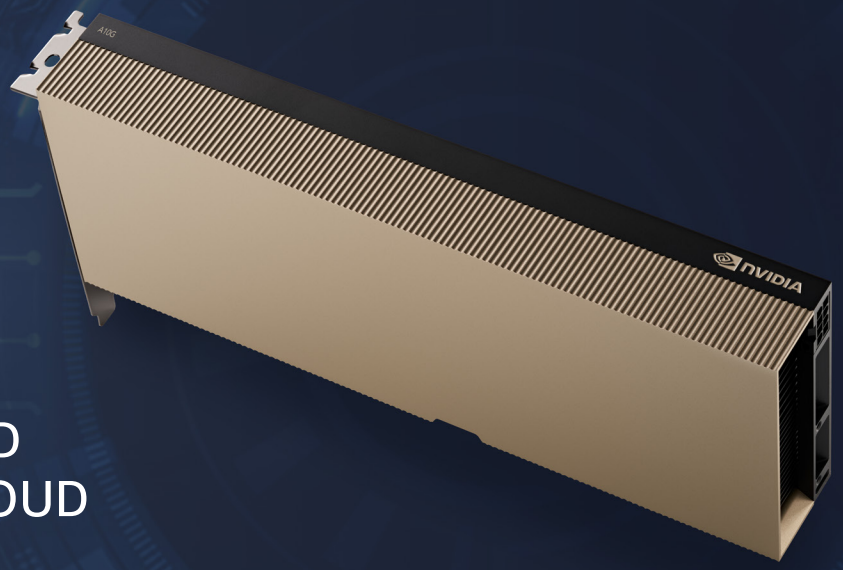


NVIDIA A10G TENSOR CORE GPU

ACCELERATED COMPUTE AND GRAPHICS FOR THE AWS CLOUD

Accelerated Graphics and AI from the Cloud

The NVIDIA A10G Tensor Core GPU brings high end graphics, video and AI to the cloud, delivering the solutions that designers, game developers, engineers, artists, researchers, and scientists need to meet today's challenges and do their work from anywhere.

NVIDIA A10G brings next-generation NVIDIA RTX™ technology to the cloud, with NVIDIA RTX Virtual Workstations (vWS) to deliver the most advanced professional visualization workloads like interactive video rendering, video editing, computer-aided design, photorealistic simulations, and 3D visualization. The NVIDIA Gaming AMI driver enables graphics-rich cloud gaming.

Built on the latest NVIDIA Ampere architecture, the A10G combines second-generation RT Cores, third-generation Tensor Cores, and new streaming microprocessors with 24 gigabytes (GB) of GDDR6 memory in a 300W power envelope for the most demanding graphics, rendering and real-time AI inference performance.

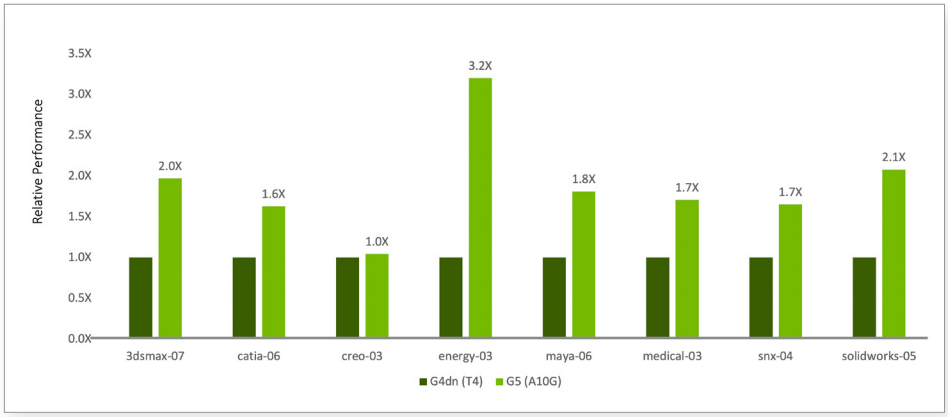
NVIDIA A10G builds on the rich ecosystem of AI frameworks from the NVIDIA NGC™ catalog, CUDA-X™ libraries, over 2.3 million developers, and over 1,800 GPU-optimized applications to help enterprises solve the most critical challenges in their business.

Specifications

FP32	**35 TF**
TF32 Tensor Core	**35 TF | 70 TF***
BFLOAT16 Tensor Core	**70 TF | 140 TF***
FP16 Tensor Core	**70 TF | 140 TF***
INT8 Tensor Core	**140 TOPS | 280 TOPS***
INT4 Tensor Core	**280 TOPS | 560 TOPS***
RT Cores	**80**
Encode / Decode	**1 encoder 2 decoder (+AV1 decode)**
GPU Memory	**24 GB GDDR6**
GPU Memory Bandwidth	**600 GB/s**
Interconnect	**PCIe Gen4: 64 GB/s**
Form Factor	**2-slot FHFL**
Max TDP Power	**300W**
vGPU Software Support	**NVIDIA RTX™ vWS**

