Datasheet

NVIDIA DGX SuperPOD

Turnkey data center solution for the AI enterprise.

NVIDIA DGX SuperPOD™ brings together leadership-class infrastructure with agile, scalable performance for the most challenging AI and high-performance computing (HPC) workloads. **DGX SuperPOD** delivers a full-service experience with industry-proven results in weeks instead of months. It's not just a collection of hardware, but a full-stack data center platform that includes industry-leading computing, storage, networking, software, and infrastructure management optimized to work together and provide maximum performance at scale. To make it even easier, DGX SuperPOD comes with NVIDIA Infrastructure Specialists who ensure smooth deployment and operation.

Solving the Challenge of Large-Scale, Multi-Node AI Infrastructure

Part of the **DGX platform**, NVIDIA DGX SuperPOD is designed to deliver unmatched levels of multi-node training. Traditional large compute clusters are constrained by the complexity of scaling inter-GPU communications as configurations become larger and computation is parallelized over more and more nodes. This results in diminishing performance returns. DGX SuperPOD solves this scaling problem by optimizing every component in the system for the unique demands of multi-node AI infrastructure. **NVIDIA's Eos**, one of the fastest supercomputers in the world, and other clusters based on the DGX SuperPOD architecture perennially make the top tier of the **TOP500** and **Green500** lists[1] and set MLPerf benchmark records.[2]

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, **NVIDIA AI Enterprise**, offering a suite of software to streamline AI development and deployment, is optimized to run on DGX systems.

Use **NVIDIA NIM™** inference microservices for optimal model deployment, offering speed, ease of use, manageability, and security.

NVIDIA DGX SuperPOD

Hardware

- NVIDIA DGX™ B200 or DGX H200 systems
- NVIDIA Networking
- High-performance storage

Software

- NVIDIA AI Enterprise
- NVIDIA DGX software

Lifecycle Services*

- **Plan/Deploy****
  - Capacity planning
  - Data center design
  - Performance projection
  - Site eval/prep
  - Installation
  - Post-install testing
  - Provisioning/management
- **Train/Optimize**
  - Application perf testing
  - Site documentation package
  - User/DevOps training
  - Workload-based NVIDIA Deep Learning Institute training
  - Custom system runbook
  - Hand-over session

* A combination of NVIDIA and partner services

** Deployed on-prem or in a DGX-Ready Data Center

NVIDIA DGX SuperPOD, Tested and Proven

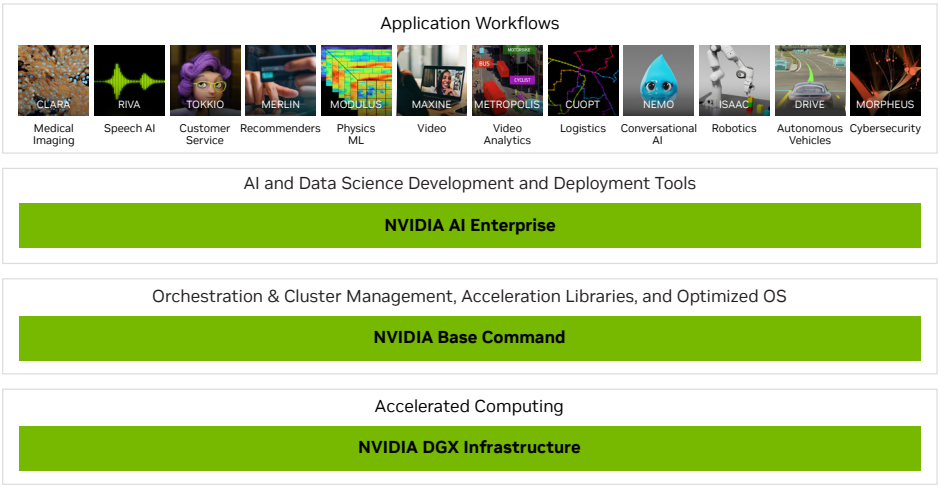
DGX SuperPOD isn't just AI infrastructure done the NVIDIA way—it's a predictable solution that meets the performance and reliability needs of enterprises. NVIDIA does all the leg work, testing DGX SuperPOD extensively and pushing it to the farthest limits with real-world enterprise AI workloads, so you don't have to worry about application performance.

A Complete Lifecycle of Expertise, Backed by NVIDIA

More than an architecture design, enterprises need a faster path to making accelerated computing infrastructure operationally useful to their businesses. They need an implementation experience that's turnkey, fast, and optimized around their IT environment—so data scientists can be up and running on day one—and continues to improve over time.

With NVIDIA DGX SuperPOD, enterprises benefit from full lifecycle infrastructure services spanning everything from install to infrastructure management to scaling workloads to streamlined production AI. And true to the promise of DGX SuperPOD, it continually gets better. NVIDIA engineers are continuously innovating and updating the software that powers DGX SuperPOD so every system runs faster than the day it was commissioned. With the DGX SuperPOD, customers also get direct access to NVIDIA **DGXperts**, a global team of AI-fluent practitioners that offer prescriptive guidance and design expertise to help fast-track AI transformation.

NVIDIA DGX - AI Software Stack



High-Performance Infrastructure in a Single Solution—Optimized for AI

NVIDIA DGX SuperPOD brings together a design-optimized combination of AI computing, network fabric, storage, and software. Its compute foundation is built on **NVIDIA DGX B200** or **DGX H200 systems**, which provide unprecedented compute density, performance, and flexibility. NVIDIA DGX B200 or DGX H200 systems feature the world's most advanced accelerators, enabling enterprises to consolidate training, inference, and analytics in a unified, easy-to-deploy AI infrastructure.

DGX SuperPOD's high-performance network fabric leverages ultra-low-latency **NVIDIA InfiniBand networking**. This powerful technology delivers the highest performance and scalability for the largest AI workloads, with reduced operational costs and infrastructure complexity.

AI supercomputers also require extremely high-speed storage to run at peak capacity. In a well-architected system, storage solutions need to handle a variety of data types—such as text, tabular data, audio, and video—in parellel—with unwavering performance. Certified storage for NVIDIA DGX SuperPOD is carefully selected and tested for the unique demands of AI workloads and then optimized for each environment to ensure success.

The Experience That Fuels AI Success

DGX SuperPOD incorporates NVIDIA's unmatched experience in designing and using AI supercomputers, driven by thousands of NVIDIA researchers and engineers who use this platform to bring new innovations to market. DGX SuperPOD delivers a turnkey data center solution and can be deployed by customers in their own data center or from a variety of **managed service providers**, **colocation options**, **private cloud offerings**, or **AI consulting partners**.

NVIDIA's DGX Supercomputer Eos is used for NVIDIA's research and development.

Ready to Get Started?

To learn more about NVIDIA DGX SuperPOD, visit:
www.nvidia.com/dgx-superpod

1. See top500.org for more information
2. See mlperf.org to read more.

Partner Logo

NVIDIA