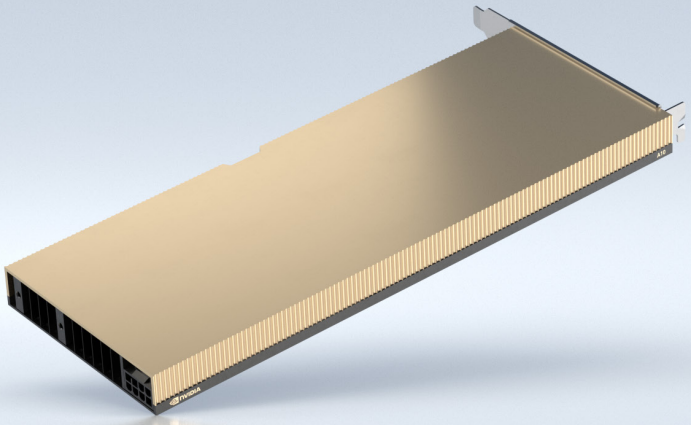
DATASHEET



NVIDIA A10

Accelerated Graphics and Video with AI for Mainstream Enterprise Servers

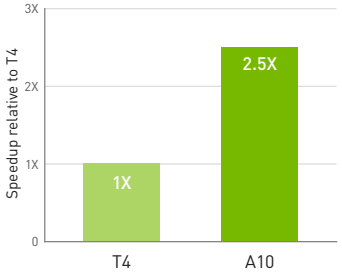
Enrich Graphics and Video Applications with Powerful AI

The NVIDIA A10 Tensor Core GPU combines with NVIDIA RTX Virtual Workstation (vWS) software to bring mainstream graphics and video with AI services to mainstream enterprise servers, delivering the solutions that designers, engineers, artists, and scientists need to meet today's challenges. Built on the latest NVIDIA Ampere architecture, the A10 combines second-generation RT Cores, third-generation Tensor Cores, and new streaming microprocessors with 24 gigabytes (GB) of GDDR6 memory—all in a 150W power envelope—for versatile graphics, rendering, AI, and compute performance. From virtual workstations, accessible anywhere in the world, to render nodes to the data centers running a variety of workloads, A10 is built to deliver optimal performance in a single-wide, full-height, full-length PCIe form factor.

NVIDIA A10 is supported as part of NVIDIA-Certified Systems™, in the on-prem data center, in the cloud, and at the edge. NVIDIA A10 builds on the rich ecosystem of AI frameworks from the NVIDIA NGC™ catalog, CUDA-X™ libraries, over 2.3 million developers, and over 1,800 GPU-optimized applications to help enterprises solve the most critical challenges in their business.

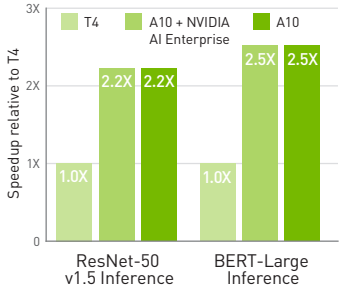
A10 delivers up to 2.5X faster virtual workstation performance compared to T4[1]

SPECviewperf 2020



A10 delivers up to 2.5X more inference performance compared to T4[2]

NVIDIA AI Enterprise delivers equivalent performance to bare metal



SPECIFICATIONS

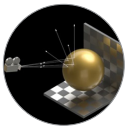
FP32	31.2 TF
TF32 Tensor Core	62.5 TF | 125 TF*
BFLOAT16 Tensor Core	125 TF | 250 TF*
FP16 Tensor Core	125 TF | 250 TF*
INT8 Tensor Core	250 TOPS | 500 TOPS*
INT4 Tensor Core	500 TOPS | 1000 TOPS*
RT Cores	72
Encode / Decode	1 encoder 2 decoders (+AV1 decode)
GPU Memory	24 GB GDDR6
GPU Memory Bandwidth	600 GB/s
Interconnect	PCIe Gen4: 64 GB/s
Form Factor	1-slot FHFL
Max TDP Power	150W
vGPU Software Support	NVIDIA vPC/vApps, NVIDIA RTX™ vWS, NVIDIA AI Enterprise
Secure and Measured Boot with Hardware Root of Trust	Yes (optional)
NEBS Ready	Level 3
Power Connector	PEX 8-pin

*with sparsity

A Look Inside the NVIDIA Ampere Architecture

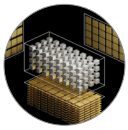
NVIDIA AMPERE ARCHITECTURE CUDA CORES

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).



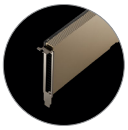
SECOND-GENERATION RT CORES

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.



