NVIDIA DGX SUPERPOD INSTANT INFRASTRUCTURE FOR AI LEADERSHIP

NVIDIA DGX SuperPOD™ is a first-of-its-kind AI supercomputing infrastructure that delivers groundbreaking performance, deploys in weeks as a fully integrated system, and is designed to solve the world's most challenging AI problems.

AI is transforming our planet and every facet of life as we know it, fueled by the next generation of leading-edge research. Organizations that want to lead in an AI-powered world know that the race is on to tackle the most complex AI models that demand unprecedented scale.

Leadership-Class AI Infrastructure

Our biggest challenges can only be answered with groundbreaking research that requires supercomputing power on an unmatched scale. Organizations that are ready to lead need to attract the world's best AI talent to fuel innovation and the leadership-class infrastructure that can get them there now, not months from now. From the experience of thousands of deployed AI environments, and the evolution of our own NVIDIA DGX SATURNV, we developed the industry-leading design formula for AI supercomputing, architected to shrink months of problem-solving into just minutes.

Building a world-class AI supercomputer normally takes six months or longer. In just two weeks, NVIDIA designed, built, and tested DGX SuperPOD. The plug-in, power-up simplicity of DGX-2, combined with an AI networking fabric from Mellanox and the proven architecture of NVIDIA DGX SATURNV, dramatically compressed time to results. Today, DGX SuperPOD powers NVIDIA's own innovations while delivering to the marketplace an instant AI infrastructure that can solve humanity's biggest challenges.

NVIDIA DGX SUPERPOD AT A GLANCE

Configuration	**64 nodes of NVIDIA DGX-2™** **1,024 NVIDIA® Tesla® V100 Tensor Core GPUs (DGX SuperPOD total)**
NVIDIA CUDA® Cores	**5,242,880 (DGX SuperPOD total)**
NVIDIA Tensor Cores	**655,360 (DGX SuperPOD total)**
NVIDIA NVSwitches™	**768 (DGX SuperPOD total)**
System Memory	**96 TB DDR4 (DGX SuperPOD total)** **32 TB GPU high-bandwidth memory (DGX SuperPOD total)**
Networking	**Mellanox CS7500 InfiniBand Switch (Compute)** **Mellanox CS7520 InfiniBand Switch (Storage)**
Storage	**Certified high-performance GPFS storage solutions (check with NVIDIA team)**

See the DGX-2 datasheet for node-level specifications.

DGX SuperPOD: Solving the Challenge of Extreme Multi-Node AI Training

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 increasing impact to performance associated with 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 gained by incremental compute nodes. NVIDIA DGX SuperPOD has demonstrated world-record-breaking performance and versatility in MLPerf 0.6[1], setting eight records in AI performance. These results offer proof of DGX SuperPOD's ability to solve this scaling problem with an ultra-dense compute solution that taps into the innovative architecture of DGX-2 combined with high-performance networking from Mellanox and high-performance storage.

DGX-2: Built for the World's Most Complex AI Challenges

As the compute foundation of DGX SuperPOD, DGX-2 incorporates the latest innovations in GPU-accelerated computing, including the integration of 16 NVIDIA V100 Tensor Core GPUs—the world's most advanced data center accelerator. The GPUs are fully interconnected using revolutionary NVIDIA NVSwitch technology, enabling direct communications between any GPU pair without bottlenecks. DGX-2 leverages the NVIDIA DGX software stack, which is optimized for maximum GPU-accelerated performance, for the world's most popular AI and deep learning applications.

DGX SuperPOD delivers over 128 petaFLOPS (PFLOPS) of AI performance, with 64 DGX-2 systems. This computing and engineering feat came together in just weeks, capturing a top spot among the world's most powerful supercomputers.

Every aspect of DGX SuperPOD is designed to tackle the world's most complex AI challenges, powering AI applications from speech-to-vision to robotics to autonomous systems. DGX SuperPOD has the performance and scale that can deliver insights to your team in minutes instead of days or weeks.

Mellanox Terabit-Speed AI Network Fabric

DGX SuperPOD can tackle the largest multi-node training problems thanks to a state-of-the-art, ultra-high-speed, low-latency fabric that interconnects each DGX-2 system, using terabit-speed InfiniBand networking to each DGX-2 node.

Access Our Global Team of AI Experts

DGX SuperPOD is more than great hardware. It's built on AI expertise that continually delivers higher performance, 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.

DGX SuperPOD simplifies the design, deployment, and operationalization of massive AI infrastructure with a validated design that's offered as a turnkey solution through our value-added resellers. Now, every enterprise can scale AI to address their most important challenges with a proven approach that's backed by 24x7 enterprise-grade support.

To learn more about NVIDIA DGX SuperPOD, visit www.nvidia.com/dgx-pod
To learn more about NVIDIA DGX-2, visit www.nvidia.com/DGX-2

[1] MLPerf 0.6 submission information: Per accelerator comparison using reported performance for NVIDIA DGX-2H servers (16 Tesla V100 GPUs) compared to other submissions at same scale except for MiniGo where the NVIDIA DGX-1 servers (8 Tesla V100 GPUs) submission was used.

MLPerf ID Max Scale: Mask R-CNN: 0.6-23, GNMT: 0.6-26, MiniGo: 0.6-11 | MLPerf ID Per Accelerator: Mask R-CNN, SSD, GNMT, Transformer: all use 0.6-20, MiniGo: 0.6-10. See mlperf.org for more information.