Datasheet

NVIDIA DGX B300

Setting a new bar for real-time AI performance, from training to inference.

Delivering Unprecedented Efficiency to Every Enterprise

A dual transformation is forging the future of business. First, we've entered the era of AI reasoning, where models perform complex, multi-step thinking. Simultaneously, the data center is evolving into an AI factory, a new class of facility built to manufacture intelligence at scale. This new reality requires a fundamental rethinking of enterprise infrastructure.

For enterprises, this dual shift presents profound challenges. AI reasoning requires immense computation, memory, and bandwidth. At the same time, building and operationalizing an AI factory is complex, and many companies struggle with critical gaps in expertise, system integration, and rising energy costs. They're finding they lack the specialized teams and tools needed to run this sophisticated new ecosystem with the efficiency and resilience of a hyperscaler.

NVIDIA DGX™ B300 is the foundational building block for this era of AI, empowering pioneers to build their own AI factories capable of tackling the demands of AI reasoning. Powered by the NVIDIA Blackwell Ultra architecture, DGX B300 is an AI powerhouse purpose-built to deliver hyperscale performance in an enterprise-sized footprint—including 1.5x dense FP4 performance and 2x attention performance compared to DGX B200. Its redesigned, NVIDIA MGX™-compliant, air-cooled chassis allows for seamless integration into modern data centers. Paired with NVIDIA Mission Control software, DGX B300 simplifies AI operations, delivering the full-stack solution enterprises need to master the complexity of generative AI and unlock the return on their investment.

Real-Time AI Powerhouse

DGX B300 is engineered to be the foundational building block of the AI factory. It enables AI innovators of all sizes to manufacture intelligence at scale, harnessing generative AI capabilities previously reserved for global-scale AI organizations. As a fully integrated system powered by NVIDIA Blackwell Ultra GPUs, NVIDIA® ConnectX®-8 networking, and NVIDIA Mission Control software, DGX B300 delivers unprecedented performance and hyperscale-grade efficiency. By combining exceptional training performance with leadership-class, real-time inference, DGX B300 empowers every organization to build scalable infrastructure for the era of AI reasoning.

Key Features

- Built with 8x NVIDIA Blackwell Ultra SXM
- 2.1 TB of GPU memory space
- 72 petaFLOPS of training performance
- 144 petaFLOPS of inference performance
- NVIDIA networking
- Intel Xeon 6776P processors
- Foundation of NVIDIA DGX BasePOD™ and NVIDIA DGX SuperPOD™
- Leverages NVIDIA AI Enterprise and NVIDIA Mission Control™ software

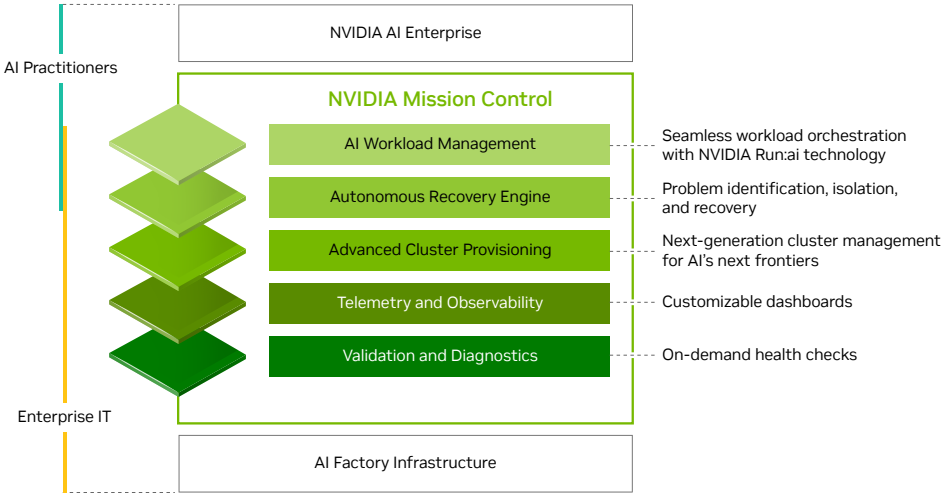
A Blueprint for Modern Data Centers

DGX B300 serves as the blueprint for the modern AI factory, starting with its redesigned, Open Compute Project (OCP)-compliant chassis that brings hyperscaler design principles to any data center. For the first time, DGX B300 can be deployed using an NVIDIA MGX-compatible chassis or a more traditional AC-powered design, ensuring DGX B300 can be used in any infrastructure environment. By pairing this leading-edge design with practical serviceability, DGX B300 enables enterprises to construct AI factories on their own terms, with unprecedented efficiency and choice.

Run Models, Automate the Essentials With NVIDIA Mission Control

Deploying an AI factory is more than an infrastructure purchase—it's an operational commitment. Many enterprises starting their AI transformation face significant complexity that leads to costly downtime and low utilization, directly hindering ROI. This is the challenge NVIDIA Mission Control is designed to solve.

Mission Control acts as a software-defined operations team, delivering the skills of a world-class AI factory operator in software to ensure enterprises get the most from their investment. It automates the full range of complex tasks—from initial cluster bring-up to daily workload management—allowing IT teams to run AI infrastructure with hyperscale-grade efficiency. To protect hardware investments, Mission Control provides critical infrastructure resiliency and maximizes productivity and uptime for AI factories, empowering developers to spend less time waiting and more time innovating.



NVIDIA DGX software stack.

DGX B300 Technical Specifications

	DGX B300
GPU	8x NVIDIA Blackwell Ultra SXM
Total GPU Memory	2.1 TB total, 62 TB/s HBM3e bandwidth
Performance	FP4 Tensor Core* - 144 PFLOPS | 108 PFLOPS FP8/FP6 Tensor Core** - 72 PFLOPS
System Memory	2 TB, configurable to 4 TB
NVIDIA NVLink™ Switch System	2x
NVIDIA NVLink Bandwidth	14.4 TB/s aggregate bandwidth
System Power	14.5 kW (Busbar) 15.1 kW (PSU)
CPU	Intel Xeon 6776P processors
Networking	8x OSFP ports serving 8x NVIDIA ConnectX-8 VPI >Up to 800 Gb/s of NVIDIA InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField®-3 DPU >Up to 400 Gb/s of NVIDIA InfiniBand/Ethernet
Management Network	1GbE onboard network interface card (NIC) with RJ45 1GbE RJ45 host baseboard management controller (BMC)
Storage	OS: 2x 1.9 TB NVMe M.2 Internal storage: 8x 3.84 TB NVMe E1.S
Software	NVIDIA AI Enterprise (optimized AI software) NVIDIA Mission Control (AI data center operations and orchestration with NVIDIA Run:ai technology) NVIDIA DGX OS (operating system) Supports Red Hat Enterprise Linux / Rocky / Ubuntu
Rack Units	10
Operating Temperature	10C°–35C°
Support	Three-year business-standard hardware and software support

* Specification shown as sparse | dense.
** Shown with sparsity. Dense is 1/2 sparse spec shown.

Ready to Get Started?

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

NVIDIA