

## NVIDIA HGX-2

### FUSING HPC AND AI COMPUTING INTO ONE UNIFIED ARCHITECTURE

#### Designed for Larger, More Complex AI Models

Deep neural networks are rapidly growing in size and complexity in response to 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 data center scale. A new approach is needed—one that delivers almost limitless AI computing to achieve faster insights that can transform the world.

#### Redefining The Future of Computing

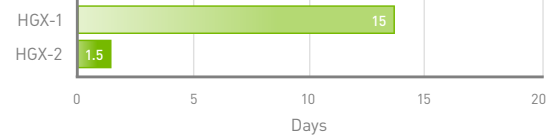
HGX-2 multi-precision computing platform allows high-precision calculations using FP64 and FP32 for scientific computing and simulations, while also enabling FP16 and Int8 for AI training and inference. This unprecedented versatility provides unique flexibility to support the future of computing.

NVIDIA's stack-to-stack enhancements across hardware, software, and libraries, provide up to 10X faster training of advanced AI in 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
Scalability	Manufacturing partners can build servers in the following configurations: > 1 baseboard (8x Tesla V100) > 2 baseboards (16x Tesla V100)

#### 10X Faster AI Training in Six Months



FairSeq, trained with WMT'14 English-French dataset in 55 epochs  
HGX-1 9/2017 software (SW) stack (run on NVIDIA DGX-1)  
HGX-2 3/2018 SW stack (run on NVIDIA DGX-2)

## 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 and solve the most complex HPC problems without being limited by compute capability.

## Best-in-Industry Performance for AI and HPC

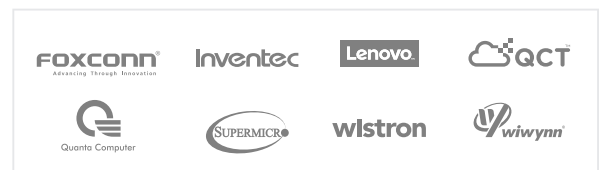
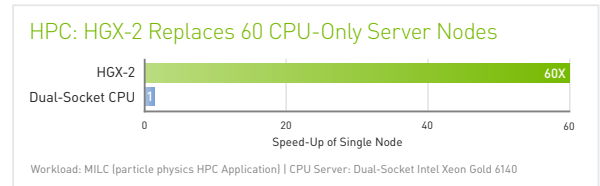
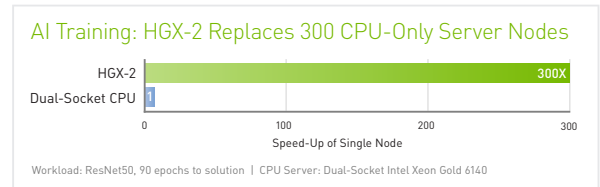
Today's most complex AI and HPC workloads demand highly parallel compute architectures. With NVIDIA's complete solution stack of hardware and software, users can solve problems at scale that were previously unsolvable. HGX-2 replaces 300 CPU servers for AI training and accelerates HPC 60X faster than 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—Foxconn, Inventec, Lenovo, QCT, Quanta, Supermicro, Wistron, and Wiyynn—to rapidly advance AI cloud computing. NVIDIA provides HGX-2 GPU baseboards, design guidelines, and early access to GPU computing technologies for partners to integrate into servers and deliver at scale to their data center ecosystem.



For more information, visit [www.nvidia.com/hgx](http://www.nvidia.com/hgx)

© 2018 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, Tesla, NVSwitch, NVLink, and CUDA 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. MAY18

