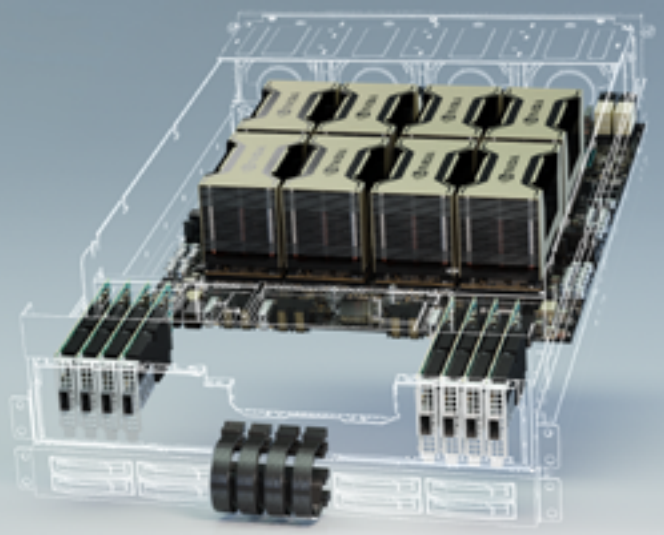
NVIDIA HGX A100

UNPRECEDENTED ACCELERATION AND VERSATILITY FOR THE HIGHEST-PERFORMING, ELASTIC DATA CENTERS



Design Versatility to Suit Any Workload

NVIDIA HGX™ A100 delivers a best-in-class server platform through GPU baseboards and a design guide that provides different configuration options. This allows unmatched versatility, enabling server manufacturers to build a range of CPU and GPU systems or cloud instances ideal for different workloads.

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 high-performance computing (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*	78TF/1.25PF*/ 2.5PF*/5POPS*	156TF/2.5PF*/ 5PF*/10POPS*	312TF/5PF*/ 10PF*/20POPS*
Memory	160 GB	320 GB	640 GB
NVIDIA NVLink	3rd generation	3rd generation	3rd generation
NVIDIA NVSwitch	N/A	2nd generation	2nd generation
GPU-to-GPU Bandwidth	600 GB/s	600 GB/s	600 GB/s
Total Aggregate Bandwidth	2.4 TB/s	4.8 TB/s	9.6 TB/s

* With sparsity

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 Multi-Instance GPU (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 acceleration for every application, big and small, with guaranteed quality of service.

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

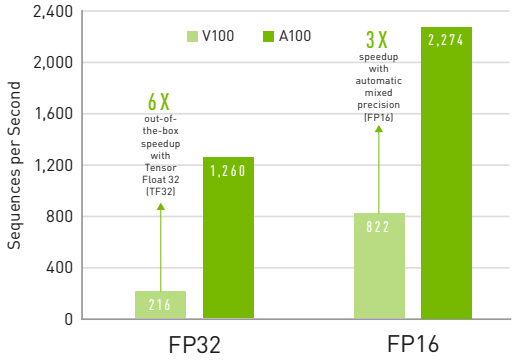
First introduced in the NVIDIA Volta™ architecture, NVIDIA Tensor Core technology has brought AI training times down from weeks to hours and provided massive acceleration to inference operations. The third generation of Tensor Cores in the NVIDIA Ampere architecture builds upon these innovations by providing 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. Abd 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

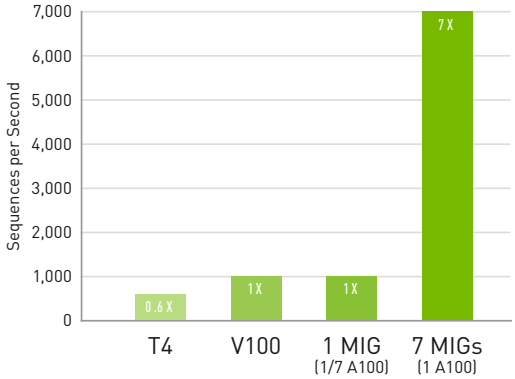
HGX A100 servers deliver the necessary compute power—along with 1.6 terabytes per second (TB/s) of memory bandwidth and the scalability of NVLink and NVSwitch—to tackle high-performance data analytics and support massive datasets. Combined with NVIDIA Mellanox Infiniband, the Magnum IO software, GPU-accelerated Spark 3.0, and NVIDIA RAPIDS™, the NVIDIA data center platform can accelerate these massive workloads at unprecedented levels of performance and efficient data center scale.

BERT Large Training



BERT Pre-Training Throughput using Pytorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 V100: DGX-1 Server with 8xV100 using FP32 and FP16 precision A100: DGX A100 Server with 8xA100 using TF32 precision and FP16 |

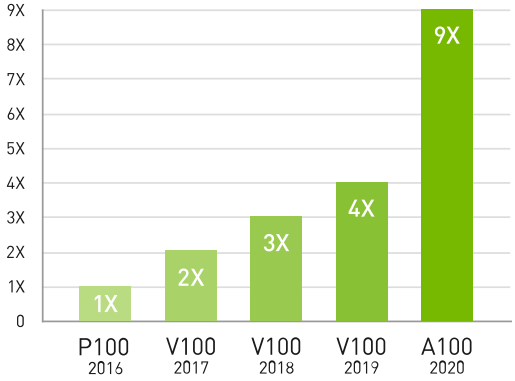
BERT Large Inference



BERT Large Inference | T4: TRT 7.1, Precision = INT8, Batch Size =256, V100: TRT 7.1, Precision = FP16, Batch Size =256 | A100 with 7 MIG instances of 1g.5gb : Pre-production TRT, Batch Size =94, Precision = INT8 with Sparsity

9X More HPC Performance in 4 Years

Throughput for Top HPC Apps



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.

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

NVIDIA