Datasheet

NVIDIA DGX SuperPOD With DGX GB200 Systems

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 SuperPOD™ with DGX™ GB200 systems 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, its efficient, liquid-cooled, rack-scale design leverages NVIDIA GB200 Grace Blackwell Superchips to tackle the trillion-parameter AI models needed for today's advanced generative AI applications.

This next generation of DGX SuperPOD 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 With Constant Uptime

DGX SuperPOD with DGX GB200 systems delivers constant uptime with 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
- 72 NVIDIA Blackwell GPUs connected as one with NVIDIA® NVLink®
- Efficient, liquid-cooled, rack-scale design
- NVIDIA networking
- Integrated predictive maintenance to maximize uptime
- Includes NVIDIA AI Enterprise and NVIDIA Base Command™ software
- Each DGX SuperPOD is fully assembled and tested at the factory to speed on-site deployment

Supercomputing for Generative AI

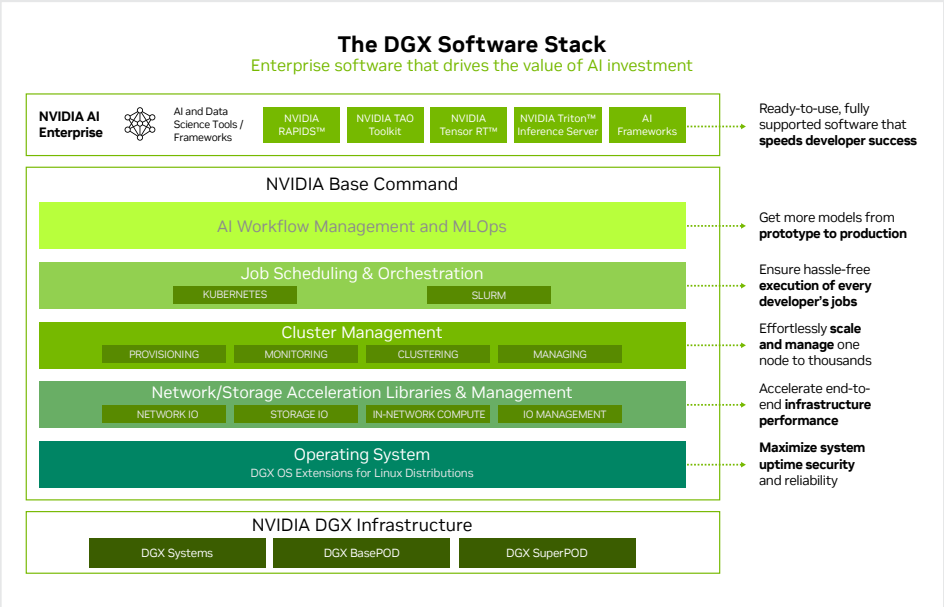
To achieve training and latency-sensitive inference on state-of-the-art trillion-parameter models, DGX SuperPOD with DGX GB200 systems can scale up to tens of thousands of NVIDIA Grace Blackwell Superchips. Ideal for large language models (LLMs),each DGX GB200 system within DGX SuperPOD features 36 NVIDIA 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 SuperPOD can optionally be configured using DGX GB200 systems with 576 NVLink-connected NVIDIA Blackwell GPUs to create a massive shared memory pool to provide significant speedups for memory-bound workloads such as recommenders, graph neural networks (GNNs), and **mixture-of-experts (MoE) LLMs** at trillion-parameter scale. DGX SuperPOD with 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 SuperPOD 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 1.8TB/s of bidirectional throughput using NVLink for GPU-to-GPU connectivity.

Integrated AI Software

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.



Technical Specifications	
	72-GPU NVLink Domain (NVL72)
FP4 AI	1,440 PFLOPS
FP8 AI	725 PFLOPS
FP16 AI	362 PFLOPS
GPU	72x NVIDIA Blackwell GPUs in Grace Blackwell Superchips
GPU Memory HBM3e	13.3TB
Total Fast Memory	30.2TB
Interconnect	72x OSFP single-port NVIDIA ConnectX®-7 VPI with 400Gb/s InfiniBand 36x 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 AI Enterprise: optimized AI software NVIDIA Base Command: orchestration, scheduling, and cluster management DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky: operating system
Enterprise Support	Three-year Enterprise Business-Standard Support for hardware and software

Ready to Get Started?

To learn more about NVIDIA DGX SuperPOD with DGX GB200 systems, visit: **nvidia.com/dgx-gb200**

NVIDIA