24 GB GDDR6

Ultra-fast GDDR6 memory, delivering 600 GB/s of bandwidth for rendering, data science, engineering simulation, and other GPU-memory intensive workloads.

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 **NVIDIA GPUDirect® for video**-enabled devices, delivering a powerful solution for live broadcast. A10 is also backwards compatible with PCI Express Gen 3 for deployment flexibility.

DATA CENTER EFFICIENCY AND SECURITY

Featuring a single-slot, full-height, full-length power efficient design, NVIDIA A10 is compatible with a wide range of servers from worldwide OEMs. The NVIDIA A10 includes secure and measured boot with hardware root-of-trust technology, ensuring that firmware isn't tampered with or corrupted.

NVIDIA A10 Tensor Core GPU is ideal for mainstream graphics and video with AI. 2nd Gen RT Cores and 3rd Gen Tensor Cores enrich graphics and video applications with powerful AI in 150W TDP for mainstream servers.

NVIDIA A10 also combines with NVIDIA virtual GPU (vGPU) software to accelerate multiple data center workloads—from graphics-rich VDI to high-performance virtual workstations to AI—in an easily managed, secure, and flexible infrastructure that can be scaled to accommodate resource needs.

EVERY DEEP LEARNING FRAMEWORK

mxnet

PYTORCH

Spark

TensorFlow

RTX FOR PROFESSIONAL APPLICATIONS

AUTODESK® REVIT®

CATIA

SOLIDWORKS

creo®

Rhinoceros®
design, model, present, analyze, realize...

SIEMENS

Learn More

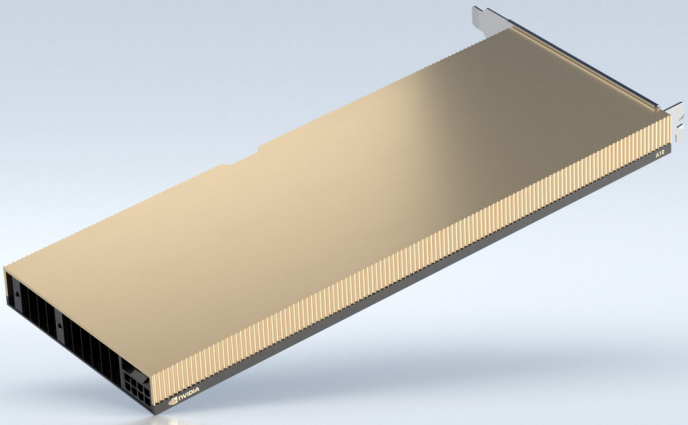
To learn more about the NVIDIA A10 Tensor Core GPU, visit **www.nvidia.com/a10**

1 Test run on a server with 2x Xeon Gold 6154 3.0GHz (3.7GHz Turbo), NVIDIA RTX vWS software, VMware ESXi 7 U2, host/guest driver 461.33. | SPECviewperf 2020 Subtest, and HD 3dsmax-07 composite.

2 BERT Large inference NVIDIA TensorRT7.2, Seq Length =128, batch size =128; NGC Container: 21.02-py3 | ResNet-50 v1.5: NVIDIA TensorRT7.2, INT8 precision batch size = 128 NGC Container: 20.12-py3 | NVIDIA A10 with vCS software, VMware ESXi 7 U2 host/guest driver 461.33

FPO

NVIDIA



NVIDIA A10

Accelerated Graphics and Video with AI for Mainstream Enterprise Servers

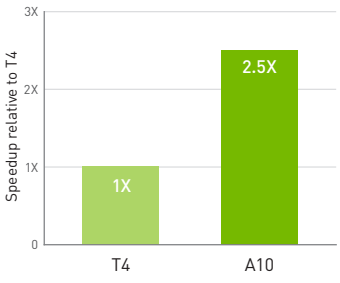
Enrich Graphics and Video Applications with Powerful AI

The NVIDIA A10 Tensor Core GPU combines with NVIDIA RTX Virtual Workstation (vWS) software to bring mainstream graphics and video with AI services to mainstream enterprise servers, delivering the solutions that designers, engineers, artists, and scientists need to meet today's challenges. Built on the latest NVIDIA Ampere architecture, the A10 combines second-generation RT Cores, third-generation Tensor Cores, and new streaming microprocessors with 24 gigabytes (GB) of GDDR6 memory—all in a 150W power envelope—for versatile graphics, rendering, AI, and compute performance. From virtual workstations, accessible anywhere in the world, to render nodes to the data centers running a variety of workloads, A10 is built to deliver optimal performance in a single-wide, full-height, full-length PCIe form factor.

