The NVIDIA A40 accelerates the most demanding visual computing workloads from the data center, combining the latest NVIDIA Ampere architecture RT Cores, Tensor Cores, and CUDA® Cores with 48 GB of graphics memory. From powerful virtual workstations accessible from anywhere to dedicated render nodes, NVIDIA A40 brings next-generation NVIDIA RTX™ technology to the data center for the most advanced professional visualization workloads.
A Look Inside the NVIDIA Ampere Architecture

**NVIDIA AMPERE ARCHITECTURE CUDA CORES**
Double-speed processing for single-precision floating point (FP32) operations and improved power efficiency provide significant performance gains in graphics and compute workflows such as complex 3D computer-aided design (CAD) and computer-aided engineering (CAE).

**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, architectural design evaluations, 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 TENSOR CORES**
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 any code changes. Hardware support for structural sparsity provides up to double the throughput for inferencing. Tensor Cores also bring AI to graphics with capabilities like deep learning super sampling (DLSS), AI denoising, and enhanced editing for select applications.

**48 GB GDDR6 MEMORY WITH NVLINK**
Ultra-fast GDDR6 memory, scalable up to 96 GB with NVLink®, gives data scientists, engineers, and creative professionals the large memory necessary to work with massive datasets and workloads like data science and simulation.

**PCI EXPRESS GEN 4**
PCI Express Gen 4 doubles the bandwidth of PCIe Gen 3, improving data-transfer speeds from CPU memory for data-intensive tasks like AI, data science, and 3D design. Faster PCIe performance also accelerates GPU direct memory access (DMA) transfers, providing faster input/output communication of video data between the GPU and GPUDirect® for Video-enabled devices to deliver a powerful solution for live broadcast. A40 is backwards compatible with PCI Express Gen 3 for deployment flexibility.

**DATA CENTER EFFICIENCY AND SECURITY**
Featuring a dual-slot, power-efficient design, NVIDIA A40 is up to 2X as power efficient as the previous generation and compatible with a wide range of servers from worldwide OEMs. The NVIDIA A40 includes secure and measured boot with hardware root-of-trust technology, ensuring that firmware isn’t tampered with or corrupted.

The NVIDIA A40 GPU delivers state-of-the-art visual computing capabilities, including real-time ray tracing, AI acceleration, and multi-workload flexibility to accelerate deep learning, data science, and compute-based workloads. Virtual workstations powered by NVIDIA A40 and NVIDIA RTX Virtual Workstation (vWS) and NVIDIA Virtual Compute Server software benefit from extensive testing across a broad range of industry applications and professional software for optimal performance and stability.

**EVERY DEEP LEARNING FRAMEWORK**
To learn more about the NVIDIA A40 GPU, visit [www.nvidia.com/a40](http://www.nvidia.com/a40)

Learn more

© 2021 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA, GRID, GPUDirect, NVLink, OpenACC, Quadro, and RTX 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. DEC21