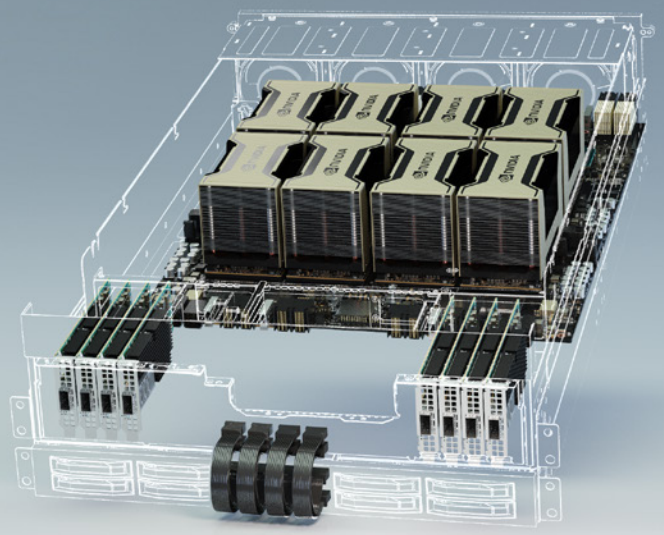
NVIDIA HGX A100

THE MOST POWERFUL END-TO-END AI SUPERCOMPUTING PLATFORM



Purpose-Built for the Convergence of Simulation, Data Analytics, and AI

Massive datasets, exploding model sizes, and complex simulations require multiple GPUs with extremely fast interconnections. The NVIDIA HGX™ platform brings together the full power of NVIDIA GPUs, NVIDIA® NVLink®, NVIDIA Mellanox® InfiniBand® networking, and a fully optimized NVIDIA AI and HPC software stack from NGC™ to provide highest application performance. With its end-to-end performance and flexibility, NVIDIA HGX enables researchers and scientists to combine simulation, data analytics, and AI to advance scientific progress.

With a new generation of A100 80GB GPUs, a single HGX A100 now has up to 1.3 terabytes (TB) of GPU memory and a world's-first 2 terabytes second (TB/s) of memory bandwidth, delivering unprecedented acceleration for emerging workloads, fueled by exploding model sizes and massive data-sets.

Third-Generation NVIDIA NVLink Creates a Single Super GPU

Scaling applications across multiple GPUs requires extremely fast movement of data. The third generation of NVIDIA NVLink in the NVIDIA A100 Tensor Core GPU doubles the GPU-to-GPU direct bandwidth to 600 gigabytes per second (GB/s), almost 10X higher than PCIe Gen4. Third-generation NVLink is available in four-GPU and eight-GPU HGX A100 servers from leading computer makers.

Second-Generation NVIDIA NVSwitch Drives Full-Bandwidth Computing

NVIDIA NVSwitch™ powered by 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.

SYSTEM SPECIFICATIONS (PEAK PERFORMANCE)

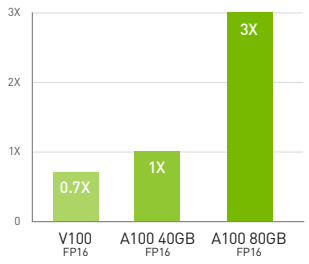
	4-GPU	8-GPU	16-GPU
GPUs	4x NVIDIA A100	8x NVIDIA A100	16x NVIDIA A100
HPC and AI Compute FP64/ TF32*/FP16*/ INT8*	78 TF/1.25PF*/ 2.5 PF*/5 POPS*	156 TF/2.5 PF*/ 5 PF*/10 POPS*	312 TF/5 PF*/ 10 PF*/20 POPS*
Memory	Up to 320 GB	Up to 640 GB	Up to 1,280 GB
NVIDIA NVLink	3rd generation	3rd generation	3rd generation
NVIDIA NVSwitch	N/A	2nd generation	2nd generation
NVSwitch GPU-to-GPU Bandwidth	N/A	600 GB/s	600 GB/s
Total Aggregate Bandwidth	2.4 TB/s	4.8 TB/s	9.6 TB/s

* With sparsity

Incredible Performance Across Workloads

Up to 3X Higher AI Training on Largest Models

DLRM Training

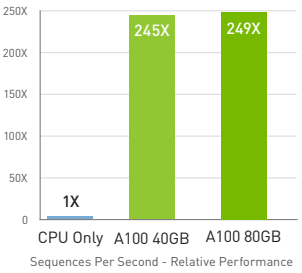


Time Per 1,000 Iterations - Relative Performance

DLRM on HugeCTR framework, precision = FP16 | NVIDIA A100 80GB batch size = 48 | NVIDIA A100 40GB batch size = 32 | NVIDIA V100 32GB batch size = 32.

Up to 249X Higher AI Inference Performance over CPUs

BERT-LARGE Inference

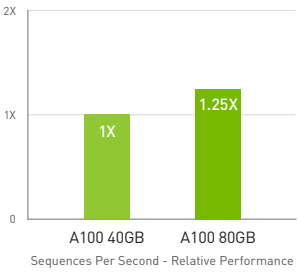


Sequences Per Second - Relative Performance

BERT-Large Inference | CPU only: Dual Xeon Gold 6240 @2.60 GHz, precision = FP32, batch size = 128 | V100: NVIDIA TensorRT™ (TRT) 7.2, precision = INT8, batch size = 256 | A100 40GB and 80GB, batch size = 256, precision = INT8 with sparsity.

