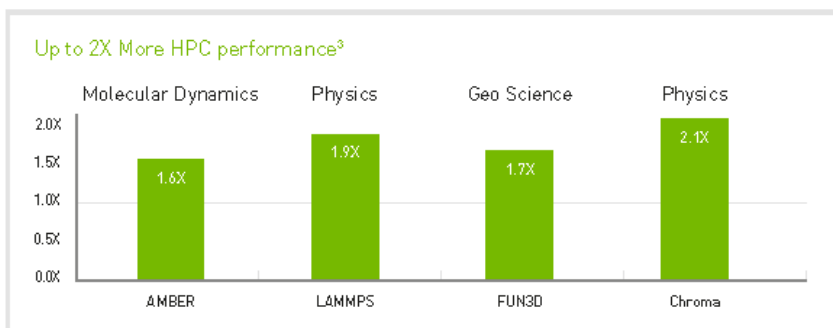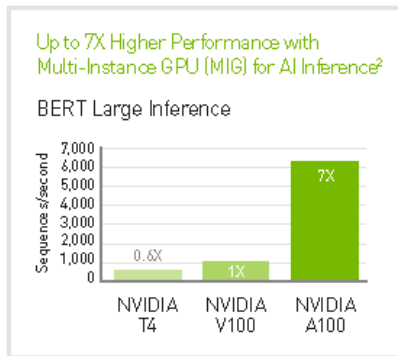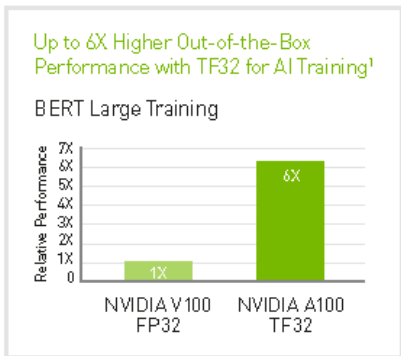# NVIDIA A100
TENSOR CORE GPU

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.
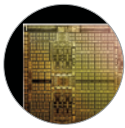
SPECIFICATIONS (PEAK PERFORMANCE)

Part Number	TCSA100M-PB
EAN	3536403378035
GPU Architecture	NVIDIA Ampere
NVIDIA Tensor Cores	640
NVIDIA CUDA® Cores	5,120
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*
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
ECC	Yes
System Interface	PCIe Gen4
Form Factor	PCIe Full Height
Multi-Instance GPU	Up to 7 GPU instances
Max Power Comsumption	250 W
Thermal Solution	Passive
Compute APIs	CUDA, DirectCompute, OpenCL™, OpenACC

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

BERT Large Training

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

BERT Large Inference


Up to 2X More HPC performance[3]

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.
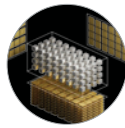
THIRD-GENERATION TENSOR CORES

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.
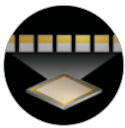
MULTI-INSTANCE GPU (MIG)

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.
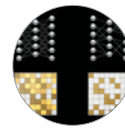
STRUCTURAL SPARSITY

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

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

To learn more about the NVIDIA A100 Tensor Core GPU, visit **www.pny.eu**