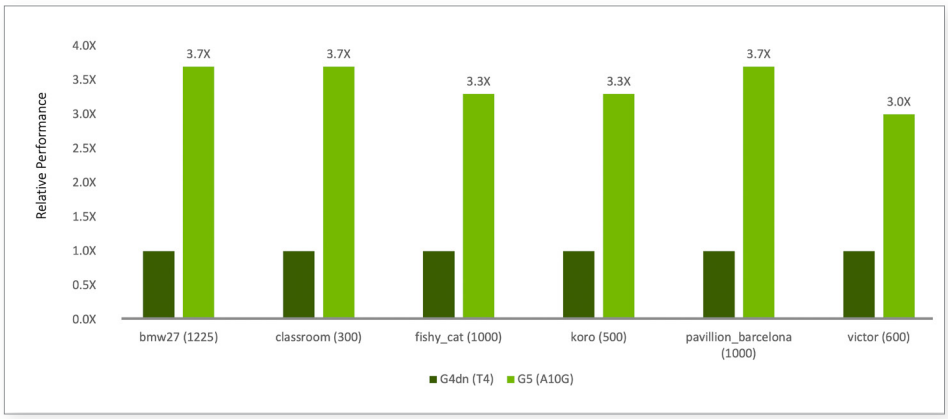
*with sparsity

3.2x Better professional graphics performance



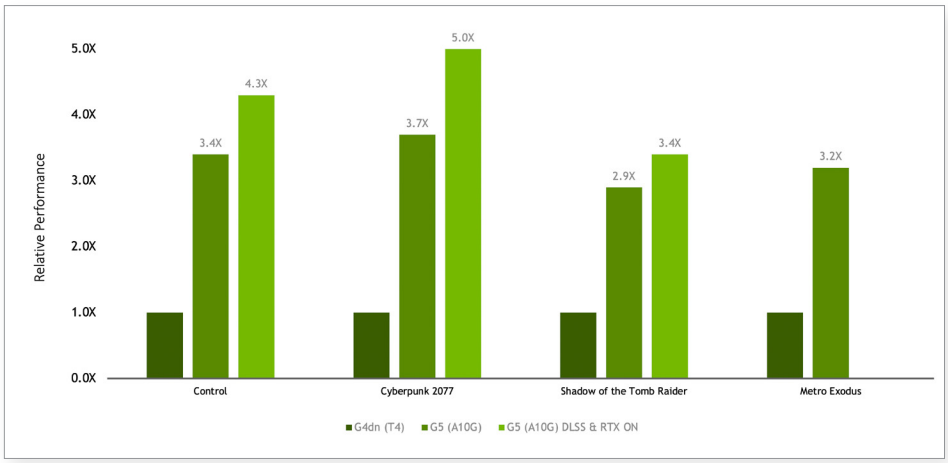
Relative SPECviewperf 2020 performance at 1920x1080 comparing AWS G4dn with NVIDIA T4 vs. AWS G5 with NVIDIA A10G.

3.7x Better rendering performance



Relative Blender Cycles performance comparing AWS G4dn with NVIDIA T4 vs. AWS G5 with NVIDIA A10G.

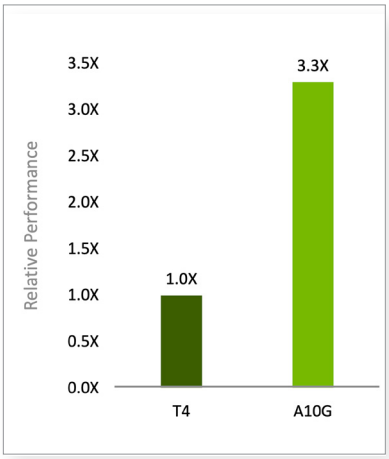
5x Better gaming performance with NVIDIA RTX and DLSS



Relative gaming performance comparing AWS G4dn with NVIDIA T4 GPUs vs. AWS G5 with NVIDIA A10G GPUs vs. AWS G5 with A10G GPUs (4K) with DLSS and RTX On. DLSS support on “Metro Exodus” is coming soon.

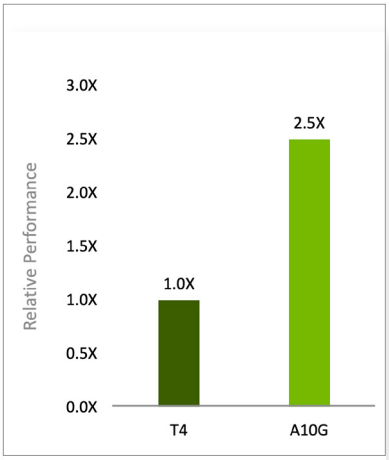
3.3x Better inferencing performance

BERT Large Inference



Relative BERT Large Inference performance comparing T4 vs. A10G. NVIDIA TensorRT 8.0 Seq. length=384, INT8 precision batch size=128, NGC Container: 21.11.

ResNet-50 Inference



Relative ResNet-50 Inference performance comparing T4 vs. A10G. ResNet-50 v1.5: NVIDIA TensorRT 8.0, INT8 precision batch size=128 NGC Container: 21.11.

A Look Inside the NVIDIA Ampere Architecture

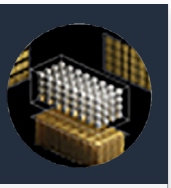
NVIDIA Ampere Architecture CUDA Cores

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 accelerates 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®

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. A10G is also backwards compatible with PCI Express Gen 3 for deployment flexibility.

Best platform for rendering

NVIDIA RTX is supported by the industry's top rendering applications, giving users the ability to utilize GPU-accelerated rendering in the package of their choice.

NVIDIA A10G Tensor Core GPU is ideal for high end graphics, video and AI. 2nd Gen RT Cores and 3rd Gen Tensor Cores enrich graphics, video and AI workloads in 300W TDP.

Every Deep Learning Framework

RTX for Professional Applications

To learn more about the NVIDIA A10G Tensor Core GPU, visit **https://aws.amazon.com/ec2/instance-types/g5/**