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 datasets.

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 the NVIDIA NGC™ catalog. 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.

NVIDIA A100 TENSOR CORE GPU SPECIFICATIONS

<table>
<thead>
<tr>
<th></th>
<th>A100 40GB PCIe</th>
<th>A100 80GB PCIe</th>
<th>A100 40GB SXM</th>
<th>A100 80GB SXM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP64</td>
<td>9.7 TFLOPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP64 Tensor Core</td>
<td>19.5 TFLOPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP32</td>
<td>19.5 TFLOPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tensor Float 32 (TF32)</td>
<td>156 TFLOPS</td>
<td>312 TFLOPS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BFLOAT16 Tensor Core</td>
<td>312 TFLOPS</td>
<td>624 TFLOPS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP16 Tensor Core</td>
<td>312 TFLOPS</td>
<td>624 TFLOPS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INT8 Tensor Core</td>
<td>624 TOPS</td>
<td>1248 TOPS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU Memory</td>
<td>40GB HBM2</td>
<td>80GB HBM2e</td>
<td>40GB HBM2</td>
<td>80GB HBM2e</td>
</tr>
<tr>
<td>GPU Memory Bandwidth</td>
<td>1,555GB/s</td>
<td>1,935GB/s</td>
<td>1,555GB/s</td>
<td>2,039GB/s</td>
</tr>
<tr>
<td>Max Thermal Design Power (TDP)</td>
<td>250W</td>
<td>300W</td>
<td>400W</td>
<td>400W</td>
</tr>
<tr>
<td>Multi-Instance GPU</td>
<td>Up to 7 MIGs @ 5GB</td>
<td>Up to 7 MIGs @ 10GB</td>
<td>Up to 7 MIGs @ 5GB</td>
<td>Up to 7 MIGs @ 10GB</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe</td>
<td>SXM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interconnect</td>
<td>NVIDIA® NVLink® Bridge for 2 GPUs: 600GB/s ** PCIe Gen4: 64GB/s</td>
<td>NVLink: 600GB/s PCIe Gen4: 64GB/s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server Options</td>
<td>Partner and NVIDIA-Certified Systems™ with 1-8 GPUs</td>
<td>NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4, 8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* With sparsity
** SXM4 GPUs via HGX A100 server boards; PCIe GPUs via NVLink Bridge for up to two GPUs
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 40GB**

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

---

**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.

**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.

**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.

**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.

---

**NVIDIA A100 TENSOR CORE GPU | DATA SHEET | JUN21 | 2**

- **Groundbreaking Innovations**
- **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 40GB**
- **Up to 1.8X Higher Performance for HPC Applications**

---

**Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma Fine Tuner, Quantum Espresso [AUSURF112-JR], Random Forest FP32 [make_blobs (160000 x 64: 10)], TensorFlow Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma Fine Tuner, Quantum Espresso [AUSURF112-JR], Random Forest FP32 [make_blobs (160000 x 64: 10)], TensorFlow
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 2,000 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.

OPTIMIZED SOFTWARE AND SERVICES FOR ENTERPRISE

EVERY DEEP LEARNING FRAMEWORK

mxnet  PYTORCH  Apache Spark  TensorFlow

2,000+ GPU-ACCELERATED APPLICATIONS

Altair nanoFluidX  Altair ultraFluidX  AMBER  ANSYS Fluent

DS SIMULIA Abaqus  GAUSSIAN  GROMACS  NAMD

OpenFOAM  VASP  WRF

Arrow Contact Information
Email: IntelligentSolutions@arrow.com, Online: www.arrow.com/AIS/nvidia

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, DGX, HGX, NGC, NVIDIA-Certified Systems, NVLink, NVSwitch, and Volta are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. All other trademarks are property of their respective owners. JUN21