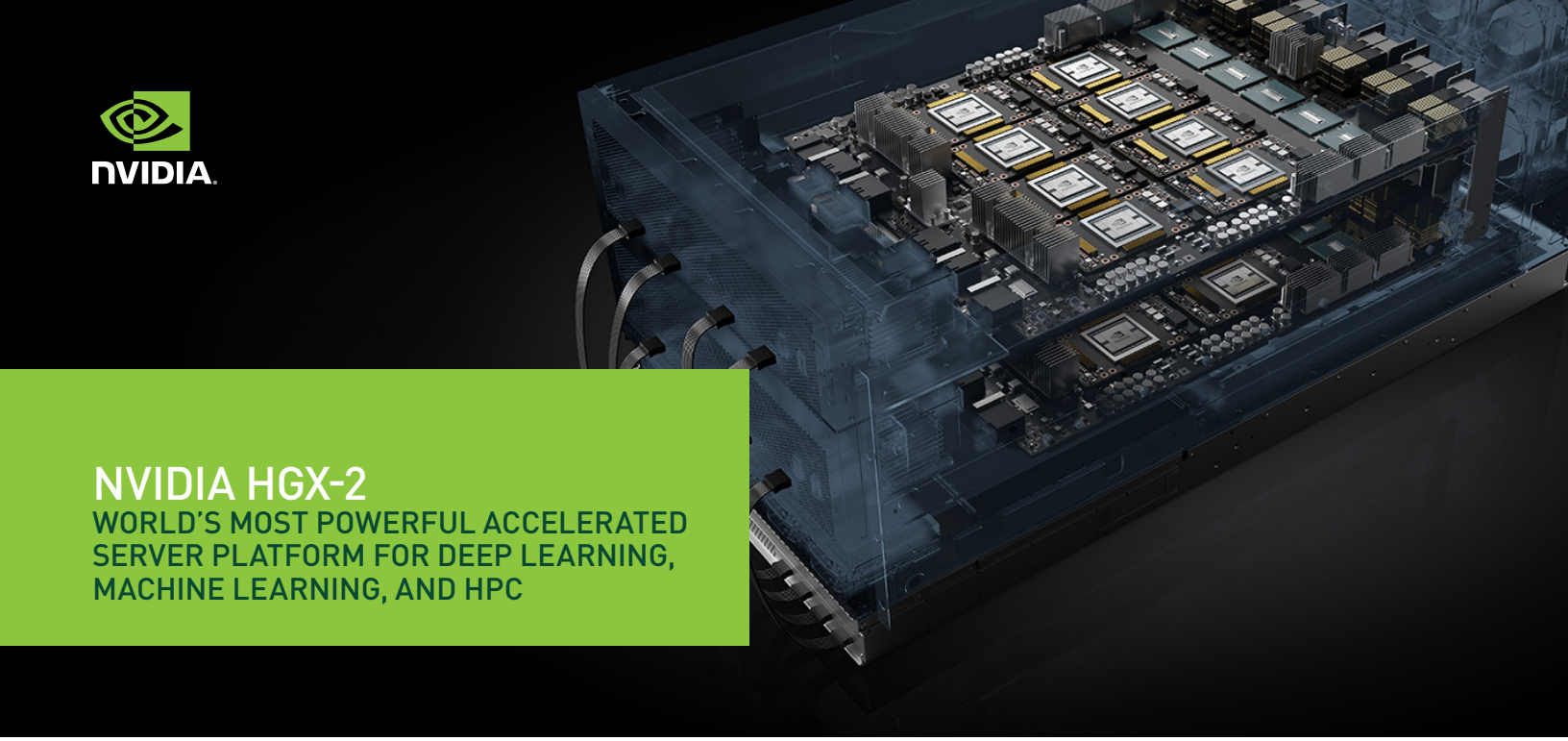


NVIDIA HGX-2

WORLD'S MOST POWERFUL ACCELERATED SERVER PLATFORM FOR DEEP LEARNING, MACHINE LEARNING, AND HPC

Designed for Larger, More Complex AI Models

AI deep learning neural networks and AI machine learning datasets are exploding in size and complexity to solve the most pressing challenges - in business and research.

The computational demands needed to support today's modern AI workloads have outpaced traditional data center architectures. As developers build increasingly large, accelerated computing clusters, they're pushing the limits of node density and data center scale. A new approach is needed—one that delivers almost limitless AI computing capability to achieve faster insights that can transform the world.

