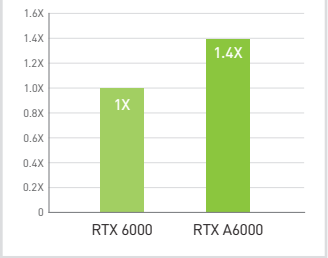
NVIDIA RTX A6000

Powering the World's Highest-Performing Workstations.

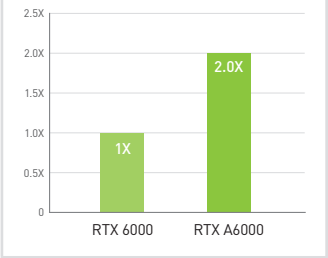
Amplified Performance for Professionals

The NVIDIA RTX™ A6000, built on the NVIDIA Ampere architecture, delivers everything designers, engineers, scientists, and artists need to meet the most graphics and compute-intensive workflows. The RTX A6000 is equipped with the latest generation RT Cores, Tensor Cores, and CUDA® cores for unprecedented rendering, AI, graphics, and compute performance. Certified with a broad range of professional applications, tested by leading independent software vendors (ISVs) and workstation manufacturers, and backed by a global team of support specialists, NVIDIA RTX is the visual computing solution of choice for demanding enterprise deployments.

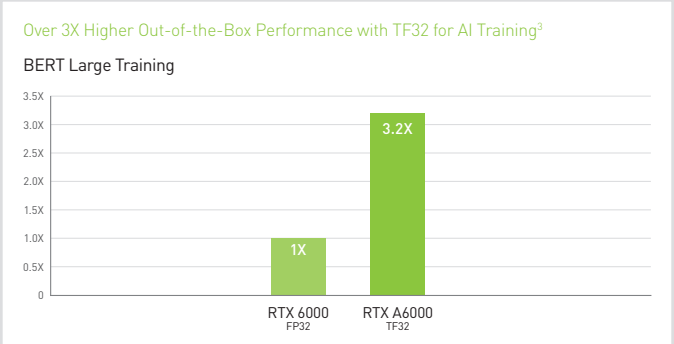
Up to 40% Faster Graphics Performance[1]



Up to 2X Faster Rendering Performance[2]



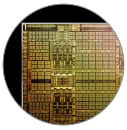
Over 3X Higher Out-of-the-Box Performance with TF32 for AI Training[3]



SPECIFICATIONS

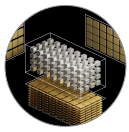
GPU memory	**48GB GDDR6**
Memory interface	**384-bit**
Memory bandwidth	**768 GB/s**
Error-correcting code (ECC)	**Yes**
NVIDIA Ampere architecture-based CUDA Cores	**10,752**
NVIDIA third-generation Tensor Cores	**336**
NVIDIA second-generation RT Cores	**84**
Single-precision performance	**38.7 TFLOPS[7]**
RT Core performance	**75.6 TFLOPS[7]**
Tensor performance	**309.7 TFLOPS[8]**
NVIDIA NVLink	**Connects two NVIDIA RTX A6000 GPUs[12]**
NVIDIA NVLink bandwidth	**112.5 GB/s (bidirectional)**
System interface	**PCIe 4.0 x16**
Power consumption	**Total board power: 300 W**
Thermal solution	**Active**
Form factor	**4.4" H x 10.5" L, dual slot, full height**
Display connectors	**4x DisplayPort 1.4a[9]**
Max simultaneous displays	**4x 4096 x 2160 @ 120 Hz, 4x 5120 x 2880 @ 60 Hz, 2x 7680 x 4320 @ 60 Hz**
Power connector	**1x 8-pin CPU**
Encode/decode engines	**1x encode, 2x decode (+AV1 decode)**
VR ready	**Yes**
vGPU software support	**NVIDIA vPC/vApps, NVIDIA RTX Virtual Workstation**
vGPU profiles supported	**See the Virtual GPU Licensing Guide**
Graphics APIs	**DirectX 12 Ultimate, Shader Model 6.6, OpenGL 4.6[10], Vulkan 1.3[10]**
Compute APIs	**CUDA 11.6, DirectCompute, OpenCL 3.0**

Groundbreaking Innovations



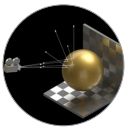
NVIDIA AMPERE ARCHITECTURE

NVIDIA® RTX™ technology revolutionized professional visual computing forever. The NVIDIA Ampere architecture builds on the power of RTX to significantly enhance the performance of rendering, graphics, AI, and compute workloads. Engineered to perfection and featuring cutting-edge innovations, the NVIDIA Ampere architecture takes RTX to new heights for professional workloads.



