Datasheet

NVIDIA DGX B200

A complete AI platform for training, fine-tuning, and inference.

Powering the Next Generation of AI

Artificial intelligence is transforming almost every business by automating tasks, enhancing customer service, generating insights, and enabling innovation. It's no longer a futuristic concept but a reality that's fundamentally reshaping how businesses operate. However, as AI workloads continue to develop, they're beginning to require significantly more compute capacity for both training and inference than most enterprises have available. To leverage AI, enterprises need high-performance computing, storage, and networking capabilities that are reliable and efficient.

NVIDIA DGX™ B200 is a complete AI platform that defines the next chapter of generative AI by taking full advantage of NVIDIA Blackwell GPUs and high-speed interconnects. Configured with eight NVIDIA Blackwell GPUs, DGX B200 delivers unparalleled generative AI performance with a massive 1.4 terabytes (TB) of GPU memory, 64 terabytes per second (TB/s) of HBM3e memory bandwidth, and 14.4 TB/s of all-to-all GPU bandwidth, making it uniquely suited to handle any enterprise AI workload.

With NVIDIA DGX B200, enterprises can equip their data scientists and developers with a universal AI supercomputer to accelerate their time to insight and fully realize the benefits of AI for their businesses.

One Platform for Develop-to-Deploy Pipelines

As AI workflows have become more sophisticated, so too has the need for enterprises to handle large datasets at all stages of the AI pipeline, from training to fine-tuning to inference. This requires massive amounts of compute power. With NVIDIA DGX B200, enterprises can arm their developers with a single, unified platform built to accelerate their workflows. Supercharged for the next generation of generative AI, DGX B200 enables businesses to infuse AI into their daily operations and customer experiences.

Key Features

NVIDIA DGX B200

- Built with eight NVIDIA Blackwell GPUs
- 1.4TB of GPU memory space
- 72 petaFLOPS of training performance
- 144 petaFLOPS of inference performance
- NVIDIA networking
- Dual 5th generation Intel® Xeon® Scalable Processors
- Foundation of NVIDIA DGX BasePOD™ and NVIDIA DGX SuperPOD™
- Leverages NVIDIA AI Enterprise and NVIDIA Mission Control software

Powerhouse of AI Performance

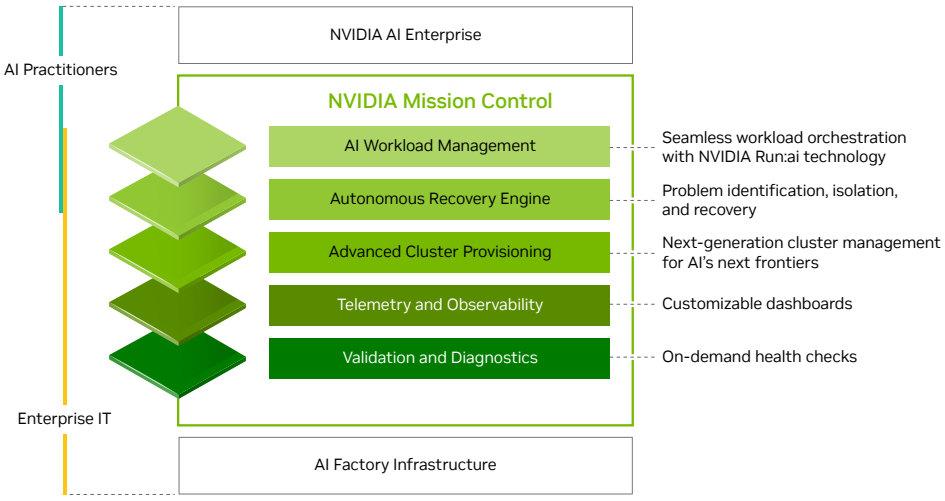
NVIDIA is dedicated to designing the next generation of the world's most powerful supercomputers, built to tackle the most complex AI problems that enterprises face. Powered by the NVIDIA Blackwell architecture's advancements in computing, DGX B200 delivers 3X the training performance and 15X the inference performance of DGX H100. DGX B200 offers high-speed scalability for NVIDIA DGX BasePOD and NVIDIA DGX SuperPOD, delivering top-of-the-line performance in a turnkey AI infrastructure solution.

Proven Infrastructure Standard

NVIDIA DGX B200 is the world's first system with the NVIDIA Blackwell GPU, delivering breakthrough performance for the world's most complex AI problems, such as large language models and natural language processing. DGX B200 offers a fully optimized hardware and software platform that leverages the complete NVIDIA AI software stack, a rich ecosystem of third-party support, and access to expert advice from NVIDIA professional services, allowing organizations to solve the biggest and most complex business problems with AI.

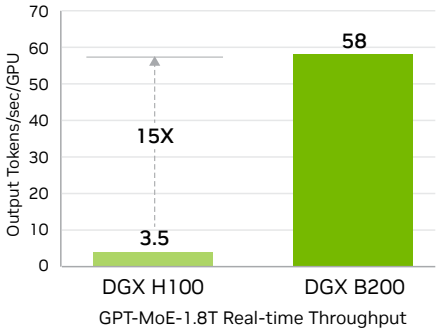
Run Models, Automate the Essentials With NVIDIA Mission Control

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, offering a suite of software to streamline AI development and deployment, is optimized to run on NVIDIA DGX systems. Use NVIDIA NIM™ microservices for optimal model deployment, offering speed, ease of use, manageability, and security.