NVIDIA A10 is supported as part of NVIDIA-Certified Systems™, in the on-prem data center, in the cloud, and at the edge. NVIDIA A10 builds on the rich ecosystem of AI frameworks from the NVIDIA NGC™ catalog, CUDA-X™ libraries, over 2.3 million developers, and over 1,800 GPU-optimized applications to help enterprises solve the most critical challenges in their business.

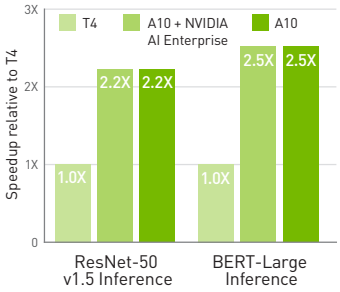
A10 delivers up to 2.5X faster virtual workstation performance compared to T4[1]

SPECviewperf 2020



A10 delivers up to 2.5X more inference performance compared to T4[2]

NVIDIA AI Enterprise delivers equivalent performance to bare metal



SPECIFICATIONS

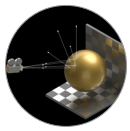
FP32	31.2 TF
TF32 Tensor Core	62.5 TF | 125 TF*
BFLOAT16 Tensor Core	125 TF | 250 TF*
FP16 Tensor Core	125 TF | 250 TF*
INT8 Tensor Core	250 TOPS | 500 TOPS*
INT4 Tensor Core	500 TOPS | 1000 TOPS*
RT Cores	72
Encode / Decode	1 encoder 2 decoders (+AV1 decode)
GPU Memory	24 GB GDDR6
GPU Memory Bandwidth	600 GB/s
Interconnect	PCIe Gen4: 64 GB/s
Form Factor	1-slot FHFL
Max TDP Power	150W
vGPU Software Support	NVIDIA vPC/vApps, NVIDIA RTX™ vWS, NVIDIA AI Enterprise
Secure and Measured Boot with Hardware Root of Trust	Yes (optional)
NEBS Ready	Level 3
Power Connector	PEX 8-pin

*with sparsity

A Look Inside the NVIDIA Ampere Architecture

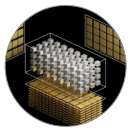
NVIDIA AMPERE ARCHITECTURE CUDA CORES

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).



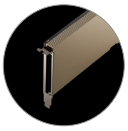
SECOND-GENERATION RT CORES

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.



24 GB GDDR6

Ultra-fast GDDR6 memory, delivering 600 GB/s of bandwidth for rendering, data science, engineering simulation, and other GPU-memory intensive workloads.

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 **NVIDIA GPUDirect® for video**-enabled devices, delivering a powerful solution for live broadcast. A10 is also backwards compatible with PCI Express Gen 3 for deployment flexibility.

DATA CENTER EFFICIENCY AND SECURITY

Featuring a single-slot, full-height, full-length power efficient design, NVIDIA A10 is compatible with a wide range of servers from worldwide OEMs. The NVIDIA A10 includes secure and measured boot with hardware root-of-trust technology, ensuring that firmware isn't tampered with or corrupted.

NVIDIA A10 Tensor Core GPU is ideal for mainstream graphics and video with AI. 2nd Gen RT Cores and 3rd Gen Tensor Cores enrich graphics and video applications with powerful AI in 150W TDP for mainstream servers.

NVIDIA A10 also combines with NVIDIA virtual GPU (vGPU) software to accelerate multiple data center workloads—from graphics-rich VDI to high-performance virtual workstations to AI—in an easily managed, secure, and flexible infrastructure that can be scaled to accommodate resource needs.

EVERY DEEP LEARNING FRAMEWORK

mxnet

PYTORCH

Apache Spark

TensorFlow

RTX FOR PROFESSIONAL APPLICATIONS

AUTODESK REVIT

creo

CATIA

Rhinoceros® design, model, present, analyze, realize...

SOLIDWORKS

SIEMENS

Learn More

To learn more about the NVIDIA A10 Tensor Core GPU, visit **www.nvidia.com/a10**

1 Test run on a server with 2x Xeon Gold 6154 3.0GHz (3.7GHz Turbo), NVIDIA RTX vWS software, VMware ESXi 7 U2, host/guest driver 461.33. | SPECviewperf 2020 Subtest, and HD 3dsmax-07 composite.

2 BERT Large inference NVIDIA TensorRT7.2, Seq Length =128, batch size =128; NGC Container: 21.02-py3 | ResNet-50 v1.5: NVIDIA TensorRT7.2, INT8 precision batch size = 128 NGC Container: 20.12-py3 | NVIDIA A10 with vCS software, VMware ESXi 7 U2 host/guest driver 461.33

FPO

NVIDIA