Datasheet

NVIDIA Elite Partner

XENON High Performance Computing

NVIDIA DGX GB200

Advanced AI infrastructure for generative AI.

The Era of Trillion-Parameter AI

Enterprises of all sizes are using generative AI to develop chatbots and copilots, personalize content, accelerate drug discovery, create visual applications, and more. Today's start-of-the-art foundation models have trillions of parameters and train on as much as a petabyte of data. This new generation of highly capable AI models needs training and inference infrastructure with thousands of GPUs to iterate more efficiently on new ideas, speed up time to result, and achieve near-real-time inference.

Enterprise-Class Generative AI Infrastructure

NVIDIA DGX™ GB200 lets enterprises achieve unprecedented performance and predictable uptime, dramatically boosting utilization and productivity and increasing the ROI of their AI initiatives. It creates a new standard for AI performance, reliability, and scalability.

With scalability up to tens of thousands of GPUs with NVIDIA DGX SuperPOD, the efficient liquid-cooled DGX GB200 rack-scale design leverages NVIDIA GB200 Grace Blackwell Superchips to tackle the state-of-the-art AI models needed for today's advanced generative AI applications.

DGX GB200 is purpose-built to deliver extreme performance and consistent uptime for superscale generative AI training and inference workloads. Built on NVIDIA's own internal cluster designs, the full-stack resilience capabilities—available for the first time in enterprise AI infrastructure—allow enterprises to focus on innovation rather than operational complexity.

Maximize Developer Productivity

DGX GB200 delivers full-stack resilience for AI infrastructure. The intelligent control plane constantly tracks thousands of data points across hardware, software, and data center infrastructure to ensure continuous operation and data integrity. It features automatic failover using standby hardware and a robust checkpoint and restart mechanism—avoiding downtime, even when system administrators are unavailable.

Key Features

- Built on NVIDIA GB200 Grace™ Blackwell Superchips
- Scalable up to tens of thousands of GB200 Superchips with NVIDIA DGX SuperPOD™.
- 72 NVIDIA Blackwell GPUs connected as one with NVIDIA® NVLink™
- Efficient, liquid-cooled, rack-scale design
- NVIDIA networking
- Leverages NVIDIA AI Enterprise and NVIDIA Mission Control software

Supercomputing for Generative AI

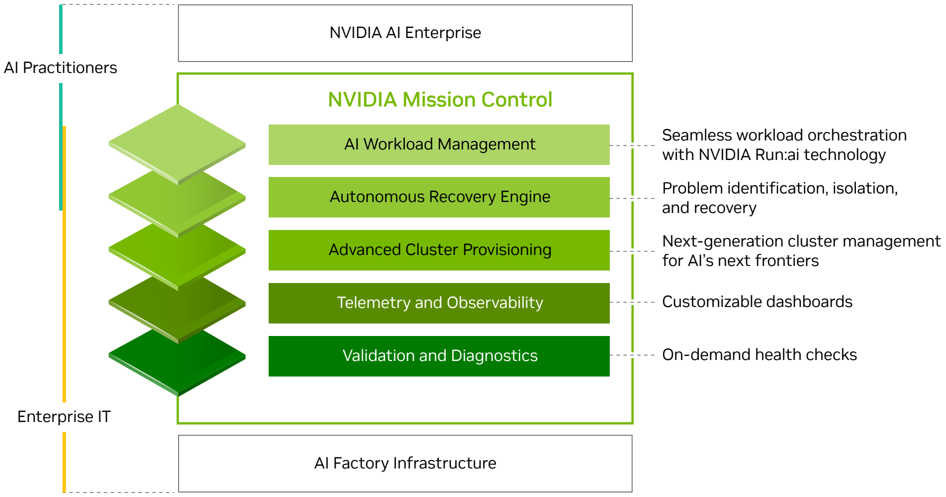
To achieve training and latency-sensitive inference on state-of-the-art parameter models, DGX GB200 can scale up to tens of thousands of NVIDIA Grace Blackwell Superchips. Ideal for large language models (LLMs), each DGX GB200 system features 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs connected as one with fifth-generation NVIDIA NVLink, delivering 1.4 exaFLOPS of AI performance, 30 terabytes (TB) of fast memory, and 130 terabytes per second (TB/s) of bidirectional GPU bandwidth. DGX GB200 enables enterprises to effortlessly perform training and inference on the largest generative AI models today and into the future.

Built on NVIDIA Grace Blackwell

With a groundbreaking 4-nanometer fabrication process, fifth-generation NVLink, and a second-generation Transformer Engine, the NVIDIA Grace Blackwell Superchips powering DGX GB200 are integrated into a liquid-cooled, rack-scale design that creates the world's most efficient AI supercomputer for generative AI. Each superchip features two high-performance NVIDIA Blackwell GPUs and an NVIDIA Grace CPU. Every Blackwell GPU in a GB200 Superchip delivers 1.8 TB/s of bidirectional throughput using NVLink for GPU-to-GPU connectivity.

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.


AI Practitioners
NVIDIA AI Enterprise
NVIDIA Mission Control
AI Workload Management
Seamless workload orchestration with NVIDIA Run:ai technology
Autonomous Recovery Engine
Problem identification, isolation, and recovery
Advanced Cluster Provisioning
Next-generation cluster management for AI's next frontiers
Telemetry and Observability
Customizable dashboards
Validation and Diagnostics
On-demand health checks
Enterprise IT
AI Factory Infrastructure

Technical Specifications

	DGX GB200
GPU	72x NVIDIA Blackwell GPUs, 36x NVIDIA Grace CPUs
CPU Cores	2,592 Arm® Neoverse V2 cores
GPU Memory | Bandwidth	13.4 TB HBM3e | 576 TB/s
Total Fast Memory	30.2 TB
Performance	FP4 Tensor Core: 1,440 PFLOPS | 720 PFLOPS* FP8/FP6 Tensor Core - 720 PFLOPS | 360 PFLOPS*
Interconnect	72x OSFP single-port NVIDIA ConnectX®-7 VPI with 400 Gb/s NVIDIA InfiniBand 36x dual-port NVIDIA BlueField®-3 VPI with 200 Gb/s InfiniBand and Ethernet
NVIDIA NVLink Switch System	9x L1 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
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 / Red Hat Enterprise Linux / Rocky: Operating system
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 GB200, visit:
https://xenon.com.au/product/nvidia-dgx-gb200/
nvidia.com/dgx-gb200

Contact XENON today!
www.xenon.com.au | info@xenon.com.au | 1300 888 030

NVIDIA Elite Partner

XENON® High Performance Computing