Datasheet

NVIDIA DGX B300

Setting a new bar for AI factory performance, from training to inference.

Delivering Unprecedented Performance to Every Enterprise

The adoption of AI across industries has seen exponential growth over a short period of time, signaling a fundamental shift in the way businesses are approaching their AI transformation. Organizations that have been slow to adopt technological advancements out of skepticism are racing to outfit their data centers with the right infrastructure and augment their teams with the right talent to leverage AI, but many of these organizations are finding that deploying AI is not as simple as they may have planned.

While the promise of generative AI is transformative, these enterprises are facing several common challenges in adopting and scaling these technologies. Among them are finding the right solution for integration complexities, filling critical gaps in expertise, and managing energy consumption and costs. These organizations are finding that they are not equipped to scale and operate in the same way hyperscalers do.

NVIDIA DGX™ B300, the building block of NVIDIA DGX SuperPOD, is a purpose-built AI infrastructure solution tailored to meet the computational demands of AI reasoning, leveraging full-stack software to ease the burden on enterprises to deploy AI in a streamlined manner. Powered by NVIDIA Blackwell Ultra GPUs, DGX B300 delivers 144 petaFLOPS for inference and 72 petaFLOPS for training, all in a new form factor designed to fit seamlessly into the modern data center and is compatible with NVIDIA MGX™ and traditional enterprise racks. With DGX B300, any enterprise can perform training and inference on diverse AI workloads with an unprecedented level of efficiency.

Real-Time AI Powerhouse

DGX B300 represents a significant leap forward in real-time inference capabilities, enabling companies of all sizes to harness AI performance previously reserved for hyperscalers. As the world's first fully integrated system powered by NVIDIA Blackwell Ultra GPUs and NVIDIA ConnectX-8 networking, and with optimized NVIDIA Mission Control software, DGX B300 delivers 11x the inference performance and 4x the training performance compared to the previous generation. DGX B300 empowers every organization to unlock new possibilities in the era of AI reasoning.

Key Features

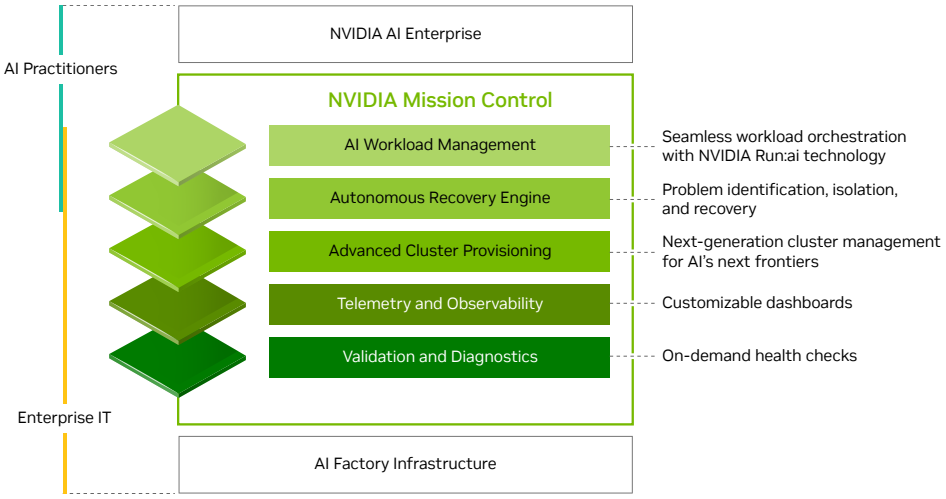
- Built with NVIDIA Blackwell Ultra GPUs
- 2.3TB of GPU memory space
- 72 petaFLOPS of training performance
- 144 petaFLOPS of inference performance
- NVIDIA networking
- Dual Intel Xeon Processors
- Foundation of NVIDIA DGX BasePOD™ and NVIDIA DGX SuperPOD™
- Leverages NVIDIA AI Enterprise and NVIDIA Mission Control software

A Blueprint for Modern Data Centers

DGX B300 introduces a redesigned chassis that fits into NVIDIA MGX racks, ensuring compatibility and scalability in modern data centers. DGX B300's air-cooled design allows for easy integration into existing data center infrastructure. For the first time, customers can take advantage of a flexible power architecture that allows them to choose between busbar and PSU options, enabling them to choose the right fit for their existing infrastructure and sustainability goals. This new DGX design serves as a blueprint for the optimal design of accelerated computing infrastructure, offering a flexible foundation upon which they can build and deploy AI infrastructure at scale. By combining leading-edge design with practical serviceability, DGX B300 sets a new standard for adaptable, efficient, and future-ready AI infrastructure.

Run Models, Automate the Essentials With NVIDIA Mission Control

NVIDIA Mission Control powers every aspect of AI factory operations—from developer workloads to infrastructure to facilities—with the skills of a world-class operations team, now delivered as software. It brings instant agility for inference and training while providing full-stack intelligence for infrastructure resilience. Mission Control lets every enterprise run AI with hyperscale-grade efficiency accelerating AI experimentation. Additionally, NVIDIA AI Enterprise offers a software suite to streamline AI development and deployment and is optimized to run on NVIDIA DGX systems. Use NVIDIA NIM™ microservices for optimal model deployment, offering speed, ease of use, manageability, and security.



NVIDIA DGX software stack.

DGX B300 Technical Specifications

	DGX B300
GPU	NVIDIA Blackwell Ultra GPUs
Total GPU Memory	2.3TB
Performance	144 PFLOPS FP4 inference* 72 PFLOPS FP8 training*
NVIDIA NVLink Switch System	2x
NVIDIA NVLink™ Bandwidth	14.4 TB/s aggregate bandwidth
Power Consumption	~14kW
CPU	Dual Intel® Xeon® Processors
Networking	8x OSFP ports serving 8x single-port NVIDIA ConnectX-8 VPI > Up to 800Gb/s NVIDIA InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField-3 DPU > Up to 800Gb/s NVIDIA InfiniBand/Ethernet
Management Network	1GbE onboard NIC with RJ45 1GbE RJ45 Host baseboard management controller (BMC)
Storage	OS: 2x 1.9TB NVMe M.2 Internal storage: 8x 3.84TB NVMe E1.S
Software	NVIDIA DGX OS / NVIDIA Mission Control / NVIDIA Base Command Manager / NVIDIA AI Enterprise Supports Red Hat Enterprise Linux / Rocky / Ubuntu
Rack Units	10RU
Support	Three-year business-standard hardware and software support

*Shown with sparsity.

Ready to Get Started?

To learn more about DGX B300, visit nvidia.com/dgx-b300

NVIDIA