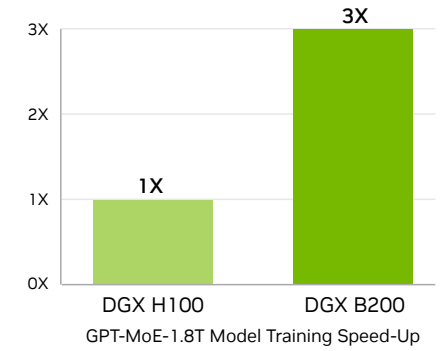
State-of-the-art AI factory software stack

Real-Time Large Language Model Inference



Projected performance subject to change. Token-to-token latency (TTL) = 50ms real time, first token latency (FTL) = 5,000ms, input sequence length = 32,768, output sequence length = 1,028, 8x eight-way DGX H100 GPUs air-cooled vs. 1x eight-way DGX B200 air-cooled, per GPU performance comparison.

Supercharged AI Training Performance



Projected performance subject to change. 32,768 GPU scale, 4,096x eight-way DGX H100 air-cooled cluster: 400G IB network, 4,096x 8-way DGX B200 air-cooled cluster: 400G IB network.

DGX B200 Technical Specifications	
GPU	8x NVIDIA Blackwell GPUs
GPU Memory	1,440GB total, 64TB/s HBM3e bandwidth
Performance	72 petaFLOPS FP8 training and 144 petaFLOPS FP4 inference
NVIDIA® NVSwitch™	2x
NVIDIA NVLink Bandwidth	14.4 TB/s aggregate bandwidth
System Power Usage	~14.3kW max
CPU	2 Intel® Xeon® Platinum 8570 Processors 112 Cores total, 2.1 GHz (Base), 4 GHz (Max Boost)
System Memory	2TB, configurable to 4TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI >Up to 400Gb/s NVIDIA InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField-3 DPU >Up to 400Gb/s InfiniBand/Ethernet
Management Network	10Gb/s onboard NIC with RJ45 100Gb/s dual-port ethernet NIC Host baseboard management controller (BMC) with RJ45
Storage	**OS:** 2x 1.9TB NVMe M.2 Internal storage: 8x 3.84TB NVMe U.2
Software	NVIDIA AI Enterprise – Optimized AI Software NVIDIA Mission Control – AI Data Center Operations and Orchestration with NVIDIA Run:ai Technology NVIDIA DGX OS / Ubuntu – Operating System
Rack Units (RU)	10 RU
System Dimensions	**Height:** 17.5in (444mm) **Width:** 19.0in (482.2mm) **Length:** 35.3in (897.1mm)
Operating Temperature	5–30°C (41–86°F)
Enterprise Support	Three-year Enterprise Business-Standard Support for hardware and software 24/7 Enterprise Support portal access Live agent support during local business hours

Ready to Get Started?

To learn more about NVIDIA DGX B200,
visit: nvidia.com/dgx-b200

NVIDIA

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 hyperscaler performance in an enterprise-sized footprint—144 petaFLOPS of FP4 inference performance for reasoning workloads. 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 NVIDIA Blackwell Ultra GPUs
- 2.3 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

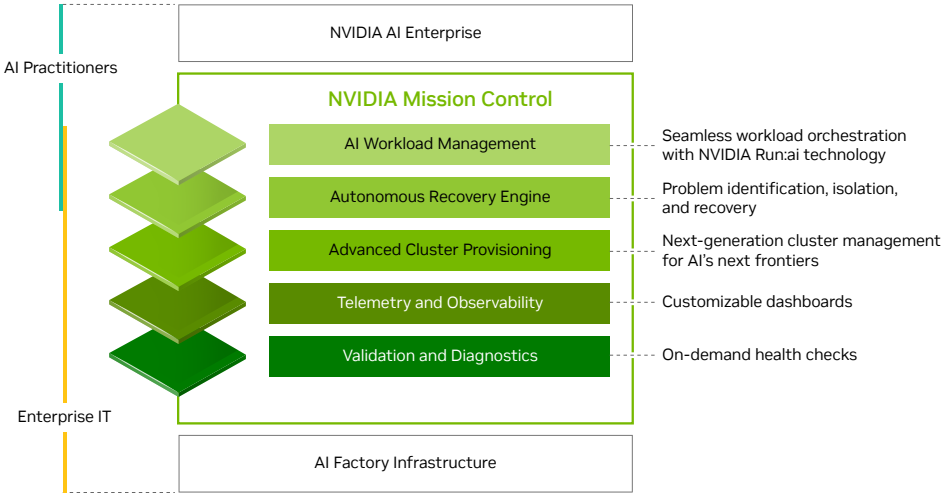
A Blueprint for Modern Data Centers

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	NVIDIA Blackwell Ultra GPUs
Total GPU Memory	2.3 TB
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	~14 kW
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

*Shown with sparsity.

Ready to Get Started?

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

NVIDIA