NVIDIA HGX-2
FUSING HPC AND AI COMPUTING INTO ONE UNIFIED ARCHITECTURE

RECORD PERFORMANCE
The HGX-2 platform is powered by NVIDIA NVSwitch™ which enables every GPU to communicate with every other GPU at full bandwidth of 2.4TB/sec to solve the largest of AI and HPC problems.

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.

EXPLOSION OF NETWORK COMPLEXITY
AI models are becoming increasingly complex and diverse, from translating languages to autonomous driving. Solving these models requires massive compute capability, large memory, and extremely fast connections between the GPUs.

16 NVIDIA® Tesla® V100 GPUs
0.5TB Aggregate High-Bandwidth GPU Memory
2 PFLOPS Total Compute

NVIDIA NVSwitch™ Direct GPU-to-GPU Connection Between All 16 GPUs

FUSING HPC AND AI COMPUTING INTO ONE UNIFIED ARCHITECTURE

EMPOWERING THE DATA CENTER ECOSYSTEM
NVIDIA works with a wide range of partners to deliver the ideal AI and HPC solution. With HGX-2, they can now accelerate to state-of-the-art AI and HPC problems.

SEE HOW HGX-2 CAN ACCELERATE YOUR AI AND HPC WORKLOADS.
www.nvidia.com/hgx