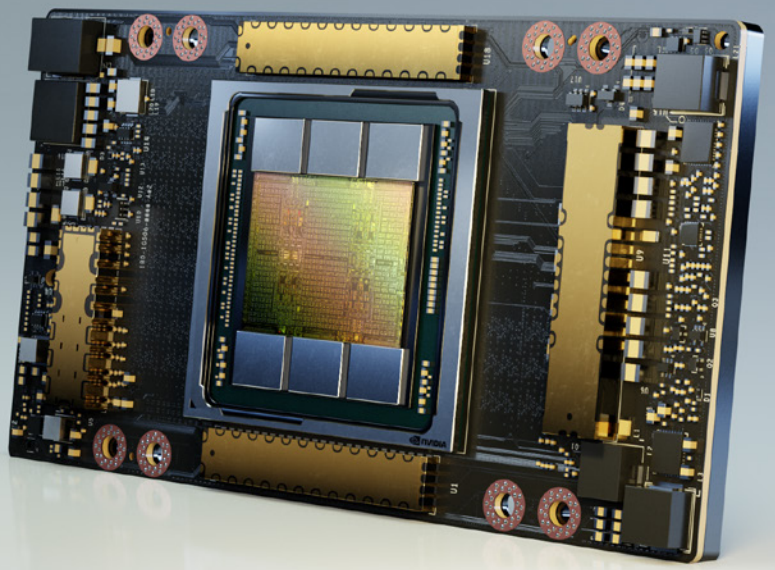


NVIDIA A100 TENSOR CORE GPU

UNPRECEDENTED SCALE AT EVERY SCALE

The Most Powerful Compute Platform for Every Workload

The NVIDIA® A100 Tensor Core GPU delivers unprecedented acceleration—at every scale—to power the world's highest-performing elastic data centers for AI, data analytics, and high-performance computing (HPC) applications. As the engine of the NVIDIA data center platform, A100 provides up to 20X higher performance over the prior NVIDIA Volta™ generation. A100 can efficiently scale up or be partitioned into seven isolated GPU instances, with Multi-Instance GPU (MIG) providing a unified platform that enables elastic data centers to dynamically adjust to shifting workload demands.

NVIDIA A100 Tensor Core technology supports a broad range of math precisions, providing a single accelerator for every workload. The latest generation A100 80GB doubles GPU memory and debuts the world's fastest memory bandwidth at 2 terabytes per second (TB/s), speeding time to solution for the largest models and most massive data sets.

A100 is part of the complete NVIDIA data center solution that incorporates building blocks across hardware, networking, software, libraries, and optimized AI models and applications from NGC™. Representing the most powerful end-to-end AI and HPC platform for data centers, it allows researchers to deliver real-world results and deploy solutions into production at scale.

SYSTEM SPECIFICATIONS

	NVIDIA A100 for NVLink		NVIDIA A100 for PCIe
Peak FP64	**9.7 TF**		**9.7 TF**
Peak FP64 Tensor Core	**19.5 TF**		**19.5 TF**
Peak FP32	**19.5 TF**		**19.5 TF**
Tensor Float 32 (TF32)	**156 TF | 312 TF***		**156 TF | 312 TF***
Peak BFLOAT16 Tensor Core	**312 TF | 624 TF***		**312 TF | 624 TF***
Peak FP16 Tensor Core	**312 TF | 624 TF***		**312 TF | 624 TF***
Peak INT8 Tensor Core	**624 TOPS | 1,248 TOPS***		**624 TOPS | 1,248 TOPS***
Peak INT4 Tensor Core	**1,248 TOPS | 2,496 TOPS***		**1,248 TOPS | 2,496 TOPS***
GPU Memory	**40GB**	**80GB**	**40GB**
GPU Memory Bandwidth	**1,555 GB/s**	**2,039 GB/s**	**1,555 GB/s**
Interconnect	**NVIDIA NVLink 600 GB/s****		**NVIDIA NVLink 600 GB/s****
	PCIe Gen4 64 GB/s		**PCIe Gen4 64 GB/s**
Multi-Instance GPU	**Various instance sizes with up to 7 MIGs @ 10 GB**		**Various instance sizes with up to 7 MIGs @ 5 GB**
Form Factor	**4/8 SXM on NVIDIA HGX™ A100**		**PCIe**
Max TDP Power	**400 W**	**400 W**	**250 W**

