NVIDIA DGX SUPERPOD DELIVERING RECORD-BREAKING SUPERCOMPUTING TO EVERY ENTERPRISE

NVIDIA DGX SuperPOD™ is the first-of-its-kind AI supercomputing infrastructure to achieve groundbreaking TOP500 performance with an integrated solution that was designed, built, and deployed in record-breaking time.

The supercomputing world is evolving, driven by a re-thinking in how massive computing resources can converge to solve mission critical business problems. Infrastructure that once took years of planning and the integration of custom-designed componentry, is giving way to a new era ruled by GPU-accelerated technology. This is enabling enterprises to use standardized components, now systemized for deployment in months or even weeks. DGX SuperPOD offers a proven, enterprise-grade solution that builds upon the insights and design best practices gained from NVIDIA® DGX™ SATURNV.

From Concept to Results in Only Three Weeks

DGX SuperPOD started as a design concept to test the bounds of AI infrastructure and performance traditionally achieved by only the largest supercomputing clusters. Using NVIDIA DGX-2H as its foundational compute building block, DGX SuperPOD integrates 96 nodes, combined in a fully connected low-latency network architecture using InfiniBand switching and ultra-high-speed storage.

Building a TOP500-class supercomputer normally takes six months or longer. In just three weeks, NVIDIA designed, built, and tested DGX SuperPOD. The plug-in, power-up simplicity of DGX-2, combined with the proven architecture based on NVIDIA DGX SATURNV, enabled the dramatically compressed time-to-results achieved.

NVIDIA DGX SUPERPOD AT A GLANCE

Configuration	**96 nodes of NVIDIA DGX-2H** **1,536 NVIDIA Tesla® V100 Tensor Core GPUs (DGX SuperPOD total)**
NVIDIA CUDA® Cores	**7,864,320 (DGX SuperPOD Total)**
NVIDIA Tensor Cores	**983,040 (DGX SuperPOD Total)**
NVSwitches	**1,152 (DGX SuperPOD Total)**
System Memory	**144TB DDR4 (DGX SuperPOD Total)** **49TB GPU High-Bandwidth Memory (DGX SuperPOD total)** **See the DGX-2H datasheet for node-level specifications**
Networking	**Mellanox CS7510 Director Switches**
Storage	**IBM Spectrum GPFS GS4S**
Performance (approx.)	**FP64: 12 PFLOPS** **FP32: 25 PFLOPS** **FP16: 200 PFLOPS**

DGX SuperPOD: Solving the Challenge of Extreme AI and HPC Scale

DGX SuperPOD is designed to tackle the most important challenges of AI and HPC at scale. Traditional large compute clusters are constrained by the increasing costs 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. DGX SuperPOD solves this scaling problem by delivering an ultra-dense compute solution that taps into the innovative architecture found within each DGX-2.

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 the maximum GPU-accelerated performance, for the world's most popular AI and deep learning applications.

Record-Shattering AI and HPC Performance

DGX SuperPOD delivers almost 200 petaFLOPS of FP16 performance and ranked number 22 on the TOP500 supercomputing list, achieving results that have previously required hundreds of systems. DGX SuperPOD provides a proven, validated design approach for every organization that needs the compute power required of today's most demanding AI challenges and opportunities.

Get a TOP500 System without the Wait

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

To learn more about NVIDIA DGX-2, visit **www.nvidia.com/DGX-2**

NVIDIA.