC922 internal components

IBM Power Systems Accelerated Compute (AC922) Server

Delivering unprecedented speed for analytics and AI, the IBM Power System AC922:

- **Yields more accuracy (up to 60x larger models[4])** - The Power AC922 is uniquely capable of supporting larger models and data sets, by accessing system from GPU-based processes and algorithms, without PCIe bottlenecks in x86-based systems.
- **Delivers faster insights** - The Power AC922 delivers up to 3.8x[5] the AI performance, vs similarly-configured x86-based systems.
- **Provides faster I/O** -The Power AC922 includes a variety of next-generation I/O architectures, including: PCIe Gen4, CAPI 2.0, OpenCAPI and NVIDIA 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-based servers.
- **CPUs unleash GPU acceleration** - Built for enterprise AI, the AC922 supports up to 5.6x[1] more I/O and 2x more threads than its x86 contemporaries. The Power AC922 is available on configurations with 16, 18, 20 and 22 cores, for up to 44 cores.

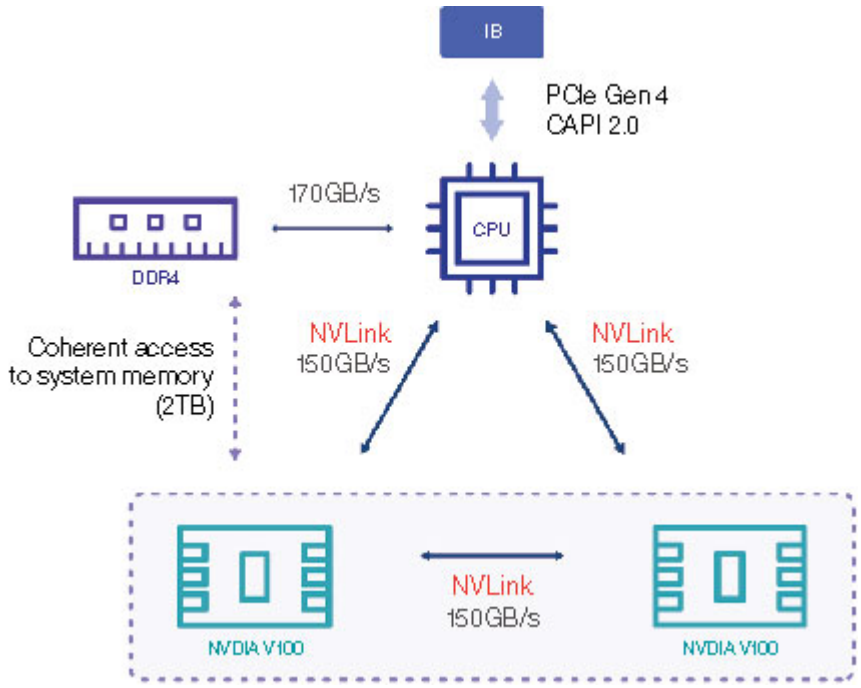
- **Includes the most advanced data center GPUs** -The Power AC922 pairs POWER9 CPUs and NVIDIA Tesla V100 with NVLink GPUs. This combination delivers up to 5.6x[1] times the performance for each pairing. The Power AC922 is the only server capable of delivering this I/O performance between CPUs and GPUs, supporting the massive throughput required for HPC, deep learning and AI workloads.
- **1st PCIe Gen4 server** - The AC922 is the industry's first server to feature the next generation of the industry standard PCIe interconnect. PCIe Gen4 delivers approximately 2x the data bandwidth[2] of the PCIe Gen3 interconnect found in x86-based servers.