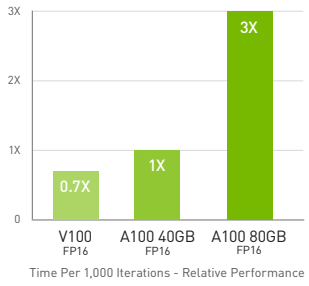
* With sparsity

** SXM GPUs via HGX A100 server boards; PCIe GPUs via NVLink Bridge for up to 2 GPUs

Incredible Performance Across Workloads

Up to 3X Higher AI Training on Largest Models

DLRM Training

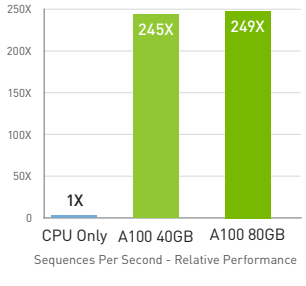


Time Per 1,000 Iterations - Relative Performance

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

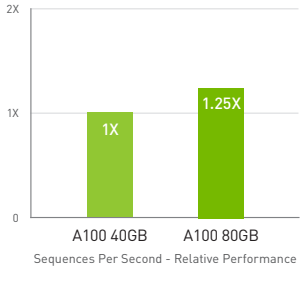


Sequences Per Second - Relative Performance

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 40GB

RNN-T Inference: Single Stream

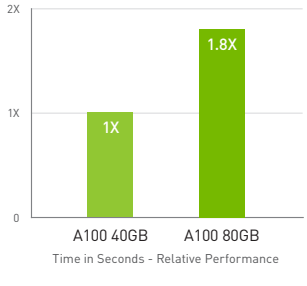


Sequences Per Second - Relative Performance

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 Espresso

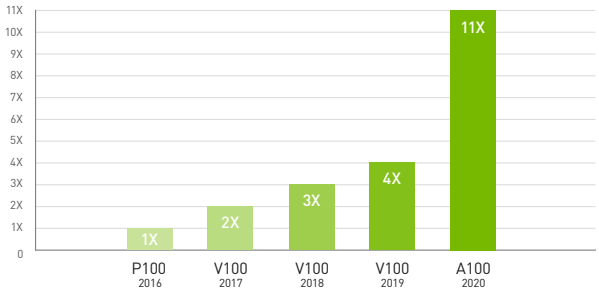


Time in Seconds - Relative Performance

Quantum Espresso measured using CNT10POR8 dataset, precision = FP64.

11X More HPC Performance in Four Years

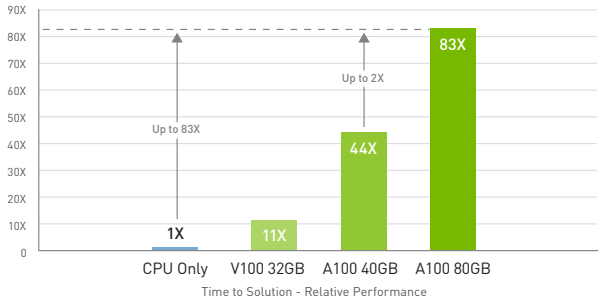
Throughput for Top HPC Apps



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



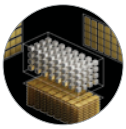
Time to Solution - Relative Performance

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

Groundbreaking Innovations

NVIDIA AMPERE ARCHITECTURE

Whether using MIG to partition an A100 GPU into smaller instances, or NVIDIA 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 versatility means IT managers can maximize the utility of every GPU in their data center, around the clock.



THIRD-GENERATION TENSOR CORES

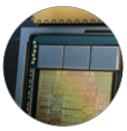
NVIDIA A100 delivers 312 teraFLOPS (TFLOPS) of deep learning performance. That's 20X the Tensor FLOPS for deep learning training and 20X the Tensor TOPS for deep learning inference, compared to NVIDIA Volta GPUs.

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.



HBM2E

With up to 80 gigabytes (GB) of high-bandwidth memory (HBM2e), A100 delivers a world's first GPU memory bandwidth of over 2TB/sec, as well as higher dynamic random-access memory (DRAM) utilization efficiency at 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.

The NVIDIA A100 Tensor Core GPU is the flagship product of the NVIDIA data center platform for deep learning, HPC, and data analytics. The platform accelerates over 1,800 applications, including every major deep learning framework. A100 is available everywhere, from desktops to servers to cloud services, delivering both dramatic performance gains and cost-saving opportunities.

EVERY DEEP LEARNING FRAMEWORK

mxnet

PYTORCH

Apache Spark

TensorFlow

1800+ GPU ACCELERATED APPLICATIONS

- Altair nanoFluidX
- Altair ultraFluidX
- AMBER
- ANSYS Fluent
- DS SIMULIA Abaqus
- GAUSSIAN
- GROMACS
- NAMD
- OpenFOAM
- VASP
- WRF

To learn more about the NVIDIA A100 Tensor Core GPU, visit **www.nvidia.com/a100**

© 2021 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA, DGX, HGX, HGX A100, NVLink, NVSwitch, OpenACC, TensorRT, and Volta are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc. All other trademarks and copyrights are the property of their respective owners. JAN21

NVIDIA