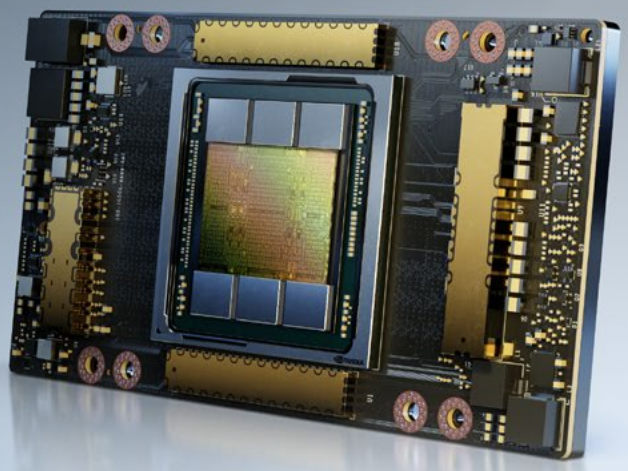
DATASHEET



NVIDIA A100 TENSOR CORE GPU

Unprecedented Acceleration 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 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 (SXM4 AND PCIE FORM FACTORS)

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

* With sparsity

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

*** 400W TDP for standard configuration. HGX A100-80GB CTS (Custom Thermal Solution) SKU can support TDPs up to 500W