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.
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. NVIDIA® DGX™ Systems also demonstrated world-class performance and versatility in MLPerf 0.6, setting eight records in AI performance. These results show that 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

---

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.

© 2019 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA DGX SATURNV, NVIDIA DGX SuperPOD, DGX, Tesla, CUDA, and NVSwitch are trademarks and/or registered trademarks of NVIDIA Corporation. All company and product names are trademarks or registered trademarks of the respective owners with which they are associated. Features, pricing, availability, and specifications are all subject to change without notice. JUL19