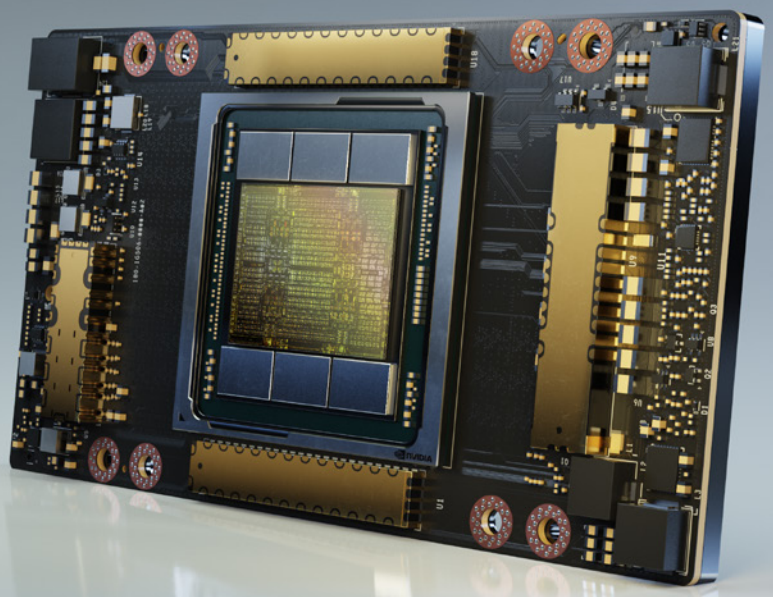
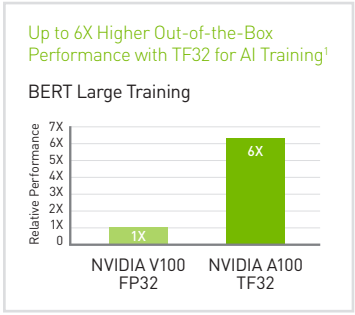
NVIDIA A100
TENSOR CORE GPU



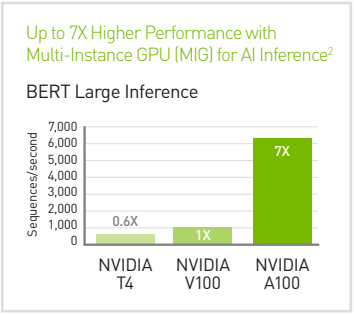
Unprecedented Acceleration at Every Scale

The NVIDIA A100 Tensor Core GPU delivers unprecedented acceleration at every scale for AI, data analytics, and HPC to tackle the world's toughest computing challenges. As the engine of the NVIDIA data center platform, A100 can efficiently scale up to thousands of GPUs or, using new Multi-Instance GPU (MIG) technology, can be partitioned into seven isolated GPU instances to accelerate workloads of all sizes. A100's third-generation Tensor Core technology now accelerates more levels of precision for diverse workloads, speeding time to insight as well as time to market.

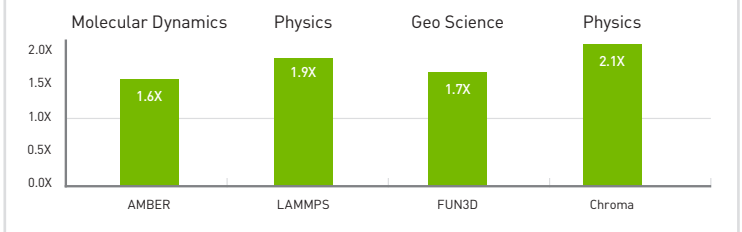
Up to 6X Higher Out-of-the-Box Performance with TF32 for AI Training[1]



Up to 7X Higher Performance with Multi-Instance GPU (MIG) for AI Inference[2]



Up to 2X More HPC performance[3]



SYSTEM SPECIFICATIONS (PEAK PERFORMANCE)

	NVIDIA A100 SXM4 for NVIDIA HGX™	NVIDIA A100 PCIe GPU
GPU Architecture	NVIDIA Ampere	
Double-Precision Performance	FP64: 9.7 TFLOPS FP64 Tensor Core: 19.5 TFLOPS	
Single-Precision Performance	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS | 312 TFLOPS*	
Half-Precision Performance	312 TFLOPS | 624 TFLOPS*	
Bfloat16	312 TFLOPS | 624 TFLOPS*	
Integer Performance	INT8: 624 TOPS | 1,248 TOPS* INT4: 1,248 TOPS | 2,496 TOPS*	
GPU Memory	40 GB HBM2	
Memory Bandwidth	1.6 TB/sec	
Error-Correcting Code	Yes	
Interconnect Interface	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**
Form Factor	4/8 SXM GPUs in NVIDIA HGX™ A100	PCIe
Multi-Instance GPU (MIG)	Up to 7 GPU instances	
Max Power Consumption	400 W	250 W
Delivered Performance for Top Apps	100%	90%
Thermal Solution	Passive	
Compute APIs	CUDA®, DirectCompute, OpenCL™, OpenACC®	

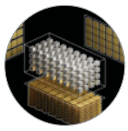
* Structural sparsity enabled

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

GROUNDBREAKING INNOVATIONS

NVIDIA AMPERE ARCHITECTURE

A100 accelerates workloads big and small. Whether using MIG to partition an A100 GPU into smaller instances, or NVLink to connect multiple GPUs to accelerate 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

A100 delivers 312 teraFLOPS (TFLOPS) of deep learning performance. That's 20X Tensor FLOPS for deep learning training and 20X 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) to unleash 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.

HBM2

With 40 gigabytes (GB) of high-bandwidth memory (HBM2), A100 delivers improved raw bandwidth of 1.6TB/sec, as well as higher dynamic random-access memory (DRAM) utilization efficiency at 95 percent. A100 delivers 1.7X higher memory bandwidth over the previous generation.

STRUCTURAL SPARSITY

AI networks are big, having millions to billions of parameters. Not all of these parameters are needed for accurate predictions, and some can be converted to zeros to make 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 700 HPC applications and every major deep learning framework. It's available everywhere, from desktops to servers to cloud services, delivering both dramatic performance gains and cost-saving opportunities.

EVERY DEEP LEARNING FRAMEWORK

Caffe2 · Microsoft Cognitive Toolkit · mxnet · PYTORCH · Apache Spark · TensorFlow · theano

700+ GPU-ACCELERATED APPLICATIONS

- AMBER
- ANSYS Fluent
- GAUSSIAN
- GROMACS
- LS-DYNA
- NAMD
- OpenFOAM
- Simulia Abaqus
- VASP
- WRF

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

1 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: NVIDIA DGX-1™ server with 8x NVIDIA V100 Tensor Core GPU using FP32 precision | A100: NVIDIA DGX™ A100 server with 8x A100 using TF32 precision.

2 BERT large inference | NVIDIA T4 Tensor Core GPU: NVIDIA TensorRT™ (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.

3 V100 used is single V100 SXM2. A100 used is single A100 SXM4. AMBER based on PME-Cellulose, LAMMPS with Atomic Fluid LJ-2.5, FUN3D with dpw, Chroma with szscl21_24_128.

© 2020 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. JUL20

NVIDIA