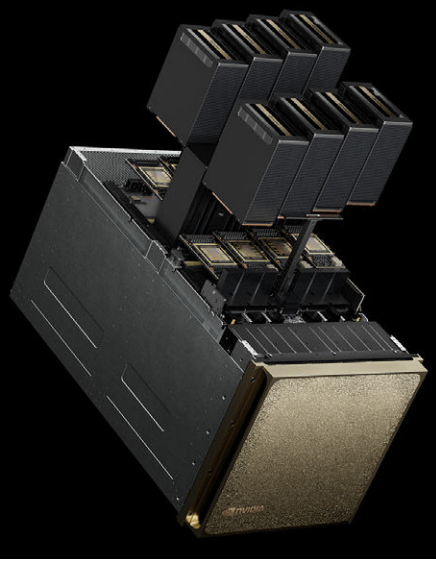
Datasheet

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
- 1.4 TB 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

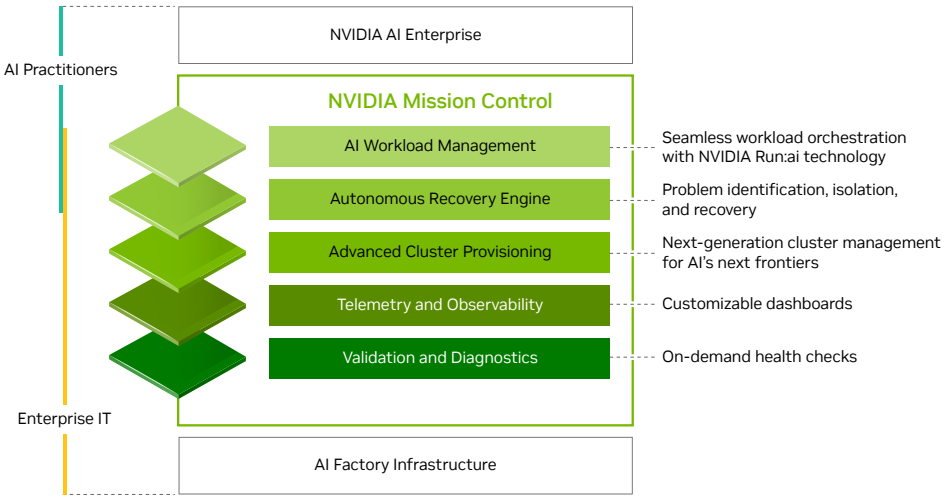
NVIDIA is dedicated to designing the next generation of the world's most powerful supercomputers, built to tackle the most complex AI problems that enterprises face. Powered by the NVIDIA Blackwell architecture's advancements in computing, DGX B200 delivers 3x the training performance and 15x the inference 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.

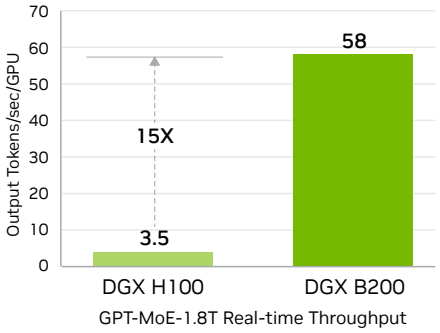
Run Models, Automate the Essentials With NVIDIA Mission Control

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

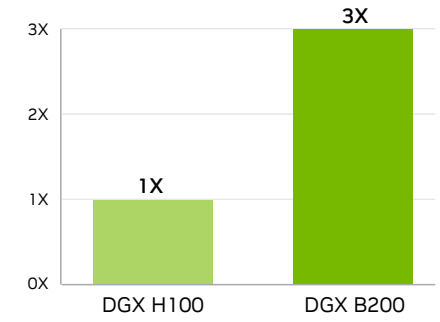
Real-Time Large Language Model Inference



GPT-MoE-1.8T Real-time Throughput

Projected performance subject to change. Token-to-token latency (TTL) = 50 ms 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



GPT-MoE-1.8T Model Training Speed-Up

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	1,440 GB total, 64 TB/s HBM3e bandwidth
Performance	FP4 Tensor Core: 144 | 72* petaFLOPS FP8 Tensor Core: 72 | 36* petaFLOPS
NVIDIA® NVSwitch™	2x
NVIDIA NVLink Bandwidth	14.4 TB/s aggregate bandwidth
System Power Usage	~14.3 kW max
CPU	2 Intel® Xeon® Platinum 8570 Processors 112 Cores total, 2.1 GHz (Base), 4 GHz (Max Boost)
System Memory	2 TB, configurable to 4 TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI >Up to 400 Gb/s NVIDIA InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField-3 DPU >Up to 400 Gb/s NVIDIA InfiniBand/Ethernet
Management Network	10 Gb/s onboard NIC with RJ45 100 Gb/s dual-port ethernet NIC Host baseboard management controller (BMC) with RJ45
Storage	**OS:** 2x 1.9 TB NVMe M.2 Internal storage: 8x 3.84 TB 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:** 17.5 in (444 mm) **Width:** 19.0 in (482.2 mm) **Length:** 35.3 in (897.1 mm)
Operating Temperature	10-35°C (50-90°F)
Enterprise Support	Three-year Enterprise Business-Standard Support for hardware and software

*Specifications shown as sparse | dense.

Ready to Get Started?

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

NVIDIA