Datasheet

NVIDIA DGX GB300

Break through performance bottlenecks to accelerate state-of-the-art models.

Purpose-Built AI Factory Infrastructure for the Era of Reasoning

Generative AI and large language models (LLMs) have experienced unprecedented growth in recent years, revolutionizing the AI landscape and forcing enterprises to rapidly develop comprehensive AI strategies. This exponential expansion has brought forth new challenges, particularly when it comes to the infrastructure requirements to support the computational intensity of frontier models.

The immense size and complexity of modern AI models require high-performance distributed computing clusters, advanced networking, and massive memory pools. As the AI landscape continues to evolve, purpose-built infrastructure solutions will play a crucial role in enabling enterprises to stay at the forefront of AI innovation, not only driving breakthroughs across industries but also unlocking new possibilities for business transformation through advanced AI applications.

NVIDIA DGX™ GB300 is a comprehensive AI infrastructure solution designed for the era of AI reasoning and uniquely suited for training and inference of state-of-the-art models. DGX GB300 is built with Grace Blackwell Ultra Superchips and can be scaled up to tens of thousands of Superchips with NVIDIA DGX SuperPOD, creating a massive shared memory space that accelerates the performance of the world's largest AI models. Its rack-scale, 100% liquid-cooled design is tailored to fit into contemporary data centers using NVIDIA MGX racks, helping ease the burden on enterprises to manage high-performance AI hardware. DGX GB300 equips organizations with an infrastructure solution that can break through performance bottlenecks and handle the most demanding AI workloads.

Built on NVIDIA Grace Blackwell Ultra

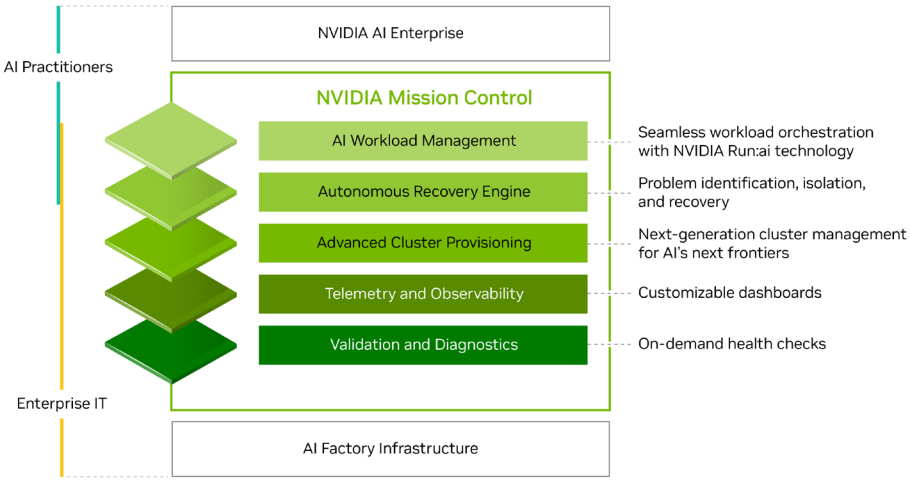
DGX GB300 represents a transformative leap in AI computing capabilities, powered by the NVIDIA Grace Blackwell Ultra architecture. This leading-edge system to delivers unprecedented inference performance while enhancing training capabilities across any AI workload. Powered by 36 Grace CPUs and 72 Blackwell Ultra GPUs that are connected into one massive GPU through an NVIDIA NVLink Switch System, DGX GB300 delivers 1.4 exaFLOPS for inference and 360 PFLOPS for training. DGX GB300 is purpose-built to optimize performance for the entire AI pipeline, from model training and post-training optimization to test-time inference, enabling enterprises to scale their infrastructure to meet the demands of the era of AI reasoning.

Key Features

- Built on NVIDIA GB300 Grace Blackwell Ultra Superchips
- Scalable up to tens of thousands of GB300 Superchips with NVIDIA DGX SuperPOD
- 72 NVIDIA Blackwell Ultra GPUs connected as one with NVIDIA® NVLink®
- Efficient, 100% liquid-cooled, rack- scale design
- NVIDIA networking
- Leverages NVIDIA AI Enterprise and NVIDIA Mission Control software

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.



NVIDIA DGX software stack

Designed for the Modern Data Center

DGX GB300 is engineered to seamlessly integrate with modern data center environments, offering unparalleled flexibility and scalability. With DGX GB300, enterprises are empowered to operate with hyperscaler-level capabilities without the need to overhaul their existing infrastructure. With a 100% liquid-cooled rack design, DGX GB300 significantly enhances energy efficiency and allows for higher power density crucial for advanced AI workloads. By combining hyperscaler-level performance with the adaptability to work within existing infrastructures, DGX GB300 represents a transformative solution for organizations seeking to harness the full potential of AI without compromising on efficiency or compatibility.

Technical Specifications

	DGX GB300
GPU	72x NVIDIA Blackwell Ultra GPUs in Grace Blackwell Ultra Superchips
CPU Cores	2,592
GPU Memory HBM3e	20.1TB
Total Fast Memory	37.9TB
Performance	1,400 petaFLOPS of FP4 AI performance*
	700 petaFLOPS of FP8 AI performance*
	360 petaFLOPS of FP16 AI performance*
Networking	72x OSFP single-port NVIDIA ConnectX®-8 VPI with 800Gb/s NVIDIA InfiniBand
	18x dual-port NVIDIA BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet
NVIDIA NVLink Switch System	9x L1 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
Software	NVIDIA Mission Control
	NVIDIA AI Enterprise
	NVIDIA DGX OS
	Supports Ubuntu
Support	Three-year business-standard hardware and software support

*Shown with sparsity.

Ready to Get Started?

To learn more about DGX GB300, visit nvidia.com/dgx-gb300

NVIDIA