Up to 1.25X Higher AI Inference Performance over A100

RNN-T Inference: Single Stream

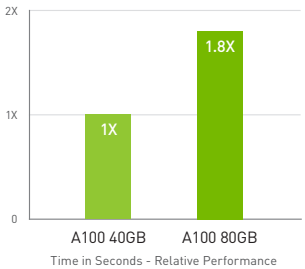


Sequences Per Second - Relative Performance

MLPerf 0.7 RNN-T measured with (1/7) MIG slices. Framework: TensorRT 7.2, dataset = LibriSpeech, precision = FP16.

Up to 1.8X Higher Performance for HPC Applications

Quantum Expresso

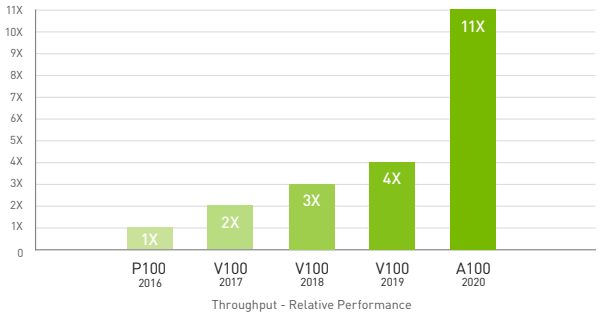


Time in Seconds - Relative Performance

Quantum Espresso measured using CNT10POR8 dataset, precision = FP64.

11X More HPC Performance in Four Years

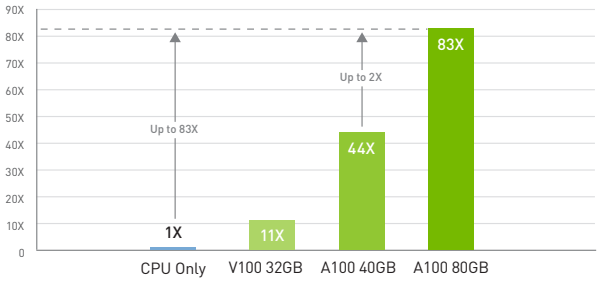
Throughput for Top HPC Apps



Throughput - Relative Performance

Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT-Large Fine Tuner], Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs [160000 x 64: 10]], TensorFlow [ResNet-50], VASP 6 [Si Huge] | GPU node with dual-socket CPUs with 4x NVIDIA P100, V100, or A100 GPUs.

Up to 83X Faster than CPU, 2X Faster than A100 40GB on Big Data Analytics Benchmark



Time to Solution - Relative Performance

Big data analytics benchmark | 30 analytical retail queries, ETL, ML, NLP on 10TB dataset | CPU: Intel Xeon Gold 6252 2.10 GHz, Hadoop | V100 32GB, RAPIDS/Dask | A100 40GB and A100 80GB, RAPIDS/Dask/BlazingSQL

Multi-Instance GPU (MIG) Delivers Seven Accelerators in a Single GPU

Every AI and HPC application can benefit from acceleration, but not every application needs the performance of a full A100 Tensor Core GPU. With MIG, each A100 can be partitioned into as many as seven GPU instances, fully isolated at the hardware level with their own high-bandwidth memory, cache, and compute cores. This allows HGX A100 systems to offer up to 112 GPU instances, giving developers access to breakthrough speed for every application, big and small, with guaranteed quality of service.

With A100 80GB, seven MIGs can be configured with 10 GB each (double the size of A100 40GB MIGs), making it now possible to perform inference on batch-size constrained models like BERT-LARGE (a natural language processing model with superhuman understanding) at much higher batch sizes, delivering up to a 1.3X increase in throughput.

Third-Generation Tensor Cores Redefine the Future of AI and HPC

First introduced in the NVIDIA Volta™ architecture, NVIDIA Tensor Core technology has reduced AI training times from weeks to hours and provided massive acceleration for inference operations. The third generation of Tensor Cores in the NVIDIA Ampere architecture builds upon these innovations by providing up to 20X more floating operations per second (FLOPS) for AI applications and up to 2.5X more FLOPS for FP64 HPC applications.

NVIDIA HGX A100 4-GPU delivers nearly 80 teraFLOPS of FP64 performance for the most demanding HPC workloads. NVIDIA HGX A100 8-GPU provides 5 petaFLOPS of FP16 deep learning compute. And the HGX A100 16-GPU configuration achieves a staggering 10 petaFLOPS, creating the world's most powerful accelerated server platform for AI and HPC.

Bandwidth and Scalability Power High-Performance Data Analytics

HGX A100 servers deliver the necessary compute power—along with an industry first 2 terabytes per second (TB/s) of memory bandwidth, along with the scalability of NVLink and NVSwitch—to tackle high-performance data analytics and support massive data-sets. Combined with NVIDIA Mellanox Infiniband, the Magnum IO software, GPU-accelerated Spark 3.0, and NVIDIA RAPIDS™, the NVIDIA data center platform can now accelerate these massive workloads at unprecedented levels of performance and efficient data center scale.

NVIDIA