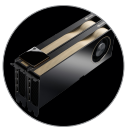
THIRD-GENERATION TENSOR CORES

New Tensor Float 32 (TF32) precision provides up to 5X the training throughput over the previous generation to accelerate AI and data science model training without requiring any code changes. Hardware support for structural sparsity doubles the throughput for inferencing. Tensor Cores also bring AI to graphics with capabilities like DLSS, AI denoising, and enhanced editing for select applications.



SECOND-GENERATION RT CORES

With up to 2X the throughput over the previous generation and the ability to concurrently run ray tracing with either shading or denoising capabilities, second-generation RT Cores deliver massive speedups for workloads like photorealistic rendering of movie content and virtual prototyping of product designs. This technology also speeds up the rendering of ray-traced motion blur for faster results with greater visual accuracy.



THIRD-GENERATION NVLINK

Third-generation NVIDIA NVLink® technology enables users to connect two GPUs together to share GPU performance and memory. With up to 112 gigabytes per second (GB/s) of bidirectional bandwidth and combined graphics memory of up to 96GB, professionals can tackle the largest rendering, AI, virtual reality, and visual computing workloads. The new NVLink connector also features a shorter Z height, which enables NVLink functionality in a wider range of chassis.

NVIDIA AMPERE ARCHITECTURE-BASED CUDA CORES

The NVIDIA Ampere architecture's CUDA® cores bring double-speed processing for single-precision floating point (FP32) operations and are up to 2X more power efficient than Turing GPUs. This provides significant performance gains for graphics workflows like 3D model development and compute workflows like desktop simulation for computer-aided engineering (CAE).

PCI EXPRESS GEN 4.0

NVIDIA Ampere architecture-based GPUs support PCI Express Gen 4.0 (PCIe Gen 4.0), which provides 2X the bandwidth of PCIe Gen 3.0. This improves data transfer speeds from CPU memory for data-intensive tasks such as AI and data science. Faster PCIe performance also accelerates GPU direct memory access (DMA) transfers, enabling faster video data transfers from GPUDirect® for video-enabled devices and faster input/output (I/O) with GPUDirect Storage.

Features

- PCI Express Gen 4
- Four DisplayPort 1.4a connectors
- AV1 decode support
- DisplayPort with audio
- VGA support[4]
- 3D stereo support with stereo connector
- NVIDIA GPUDirect® for Video support
- NVIDIA virtual GPU (vGPU) software support
- NVIDIA Quadro® Sync II[5] compatibility
- NVIDIA RTX Experience™
- NVIDIA RTX Desktop Manager software
- NVIDIA RTX IO support
- HDCP 2.2 support
- NVIDIA Mosaic[6] technology

1 Tests run on workstation with 1x Xeon Gold 6154, 3GHz (3.7GHz Turbo), Win10 x 64, NVIDIA driver version 460.48. SPECviewperf 2020, energy subtest. | 2 Tests run on workstation with 2x Xeon Gold 6126, 2.6GHz (3.7GHz Turbo), Win10 x 64, NVIDIA driver version 456.37. Autodesk VRED 221.0 GA Release. | 3 Tests run on workstation with AMD Ryzen 3900X, 3.8GHz, 4.6 Turbo, NVIDIA driver 460.17, BERT pre-training throughput using Pytorch, phase 1 sequence length 128, RTX 6000 using FP32 precision, RTX A6000 using TF32 precision. | 4 Via adapter/connector/bracket. | 5 Quadro Sync II card sold separately. | 6 Windows 10, Windows 11, and Linux. | 7 Peak rates based on GPU Boost Clock. | 8 Effective teraFLOPS (TFLOPS) using the new sparsity feature. | 9 Display ports are on by default for Quadro RTX A6000. Display ports are not active when using vGPU software. | 10 Product is based on a published Khronos specification and is expected to pass the Khronos conformance testing process when available. | 11 NVIDIA NVLink sold separately. Current conformance status can be found at **www.khronos.org/conformance**

Learn more

To learn more about the NVIDIA RTX A6000, visit **ww.nvidia.com/rtx-a6000/**