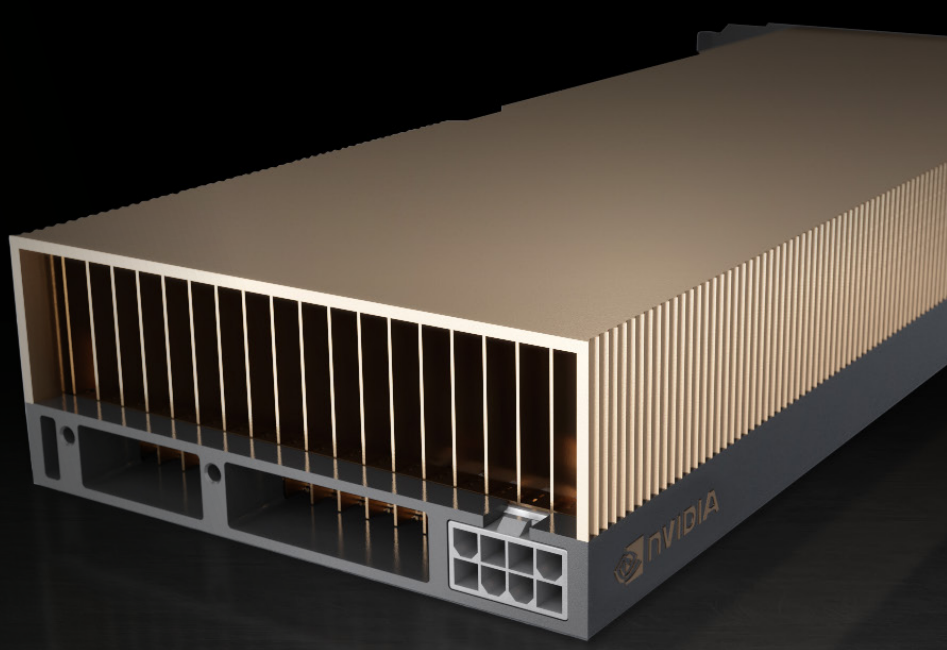
DATASHEET

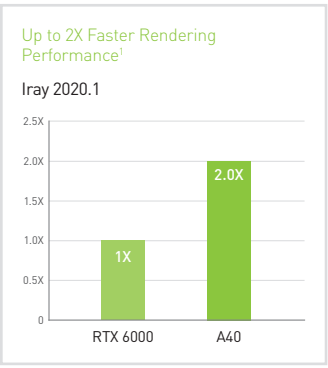


NVIDIA A40

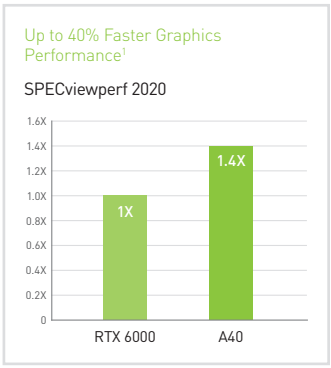
Powerful Data Center GPU For Visual Computing

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.

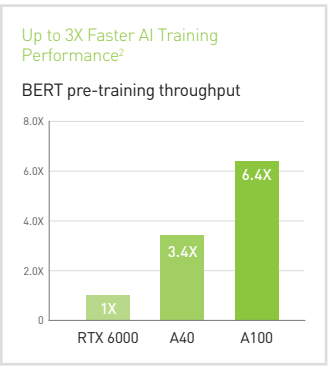
Up to 2X Faster Rendering Performance[1]



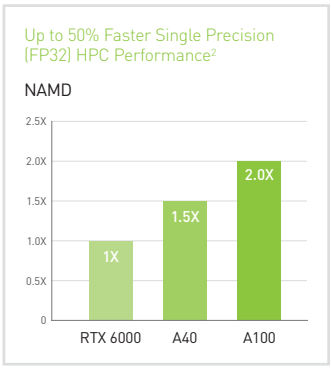
Up to 40% Faster Graphics Performance[1]



Up to 3X Faster AI Training Performance[2]



Up to 50% Faster Single Precision (FP32) HPC Performance[2]



