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.
Incredible Performance Across Workloads

**Up to 3X Higher AI Training on Largest Models**

**Up to 24X Higher AI Inference Performance over CPUs**

**Up to 1.25X Higher AI Inference Performance over A100**

**Up to 1.8X Higher Performance for HPC Applications**

**Incredible Performance Across Workloads**

**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.

**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.

**3RD GEN**

every user and application.

optimizing utilization and expanding access to

applications, and IT administrators can offer

to breakthrough acceleration for all their

own high-bandwidth memory, cache, and

GPUs via HGX A100 server boards and in PCIe

the highest application performance possible on

a single server. NVLink is available in A100 SXM

GPUs via HGX A100 server boards and in PCIe

to 600 gigabytes per second (GB/sec), unleashing

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.

**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.

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