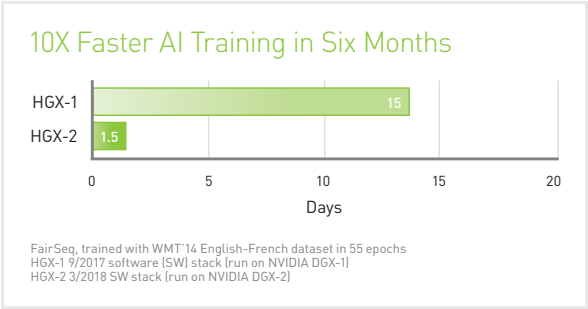
Redefining The Future of Computing

The NVIDIA HGX-2 multi-precision computing platform allows high-precision calculations using FP64 and FP32 for scientific computing and simulations, while also enabling FP16 and INT8 for deep learning and machine learning. This unprecedented versatility is uniquely positioned to support the future of computing.

Stack-to-stack enhancements from NVIDIA across hardware, software, and libraries, have accelerated AI training 10X in just six months.

SPECIFICATIONS

GPUs	**16x NVIDIA Tesla V100**
GPU Memory	**0.5TB total**
Performance	**2 petaFLOPS AI** **250 teraFLOPS FP32** **125 teraFLOPS FP64**
NVIDIA CUDA Cores	**81,920**
NVIDIA Tensor Cores	**10,240**
Communication Channel	**NVSwitch powered by NVLink** **2.4TB/sec aggregate speed**



NVIDIA NVSwitch for Full Bandwidth Computing

NVIDIA NVSwitch™ powered by NVIDIA NVLink™ creates a unified networking fabric that allows the entire node to function as a single gigantic GPU. Researchers can deploy models of unprecedented scale to solve the most complex HPC problems without being limited by compute capability.

Best-in-Industry Performance for Deep Learning, Machine Learning, and HPC

Today's most complex deep learning, machine learning, and HPC workloads demand highly parallel compute architectures. With NVIDIA's complete solution stack of hardware and software, users can solve problems at scale that were previously unsolvable. HGX-2 replaces 300 CPU servers for training, 544 CPU servers for machine learning and accelerates HPC 156X faster than a CPU-only server, making it the strongest compute node for data centers.

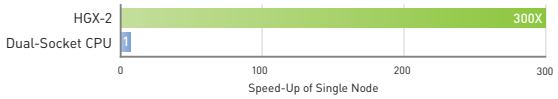
Design Versatility for the Cloud to Suit Any Workload

HGX-2 delivers a best-in-class server platform through GPU baseboards and a design guide that provides different configuration options. This allows unmatched versatility for the cloud by enabling server manufacturers to build a range of CPU and GPU machine instances ideal for different workloads.

Empowering the Data Center Ecosystem

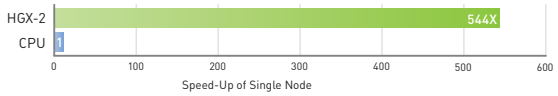
NVIDIA partners with the world's leading manufacturers to rapidly advance AI cloud computing by providing HGX-2 GPU baseboards, design guidelines and access to GPU computing technologies. Partners integrate these into servers and deliver AI and HPC at scale to their data center ecosystem.

AI Training: HGX-2 Replaces 300 CPU-Only Server Nodes



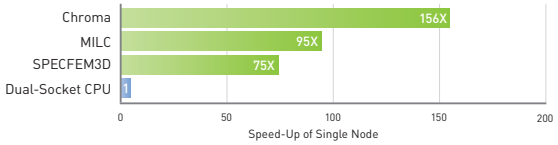
Workload: ResNet50, 90 epochs to solution | CPU Server: Dual-Socket Intel Xeon Gold 6140
Dataset: ImageNet2012

Machine Learning: HGX-2 544X Speedup Compared to CPU-Only Server Nodes



GPU Measurements Completed on DGX-2
CPU: 20 CPU cluster- comparison is prorated to 1 CPU (61 GB of memory, 8 vCPUs, 64-bit platform), Apache Spark
US Mortgage Data Fannie Mae and Freddie Mac 2006-2017 | 146M mortgages
Benchmark 200GB CSV dataset | Data preparation includes joins, variable transformations

HPC: HGX-2 Replaces Up To 156 CPU-only Server Nodes



Application(Dataset): SPECFEM3D(four material simple model), MILC(APEX Medium), and Chroma (szscl21_24_128) | CPU Server: Dual-Socket Intel Xeon Gold 6140

For more information, visit **www.nvidia.com/hgx**