NVIDIA DGX SUPERPOD SOLUTION FOR ENTERPRISE

THE FASTEST PATH TO AI INNOVATION AT SCALE

NVIDIA DGX SuperPOD™ Solution for Enterprise is a first-of-its-kind infrastructure solution that enables any organization to operationalize AI at scale. Designed for enterprises that need the fastest path to AI innovation, this turnkey solution built on the NVIDIA DGX SuperPOD reference architecture gives businesses leadership-class infrastructure that can be rapidly deployed and is supported with a complete end-to-end lifecycle of services and support, all backed by NVIDIA.

NVIDIA DGX SuperPOD Solution for Enterprise 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 platform that includes industry-leading computing, storage, networking, infrastructure management, and data science workflow 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 Solution for Enterprise 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 returns in terms of performance as systems grow. NVIDIA DGX SuperPOD Solution for Enterprise solves this scaling problem by optimizing every component in the system for the unique demands of multi-node AI infrastructure. Built on the same DGX SuperPOD architecture, Selene—NVIDIA's own NVIDIA DGX SuperPOD solution deployment—is the seventh fastest supercomputer in the world and the second most energy efficient.[1] It has earned the top spot in the MLperf benchmark suite for commercially available solutions.[2]

Intelligently Adapted and Integrated With Your Business

Data science teams need the right tools, platform, and infrastructure to streamline AI workflows and speed time to insights. IT teams need the right partner to help augment their existing infrastructure and navigate the complexities of high-performance computing, network fabric, storage architecture, and AI software that are integral to scaling AI, with flexible deployment approaches that fit business and implementation time constraints. With NVIDIA DGX SuperPOD Solution for Enterprise, NVIDIA's professional services team will help optimize the solution to any environment, including flexible deployment options tailored to your unique requirements.

DGX SUPERPOD SOLUTION FOR ENTERPRISE

HARDWARE/SOFTWARE

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

LIFECYCLE SERVICES*

Plan/Deploy**

- Capacity planning
- Data center design
- Performance projection
- Site eval/prep
- Installation
- Post-install testing
- Provisioning/management

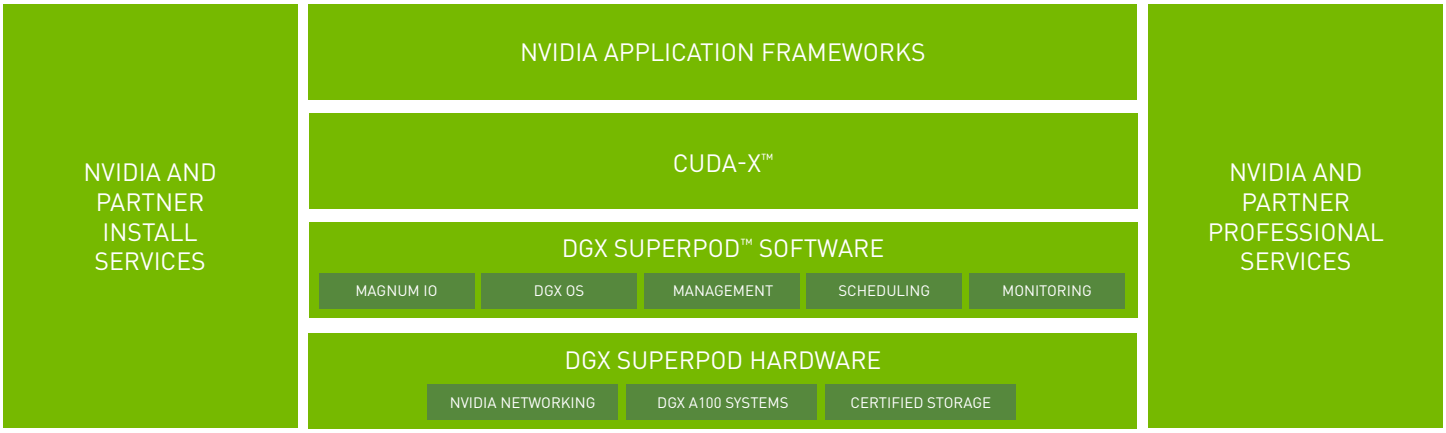
Train/Optimize

- 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

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. With NVIDIA DGX SuperPOD Solution for Enterprise, enterprises benefit from full-service data center planning and infrastructure delivery expertise that speeds deployment—from sizing, to installation, to training, to ongoing optimization and beyond—all backed by NVIDIA in combination with our DGX SuperPOD Solution for Enterprise partners.



NVIDIA DGX SuperPOD Solution for Enterprise

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

NVIDIA DGX SuperPOD Solution for Enterprise 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 features the world's most advanced accelerator, the NVIDIA A100 Tensor Core GPU, enabling enterprises to consolidate training, inference, and analytics into a unified, easy-to-deploy AI infrastructure.

NVIDIA Mellanox is the high-performance network fabric underpinning two-thirds of the world's TOP500 supercomputers, with innovative NVIDIA Mellanox InfiniBand™ in-network computing technologies, including NVIDIA Mellanox Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ and congestion control. This powerful combination delivers the highest performance and scalability, 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 to handle the enormous depth and diversity of AI data. Certified storage for NVIDIA DGX SuperPOD Solution for Enterprise is carefully selected and tested for the unique demands of AI workloads and then optimized for your environment to ensure success.

To scale AI, enterprises need to integrate optimized software and data science workflows within an IT/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 Solution for Enterprise includes fully optimized AI software from the NVIDIA NGC™ catalog and offers MLOps software from NVIDIA DGX-Ready Software partners to help organizations manage, scale, and accelerate AI and data science. This software stack provides a streamlined machine learning pipeline that enables data science practitioners and IT/DevOps teams to work together to get the highest performance and accelerate the deployment of production applications.

Our Experience Fuels Your Success

DGX SuperPOD Solution for Enterprise 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. This global team of AI experts use DGX SuperPOD every day and are ready to make your AI ambitions a reality.

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

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

NVIDIA