SPECIFICATIONS

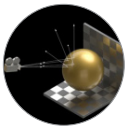
GPU architecture	**NVIDIA Ampere architecture**
GPU memory	**48 GB GDDR6 with ECC**
Memory bandwidth	**696 GB/s**
Interconnect interface	**NVIDIA® NVLink® 112.5 GB/s (bidirectional)[3] PCIe Gen4: 64GB/s**
NVIDIA Ampere architecture-based CUDA Cores	**10,752**
NVIDIA second-generation RT Cores	**84**
NVIDIA third-generation Tensor Cores	**336**
Peak FP32 TFLOPS (non-Tensor)	**37.4**
Peak FP16 Tensor TFLOPS with FP16 Accumulate	**149.7 | 299.4***
Peak TF32 Tensor TFLOPS	**74.8 | 149.6***
RT Core performance TFLOPS	**73.1**
Peak BF16 Tensor TFLOPS with FP32 Accumulate	**149.7 | 299.4***
Peak INT8 Tensor TOPS Peak INT 4 Tensor TOPS	**299.3 | 598.6*** **598.7 | 1,197.4***
Form factor	**4.4" (H) x 10.5" (L) dual slot**
Display ports	**3x DisplayPort 1.4**; Supports NVIDIA Mosaic and Quadro® Sync[4]**
Max power consumption	**300 W**
Power connector	**8-pin CPU**
Thermal solution	**Passive**
Virtual GPU (vGPU) software support	**NVIDIA vPC/vApps, NVIDIA RTX Virtual Workstation, NVIDIA Virtual Compute Server**
vGPU profiles supported	**See the Virtual GPU Licensing Guide**
NVENC | NVDEC	**1x | 2x (includes AV1 decode)**
Secure and measured boot with hardware root of trust	**Yes (optional)**
NEBS ready	**Level 3**
Compute APIs	**CUDA, DirectCompute, OpenCL™, OpenACC®**
Graphics APIs	**DirectX 12.07[5], Shader Model 5.17[5], OpenGL 4.68[6], Vulkan 1.18[6]**
MIG support	**No**

* Structural sparsity enabled
** A40 is configured for virtualization by default with physical display connectors disabled. The display outputs can be enabled via management software tools.

A Look Inside the NVIDIA Ampere Architecture

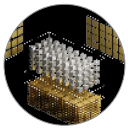
NVIDIA AMPERE ARCHITECTURE CUDA CORES

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**[3], gives data scientists, engineers, and creative professionals the large memory necessary to work with massive datasets and workloads like data science and simulation.

PCIE 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

mxnet

PYTORCH

Apache Spark

TensorFlow

RTX FOR PROFESSIONAL APPLICATIONS

Adobe Premiere Pro

SOLIDWORKS

PLM Software SIEMENS NX

AUTODESK ARNOLD

REDSHIFT

AUTODESK VRED

KeyShot by Luxion

UNREAL ENGINE

blender

octanerender

V-RAY

Learn more

To learn more about the NVIDIA A40 GPU, visit **www.nvidia.com/a40**

1 Rendering and Graphics tests run on 2x Xeon Gold 6126 2.6GHz (3.7GHz Turbo). 256GB system memory. NVIDIA Driver 461.09. Rendering test: Iray 2020.1, Render time of NVIDIA Endeavor scene. Graphics test: SPECviewperf 2020 Subtest, 4K medical-03 Composite | 2 AI and HPC tests run on AMD EPYC 7742@2.25GHz (3.4GHz Turbo). 512GB system memory. NVIDIA Driver 460.14. AI Training: BERT pre-training throughput. PyTorch (2/3) Phase 1 and (1/3) Phase 2. Precision FP32 for RTX 6000 and TF32 for A40 and A100. Sequence length for Phase 1 = 128. Phase 2 = 512. Single Precision HPC: NAMD version 3.0a7, stmv_nve_cuda; Precision=FP32; ns/day, CUDA Version: 11.1.74 | 3 Connecting two NVIDIA A40 cards with NVLink to scale performance and memory capacity to 96 GB is only possible if your application supports NVLink technology. Please contact your application provider to confirm their support for NVLink. | 4 Quadro Sync II card sold separately. Mosaic supported on Windows 10 and Linux. | 5 GPU supports DX 12.0 API, Hardware Feature Level 12 + 1. | 6 Product is based on a published Khronos specification and is expected to pass the Khronos conformance testing process when available. Current conformance status can be found at www.krhonos.org/conformance

NVIDIA