Datasheet

NVIDIA DGX B200

A unified 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 than most enterprises have available. To leverage AI, enterprises need high-performance computing, storage, and networking capabilities that are secure, reliable, and efficient.

Enter **NVIDIA DGX™ B200**, the latest addition to the **NVIDIA DGX platform**. This unified AI platform defines the next chapter of generative AI by taking full advantage of NVIDIA B200 Tensor Core GPUs and high-speed interconnects. Configured with eight B200 GPUs, DGX B200 delivers unparalleled generative AI performance with a massive 1.4 terabytes (TB) of GPU memory and 64 terabytes per second (TB/s) of memory 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, businesses can infuse AI into their daily operations and customer experiences with DGX B200.

Key Features

NVIDIA DGX B200

- Built with eight NVIDIA B200 Tensor Core 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
- Includes NVIDIA AI Enterprise and NVIDIA Base Command™ software

Powerhouse of AI Performance

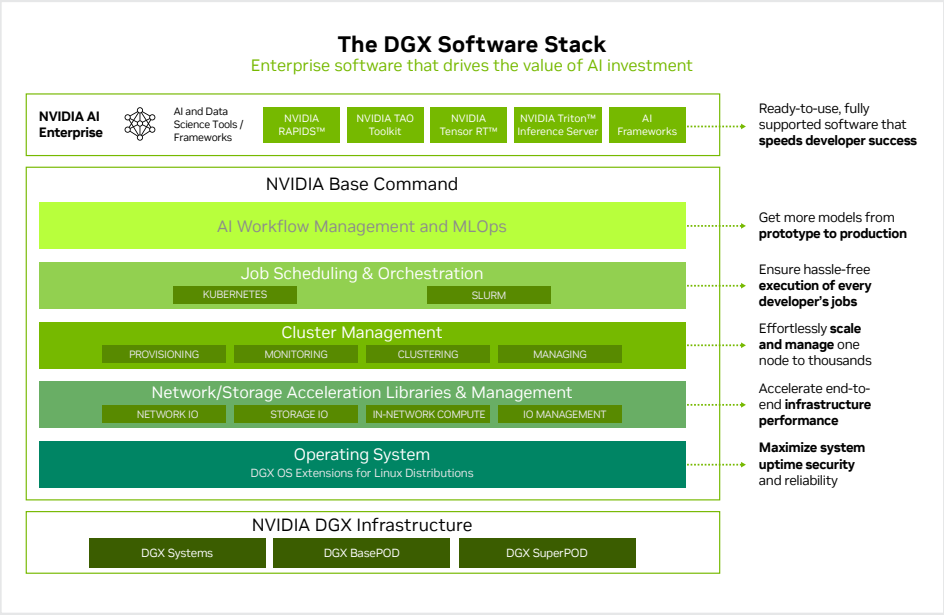
NVIDIA is dedicated to designing the next generation of the world's most powerful supercomputers, built to tackle the most complex AI problems that enterprises face. DGX B200 is the latest addition to the NVIDIA accelerated computing platform to showcase that commitment. Powered by the innovative NVIDIA Blackwell architecture's advancements in computing, DGX B200 delivers 3X the training performance and 15X the inference performance of DGX H100. As the foundation of NVIDIA DGX POD™ reference architectures, 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

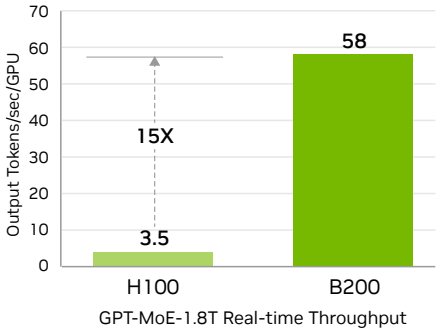
NVIDIA DGX B200 is the world's first system with the NVIDIA B200 Tensor Core GPU, delivering breakthrough performance for the world's most complex AI problems, such as large language models and natural language processing. DGX B200 offers a fully optimized hardware and software platform that includes 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.

Powered by NVIDIA Base Command

NVIDIA Base Command powers the DGX platform, enabling organizations to leverage the best of NVIDIA software innovation. Enterprises can unleash the full potential of their DGX infrastructure with a proven platform that includes enterprise-grade orchestration and cluster management, libraries that accelerate compute, storage, and network infrastructure, and an operating system optimized for AI workloads. Additionally, DGX infrastructure includes **NVIDIA AI Enterprise**, a suite of software optimized to streamline AI development and deployment.

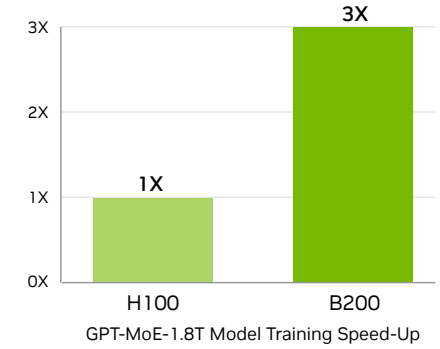


Real-Time Large Language Model Inference



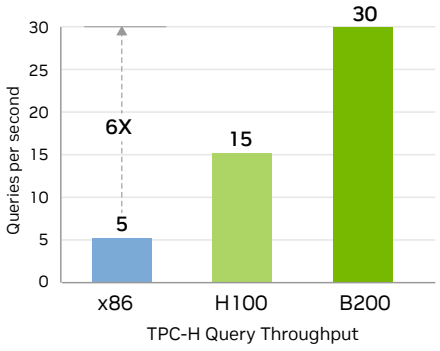
Projected performance subject to change. Token-to-token latency (TTL) = 50ms real time, first token latency (FTL) = 5,000ms, input sequence length = 32,768, output sequence length = 1,028, 8x eight-way DGX H100 GPUs air-cooled vs. 1x eight-way DGX B200 air-cooled, per GPU performance comparison.

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.

Accelerated Data Processing



Projected performance subject to change. Database join query with Snappy / Deflate compression derived from TPC-H Q4 query. 1x x86, 1x H100 GPU, and 1x B200 single GPU.

DGX B200 Technical Specifications	
GPU	8x NVIDIA B200 Tensor Core GPUs
GPU Memory	1,440GB total
Performance	72 petaFLOPS training and 144 petaFLOPS inference
NVIDIA® NVSwitch™	2x
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	Up to 4TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s 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 Base Command – Orchestration, Scheduling, and Cluster Management 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