NVIDIA® RTX™ A4000 GPU

Performance Features

NVIDIA Ampere Architecture

NVIDIA RTX A4000 is the most powerful single slot GPU solution offering high performance real-time ray tracing, AI-accelerated compute, and professional graphics rendering. Building upon the major SM enhancements from the Turing GPU, the NVIDIA Ampere architecture enhances ray tracing operations, tensor matrix operations, and concurrent executions of FP32 and INT32 operations.

CUDA Cores

The NVIDIA Ampere architecture-based CUDA cores bring up to 2.7X the single-precision floating point (FP32) throughput compared to the previous generation, providing significant performance improvements for graphics workflows such as 3D model development and compute for workloads such as desktop simulation for computer-aided engineering (CAE). The RTX A4000 enables two FP32 primary data paths, doubling the peak FP32 operations.

2nd Generation RT Cores

Incorporating 2nd generation ray tracing engines, NVIDIA Ampere architecture-based GPUs provide incredible ray traced rendering performance. A single RTX A4000 board can render complex professional models with physically accurate shadows, reflections, and refractions to empower users with instant insight. Working in concert with applications leveraging APIs such as NVIDIA OptiX, Microsoft DXR and Vulkan ray tracing, systems based on the RTX A4000 will power truly interactive design workflows to provide immediate feedback for unprecedented levels of productivity. The RTX A4000 is up to 2X faster in ray tracing compared to the previous generation. This technology also speeds up the rendering of ray-traced motion blur for faster results with greater visual accuracy.

3rd Generation Tensor Cores

Purpose-built for deep learning matrix arithmetic at the heart of neural network training and inferencing functions, the RTX A4000 includes enhanced Tensor Cores that accelerate more datatypes and includes a new Fine-Grained Structured Sparsity feature that delivers up to 2X throughput for tensor matrix operations compared to the previous generation. New Tensor Cores will accelerate two new TF32 and BFloat16 precision modes. Independent floating-point and integer data paths allow more efficient execution of workloads using a mix of computation and addressing calculations.

PCIe Gen 4

The RTX A4000 supports PCI Express Gen 4, which provides double the bandwidth of PCIe Gen 3, improving data-transfer speeds from CPU memory for data-intensive tasks like AI and data science.

Higher Speed GDDR6 Memory
Built with 16GB GDDR6 memory delivering up to 23% greater throughput for ray tracing, rendering, and AI workloads than the previous generation. The RTX A4000 provides the industry's largest graphics memory footprint to address the largest datasets and models in latency-sensitive professional applications.

Error Correcting Code (ECC) on Graphics Memory
Meet strict data integrity requirements for mission critical applications with uncompromised computing accuracy and reliability for workstations.

5th Generation NVDEC Engine[1]
NVDEC is well suited for transcoding and video playback applications for real-time decoding. The following video codecs are supported for hardware-accelerated decoding: MPEG-2, VC-1, H.264 (AVCHD), H.265 (HEVC), VP8, VP9, and AV1.

7th Generation NVENC Engine[1]
NVENC can take on the most demanding 4K or 8K video encoding tasks to free up the graphics engine and the CPU for other operations. The RTX A4000 provides better encoding quality than software-based x264 encoders.

Graphics Preemption
Pixel-level preemption provides more granular control to better support time-sensitive tasks such as VR motion tracking.

Compute Preemption
Preemption at the instruction-level provides finer grain control over compute tasks to prevent long-running applications from either monopolizing system resources or timing out.

NVIDIA RTX IO
Accelerating GPU-based lossless decompression performance by up to 100x and 20x lower CPU utilization compared to traditional storage APIs using Microsoft's new DirectStorage for Windows API. RTX IO moves data from the storage to the GPU in a more efficient, compressed form, and improving I/O performance.

Multi-GPU Technology

[1] This feature requires implementation by software applications, and it is not a stand-alone utility. Please contact quadrohelp@nvidia.com for details on availability.

NVIDIA® SLI® Technology
Leverage multiple GPUs to dynamically scale graphics performance, enhance image quality, expand display real estate, and assemble a fully virtualized system.

Display Features

NVIDIA® Mosaic Technology
Transparently scale the desktop and applications across up to 4 GPUs and 16 displays from a single workstation while delivering full performance and image quality.

DisplayPort 1.4a
Support up to four 5K monitors @ 60Hz, or dual 8K displays @ 60Hz per card. The RTX A4000 supports HDR color for 4K @ 60Hz for 10/12b HEVC decode and up to 4K @ 60Hz for 10b HEVC encode. Each DisplayPort connector can drive ultra-high resolutions of 4096x2160 @ 120 Hz with 30-bit color.

NVIDIA® RTX™ Desktop Manager[2]
Gain unprecedented end-user control of the desktop experience for increased productivity in single large display or multi-display environments, especially in the current age of large, widescreen displays.

NVIDIA® Quadro Sync II[3]
Synchronize the display and image output of up to 32 displays[iii] from 8 GPUs (connected through two Sync II boards) in a single system, reducing the number of machines needed to create an advanced video visualization environment.

Frame Lock Connector Latch
Each frame lock connector is designed with a self-locking retention mechanism to secure its connection with the frame lock cable to provide robust connectivity and maximum productivity.

OpenGL Quad Buffered Stereo Support
Provide a smooth and immersive 3D Stereo experience for professional applications.

Ultra-High-Resolution Desktop Support
Get more Mosaic topology choices with high resolution displays devices with a 32K Max desktop size.

Professional 3D Stereo Synchronization
Robust control of stereo effects through a dedicated connection to directly synchronize 3D stereo hardware to an NVIDIA RTX professional graphics card.

---

[2] Product formerly known as NVIDIA Quadro View has undergone a brand transition.
[3] Feature supported in future driver release.

Software Support

NVIDIA® RTX™ Experience[4]
NVIDIA RTX Experience delivers a suite of productivity tools to your desktop workstation, including desktop recording in up to 8K, automatic alerts for the latest NVIDIA RTX Enterprise driver updates, and access gaming features. The application is available for download here.

Software Optimized for AI
Deep learning frameworks such as Caffe2, MXNet, CNTK, TensorFlow, and others deliver dramatically faster training times and higher multi-node training performance. GPU accelerated libraries such as cuDNN, cuBLAS, and TensorRT delivers higher performance for both deep learning inference and High-Performance Computing (HPC) applications.

NVIDIA® CUDA® Parallel Computing Platform
Natively execute standard programming languages like C/C++ and Fortran, and APIs such as OpenCL, OpenACC and Direct Compute to accelerates techniques such as ray tracing, video and image processing, and computation fluid dynamics.

Unified Memory
A single, seamless 49-bit virtual address space allows for the transparent migration of data between the full allocation of CPU and GPU memory.

NVIDIA® GPUDirect for Video
GPUDirect for Video speeds communication between the GPU and video I/O devices by avoiding unnecessary system memory copies and CPU overhead.

NVIDIA Enterprise-Management Tools
Maximize system uptime, seamlessly manage wide-scale deployments, and remotely control graphics and display settings for efficient opera ions.

[4] Product formerly known as NVIDIA Quadro Experience, rebrand effective 4/21/2021.