NVIDIA DGX SUPERPOD SOLUTION FOR ENTERPRISE

Turnkey Data Center Product 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. NVIDIA DGX SuperPOD delivers a full-service experience with industry-proven results in weeks instead of months. It's not just a collection of hardware. It's a full-stack data center platform that includes industry-leading computing, storage, networking, software, and infrastructure management tools optimized to work together and provide maximum performance at scale, along with a white-glove implementation service that ensures smooth deployment and operation.

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

NVIDIA DGX SuperPOD is designed to tackle the most important challenges of AI at scale, delivering unmatched levels of multi-system 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. Using this architecture, NVIDIA created two of the fastest and most energy-efficient supercomputers in the world, which made the TOP500 and Green500 lists[1] and set multiple MLperf benchmark records.[2]

Infrastructure Management with NVIDIA Base Command Manager

To streamline operations, DGX SuperPOD features NVIDIA Base Command™ Manager. The same software used to manage thousands of NVIDIA's own systems, Base Command Manager is the best-of-breed infrastructure solution for provisioning and lifecycle management, monitoring, telemetry, logging, alerting, and scheduling.

DGX SUPERPOD SOLUTION FOR ENTERPRISE

HARDWARE/SOFTWARE

- 100-700 PFLOPS AI system
- 20-140 NVIDIA DGX A100 systems
- 1-10PB high-performance storage
- 200Gbps NVIDIA networking fabric
- NVIDIA CUDA-X™ and DGX software stack
- NVIDIA Base Command Manager

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

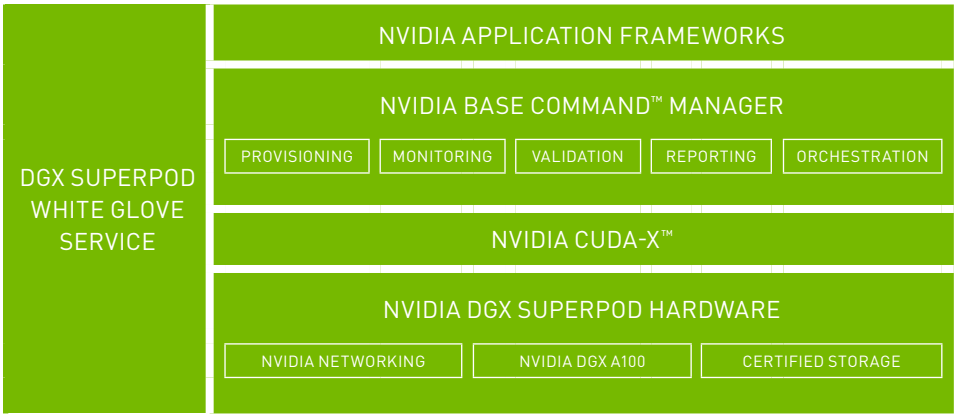
DGX SuperPOD, Tested and Proven for Every Customer

DGX SuperPOD isn't just AI infrastructure done the NVIDIA way. Every implementation is validated on a dedicated acceptance cluster at NVIDIA. The customer's design is replicated beforehand and a suite of performance results is produced—so when DGX SuperPOD is deployed on site, it runs exactly as it was intended.

For enterprises that aren't ready to purchase a DGX SuperPOD, fully managed DGX infrastructure based on the DGX SuperPOD architecture is available for short-term rental as part of NVIDIA DGX Foundry, a premium AI development platform.

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 their 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 professional 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's team of engineers is constantly innovating and improving the software that powers DGX SuperPOD—updates that are continuously delivered so every system runs faster than the day it was commissioned.



NVIDIA DGX SuperPOD Solution for Enterprise

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™ A100, the universal system for all AI workloads, which provides unprecedented compute density, performance, and flexibility. NVIDIA DGX A100 systems, available with up to 640 gigabytes (GB) of total GPU memory each, feature the world's most advanced accelerator, the NVIDIA A100 Tensor Core GPU, 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 parallel and 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.

To scale AI, enterprises need to integrate optimized software and data science workflows within an IT and DevOps approach. MLOps software streamlines AI application delivery, so data science teams and IT can more effectively manage users, models, datasets, experiments, and more, while speeding continuous application delivery. DGX SuperPOD includes fully optimized AI software from the NVIDIA NGC™ catalog to help organizations manage, scale, and accelerate AI and data science.

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. NVIDIA DGX SuperPOD delivers the turnkey data center solution for businesses focused on innovation instead of infrastructure, designed, deployed, and managed the way NVIDIA does AI.

1 See top500.org for more information | 2 See mlperf.org for more information