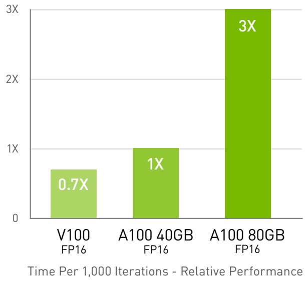
Incredible Performance Across Workloads

Up to 3X Higher AI Training on Largest Models

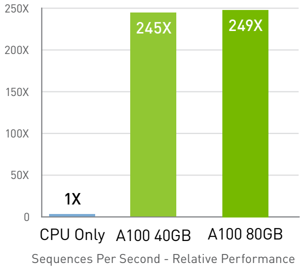
DLRM Training



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

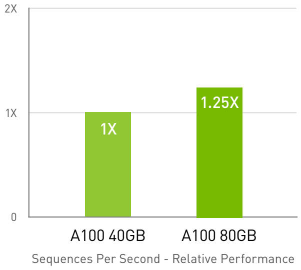
BERT-LARGE Inference



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 40GB

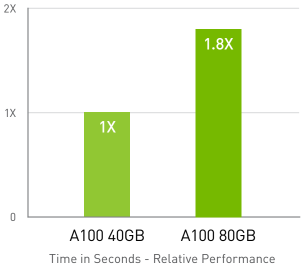
RNN-T Inference: Single Stream



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

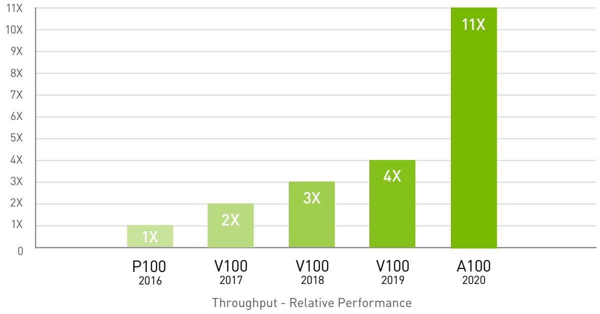
Quantum Espresso



Quantum Espresso measured using CNT10POR8 dataset, precision = FP64.

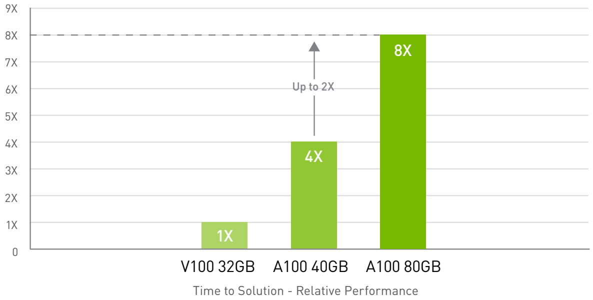
11X More HPC Performance in Four Years

Throughput for Top HPC Apps



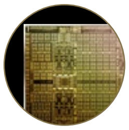
Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT-Large Fine Tuner], Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs (160000 x 64: 10)],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

2X Faster than A100 40GB on Big Data Analytics Benchmark



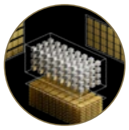
Big data analytics benchmark | GPU-BDB is derived from the TPCx-BB benchmark and is used for internal performance testing. Results from GPU-BDB are not comparable to TPCx-BB | 30 analytical retail queries, ETL, ML, NLP on 10TB dataset | V100 32GB, RAPIDS/Dask | A100 40GB and A100 80GB, RAPIDS/Dask/BlazingSQL

Groundbreaking Innovations



NVIDIA AMPERE ARCHITECTURE

Whether using MIG to partition an A100 GPU into smaller instances or NVLink to connect multiple GPUs to speed large-scale workloads, A100 can readily handle different-sized acceleration needs, from the smallest job to the biggest multi-node workload. A100's versatility means IT managers can maximize the utility of every GPU in their data center, around the clock.

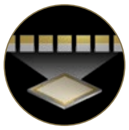


THIRD-GENERATION TENSOR CORES

NVIDIA A100 delivers 312 teraFLOPS (TFLOPS) of deep learning performance. That's 20X the Tensor floating-point operations per second (FLOPS) for deep learning training and 20X the Tensor tera operations per second (TOPS) for deep learning inference compared to NVIDIA Volta GPUs.

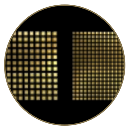
NEXT-GENERATION NVLINK

NVIDIA NVLink in A100 delivers 2X higher throughput compared to the previous generation. When combined with NVIDIA NVSwitch™, up to 16 A100 GPUs can be interconnected at up to 600 gigabytes per second (GB/sec), unleashing the highest application performance possible on a single server. NVLink is available in A100 SXM GPUs via HGX A100 server boards and in PCIe GPUs via an NVLink Bridge for up to 2 GPUs.



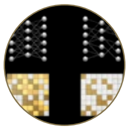
MULTI-INSTANCE GPU (MIG)

An A100 GPU 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. MIG gives developers access to breakthrough acceleration for all their applications, and IT administrators can offer right-sized GPU acceleration for every job, optimizing utilization and expanding access to every user and application.



HIGH-BANDWIDTH MEMORY (HBM2E)

With up to 80 gigabytes of HBM2e, A100 delivers the world's fastest GPU memory bandwidth of over 2TB/s, as well as a dynamic random-access memory (DRAM) utilization efficiency of 95%. A100 delivers 1.7X higher memory bandwidth over the previous generation.



STRUCTURAL SPARSITY

AI networks have millions to billions of parameters. Not all of these parameters are needed for accurate predictions, and some can be converted to zeros, making the models "sparse" without compromising accuracy. Tensor Cores in A100 can provide up to 2X higher performance for sparse models. While the sparsity feature more readily benefits AI inference, it can also improve the performance of model training.