Solution Overview

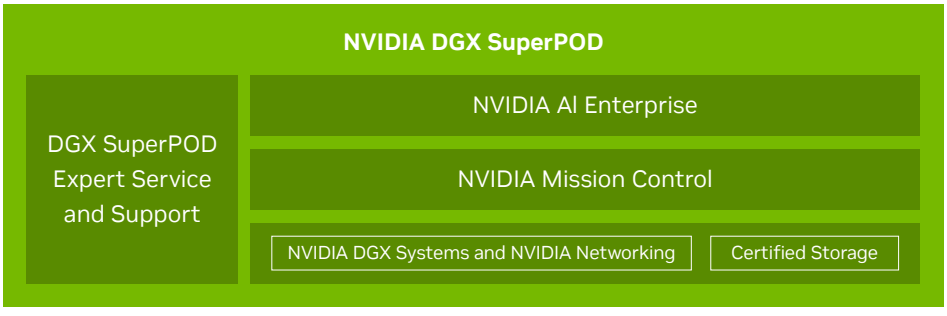
NVIDIA DGX SuperPOD

AI factory infrastructure for today's enterprise

NVIDIA DGX SuperPOD™ is an AI factory infrastructure solution that delivers uncompromising performance for every user and workload. DGX SuperPOD incorporates NVIDIA's unmatched experience in designing and using AI supercomputers and is a full-stack AI infrastructure solution that lets enterprises get up and running in weeks instead of months. DGX SuperPOD provides leadership-class accelerated infrastructure with scalable performance for the most demanding AI training and inference workloads, with industry-proven results, allowing IT to deliver performance without compromise.

Full-Stack AI Factory Infrastructure

NVIDIA DGX SuperPOD features a design-optimized combination of AI compute, network fabric, storage, and software. Its compute foundation offers high density, performance, and flexibility, interconnected with ultra-high bandwidth, low-latency NVIDIA networking. Certified storage solutions integrate a high-performance data fabric for fueling AI workloads. A fully optimized software stack that works across the entire stack provides world-class tools for developer workflows and complete cluster management to reduce operational burden and maximize utilization.



NVIDIA DGX SuperPOD architecture

DGX SuperPOD

Hardware

- NVIDIA DGX™ GB300, DGX GB200, DGX B300, or DGX B200 systems
- NVIDIA Networking
- High-performance storage

Software

- NVIDIA AI Enterprise
- NVIDIA Mission Control
- NVIDIA DGX Operating System

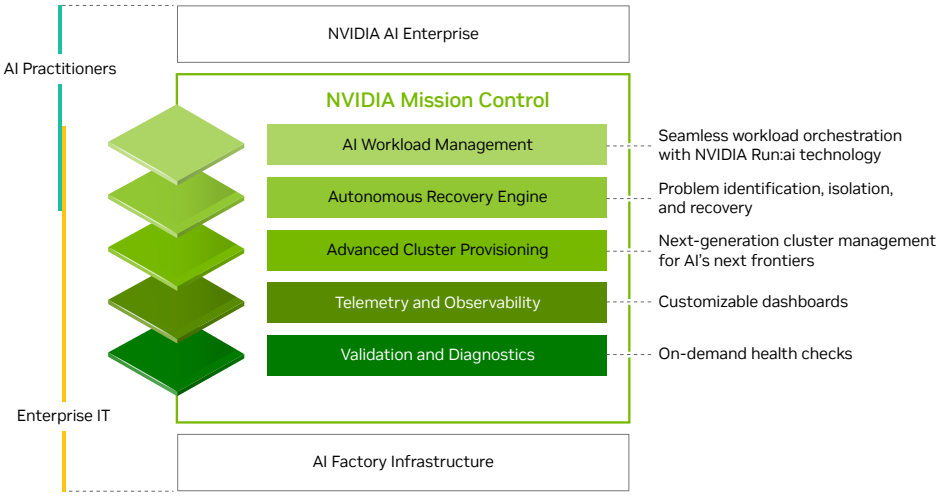
Lifecycle Services*

- **Plan/Deploy****
  - Data center/system planning and design
  - Off-site preparation
  - On-site physical and logical deployment
  - Testing and validation
- **Train/Optimize**
  - Application perf testing
  - Site documentation package
  - User/DevOps training
  - Workload-based NVIDIA Deep Learning Institute training
  - 24/7 Expert support

* A combination of NVIDIA and partner services
** Deployed on-prem or in a DGX-Ready Data Center

Run Models, Automate the Essentials With NVIDIA

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 development, offering speed, ease of use, manageability, and security.



NVIDIA Mission Control software powers NVIDIA DGX SuperPOD

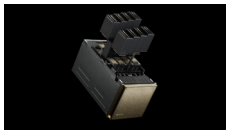
Massive Supercomputing for the Newest Frontiers of AI

Scaling to tens of thousands of NVIDIA GPUs, DGX SuperPOD tackles training and inference for state-of-the-art generative AI models. This leadership-class platform delivers the highest performance and scalability for the largest AI workloads, with reduced operational costs and infrastructure complexity. For AI inference, DGX SuperPOD offers the world's leading platform for inference at enterprise scale, delivering the largest NVIDIA NVLink domain, unmatched total GPU memory and inter-GPU memory bandwidth, and leadership-class compute performance. For AI training, supercomputing clusters based on the DGX SuperPOD architecture speed time-to-operation for enterprises and perennially rank within the top tier of the TOP500 and Green500 lists[1] and set MLPerf benchmark records.[2]

NVIDIA DGX GB300

NVIDIA DGX GB200

NVIDIA DGX B300



NVIDIA DGX B200

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 continue to improve over time.

With NVIDIA DGX SuperPOD, enterprises benefit from full lifecycle services and support from beginning to end by NVIDIA experts. Designated guidance and support cover planning and installation to validation to infrastructure management and throughout the production lifecycle. NVIDIA provides enterprises with access to experts covering the full stack of DGX SuperPOD to keep their AI workloads running at peak performance.

The Experience That Fuels AI Success

DGX SuperPOD incorporates the expertise of thousands of NVIDIA researchers and engineers who use the platform daily to bring new innovations to market. DGX SuperPOD delivers a turnkey AI factory solution and can be deployed by customers in their own data center or from a variety of managed service providers, colocation options, or private cloud offerings.

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:
microway.com/platforms/dgx/nvidia-dgx-superpod/

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

Microway | NVIDIA