Power System AC922 with 6 GPUs and water Cooling

- **Simplest AI architecture**[3] - Many AI models are large, easily outgrowing GPU memory capacity in most x86-based servers. CPU to GPU coherence in the Power AC922 addresses these concerns by allowing accelerated applications to leverage System memory as GPU memory. This acceleration simplifies programming by abstracting data movement and locality. By leveraging the 5.6x[1] faster NVIDIA NVLink interconnect, sharing memory between CPUs and GPUs doesn't bottleneck down to PCIe Gen3 speeds, as it would on x86-based servers.
- **Enterprise-ready** - Unlock a new, simpler end-to-end toolchain for AI users. Proven AI performance and scalability enables organizations to start with one node, then scale to a rack or thousands of nodes with near linear scaling efficiency.
- **Scales to meet the world's biggest AI challenges** - The Power AC922 is the backbone of the world's most powerful computers, United States Department of Energy's Summit and Sierra supercomputers, delivering hundreds of petaflops of HPC and exaflops of AI as a service

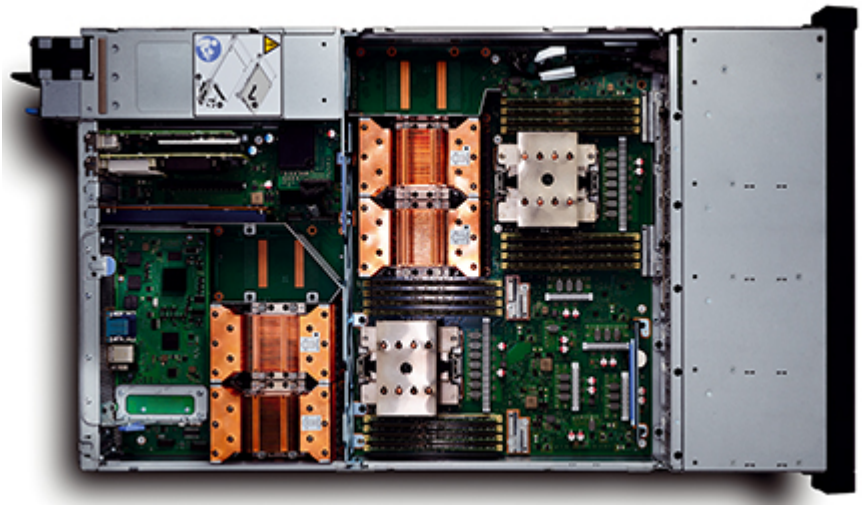
performance. With the Power AC922's efficiency and ease of AI deployment, it is also ideally-suited to address any organization's AI aspirations.



Power System AC922 with 4 GPUs and Air Cooling

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

- 2x POWER9 with NVLink 2.0 CPUs, with 16 DIMM sockets and up to 2 TB of memory
- A uniquely capable platform for GPU acceleration
  - POWER9 with NVLink 2.0 Technology: a link with up to 5.6X the performance (150 GB/sec air cooled or 100 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
- Simplified GPU Programming—full coherence and access to systems memory simplifies programming
- Advanced Mellanox ConnectX-5 InfiniBand Fabric, with industry-leading PCIe Gen4 interface
- 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 System AC922 internal view

Power System AC922 front view

[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 Gen4 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

[4]*Supports larger models – by using system memory, without the PCIe bottleneck, GPU-based processes and algorithms can access significantly larger models. 4x (DGX-2 with 16 32GB V100 GPUs vs AC922 with 2TB system memory) 60x (models run against 1 V100 GPU with 32GB memory vs AC922 with 2TB system memory)*

[5] *Enables faster insights – 3.8x speedup based on comparing an AC922 with an Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240x2240) on Caffe*

Power System AC922 (8335-GTC, 8335-GTW) at a glance	
System configurations	
Microprocessors	2x POWER9 with NVLink CPUs 16, 20 cores, or 18, 22 cores (with water 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
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
POWER Hypervisor	KVM
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 a proven partner for the journey to enterprise AI—IBM provides the most flexible and comprehensive range of technology and services needed for an enterprise's entire journey to AI.

For more information

To learn more about the Power System AC922 please contact your IBM representative or IBM Business Partner, or visit the following website: **http://ibm.com/us-en/marketplace/power-systems-ac922**

IBM Maintenance and Technical Support solutions can help you get the most out of your IT investment by reducing support costs, increasing availability and simplifying management with integrated support for your multiproduct, multivendor hardware and software environment. For more information on hardware maintenance, software support, solution support and managed support, visit: ibm.com/services/maintenance

This document contains information pertaining to the following IBM products which are trademarks and/or registered trademarks of IBM Corporation:
IBM® Power Systems™, NVIDIA® Tesla®, Watson®, POWER9™, OpenCAPI™, POWER Hypervisor™

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